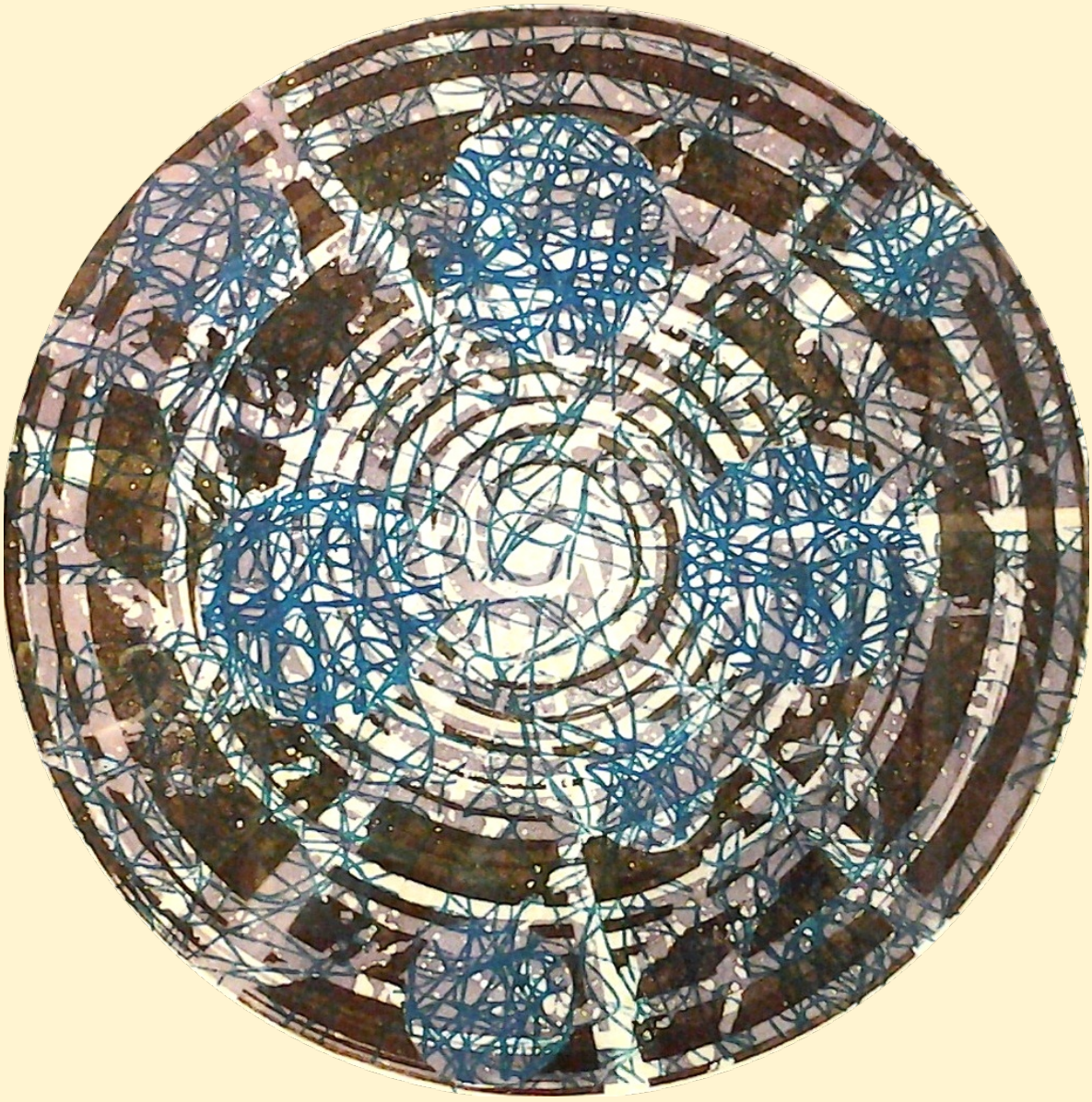
Annals of Computer Science and Information Systems
Volume 15

Proceedings of the 2018 Federated Conference on Computer Science and Information Systems

September 9–12, 2018. Poznań, Poland



Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)

PTI

IEEE

Annals of Computer Science and Information Systems, Volume 15

Proceedings of the 2018 Federated Conference on Computer Science and Information Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki (eds.)

2018, Warszawa,
Polskie Towarzystwo
Informatyczne

2018, New York City,
Institute of Electrical and
Electronics Engineers

Annals of Computer Science and Information Systems, Volume 15
Proceedings of the 2018 Federated Conference on Computer Science and Information Systems

ART: ISBN 978-83-949419-7-0, IEEE Catalog Number CFP1885N-ART
USB: ISBN 978-83-949419-6-3, IEEE Catalog Number CFP1885N-USB
WEB: ISBN 978-83-949419-5-6

ISSN 2300-5963
DOI 10.15439/978-83-949419-5-6

Ul. Solec 38/103
00-394 Warsaw, Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover art: Mandala
Bogdan Kiliński,
Elbląg, Poland

Also in this series:

Volume 17: Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-952357-0-2, ISBN USB: 978-83-952357-1-9**
Volume 16: Position Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-949419-8-7, ISBN USB: 978-83-949419-9-4**
Volume 14: Proceedings of the First International Conference on Information Technology and Knowledge Management, **ISBN WEB: 978-83-949419-2-5, ISBN USB: 978-83-949419-1-8, ISBN ART: 978-83-949419-0-1**
Volume 13: Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-2-0, ISBN USB: 978-83-922646-3-7**
Volume 12: Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-922646-0-6, ISBN USB: 978-83-922646-1-3**
Volume 11: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-946253-7-5, ISBN USB: 978-83-946253-8-2, ISBN ART: 978-83-946253-9-9**
Volume 10: Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-65750-05-1, ISBN USB: 978-83-65750-06-8**
Volume 9: Position Papers of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-93-4, ISBN USB: 978-83-60810-94-1**
Volume 8: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-90-3, ISBN USB: 978-83-60810-91-0, ISBN ART: 978-83-60910-92-7**
Volume 7: Proceedings of the LQMR Workshop, **ISBN WEB: 978-83-60810-78-1, ISBN USB: 978-83-60810-79-8**
Volume 6: Position Papers of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-76-7, ISBN USB: 978-83-60810-77-4**
Volume 5: Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-66-8, ISBN USB: 978-83-60810-67-5**

DEAR Reader, it is our pleasure to present to you Proceedings of the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Poznań, Poland, on September 9-12, 2018.

FedCSIS 2018 was Chaired by prof. Krzysztof Jassem, while dr. Paweł Skórzewski acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by the Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute Polish Academy of Sciences, Warsaw University of Technology, Wrocław University of Economics, and Adam Mickiewicz University.

FedCSIS 2018 was technically co-sponsored by: IEEE Region 8, IEEE Poland Section, IEEE Computer Society Technical Committee on Intelligent Informatics, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Gdańsk Computer Society Chapter, SMC Technical Committee on Computational Collective Intelligence, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Control System Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, ACM Special Interest Group on Applied Computing, International Federation for Information Processing, Committee of Computer Science of the Polish Academy of Sciences, Polish Operational and Systems Research Society, Mazovia Cluster ICT Poland and Eastern Cluster ICT Poland. FedCSIS 2018 was sponsored by Intel, Gambit, Samsung, Silver Bullet Labs, eSensei and Data Center PPNT.

During FedCSIS 2018, keynote lectures have been delivered by:

- Aksit, Mehmet, University of Twente, "*The Role of Computer Science and Software Technology in Organizing Universities for Industry 4.0 and Beyond*"
- Bosch, Jan, Chalmers University Technology, "*Towards a Digital Business Operating System*"
- Duch, Włodzisław, Nicolaus Copernicus University, "*Neurocognitive informatics for understanding brain functions*"
- O'Connor, Rory, V., Dublin City University, "*Demystifying the World of ICT Standardisation: An Insiders Viewpoint*"

FedCSIS 2018 consisted of the following events (conferences, symposia, workshops, special sessions). These events were grouped into FedCSIS conference areas**,** of various degree of integration. Specifically, those listed without indication of the year 2018 signify "abstract areas" with no direct paper submissions to them (but with submissions to their enclosed events).

- **AAIA'18 – 13th International Symposium Advances in Artificial Intelligence and Applications**
  - AIMaViG'18 – 3rd International Workshop on Artificial Intelligence in Machine Vision and Graphics
  - AIMA'18 – 8th International Workshop on Artificial Intelligence in Medical Applications
  - AIRIM'18 – 3rd International Workshop on AI aspects of Reasoning, Information, and Memory
  - ASIR'18 – 8th International Workshop on Advances in Semantic Information Retrieval
  - DMGATE'18 – 1st International Workshop on AI Methods in Data Mining Challenges
  - SEN-MAS'18 – 6th International Workshop on Smart Energy Networks & Multi-Agent Systems
  - WCO'18 – 11th International Workshop on Computational Optimization
- *CSS – Computer Science & Systems*
  - BEDA'18 – 1st International Workshop on Biomedical & Health Engineering and Data Analysis
  - CANA'18 – 11th Workshop on Computer Aspects of Numerical Algorithms
  - C&SS'18 – 5th International Conference on Cryptography and Security Systems
  - CPORA'18 – 3rd Workshop on Constraint Programming and Operation Research Applications
  - LTA'18 – 3rd International Workshop on Language Technologies and Applications
  - MMAP'18 – 11th International Symposium on Multimedia Applications and Processing
- *iNetSApp – International Conference on Innovative Network Systems and Applications*
  - INSERT'18 – 2nd International Conference on Security, Privacy, and Trust
  - IoT-ECAW'18 – 2nd Workshop on Internet of Things - Enablers, Challenges and Applications
- *IT4MBS – Information Technology for Management, Business & Society*
  - AITM'18 – 15th Conference on Advanced Information Technologies for Management
  - ISM'18 – 13th Conference on Information Systems Management
  - KAM'18 – 24th onference on Knowledge Acquisition and Management
- *SSD&A – Software Systems Development & Applications*
  - MDASD'18 – 5th Workshop on Model Driven Approaches in System Development
  - MIDI'18– 6th Conference on Multimedia, Interaction, Design and Innovation
  - LASD'18 – 2nd International Conference on Lean and Agile Software Development
  - SEW-38 & IWCPS-5 – Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)
- **DS-RAIT'18 – 5th Doctoral Symposium on Recent Advances in Information Technology**

The 2018 edition of an AAIA'18 Data Mining Challenge is a continuation of the topic from the previous year – data analytics related to video games. In particular, it was focused on a popular collectible card video game *Hearthstone: Heroes of Warcraft*. Awards for the winners of the contest were sponsored by: Silver Bullet Solutions, eSensei and the Mazovia Chapter of the Polish Information Processing Society. Papers resulting from the competition constitute a separate section of these Proceedings.

Each paper, found in this volume, was refereed by at least two referees and the acceptance rate of regular full papers was ~21,2% (74 papers out of 349 general submissions).

The program of FedCSIS required a dedicated effort of many people. Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to all Committee members for

their hard work in attracting and later refereeing 349 submissions (regular and data mining).

We thank the authors of papers for their great contribution to research and practice in computing and information systems. We thank the invited speakers for sharing their knowledge and wisdom with the participants. Finally, we thank all those responsible for staging the conference in Poznań. Organizing a conference of this scope and level could only be achieved by the collaborative effort of a highly capable team taking charge of such matters as conference registration system, finances, the venue, social events, catering, handling all sorts of individual requests from the authors, preparing the conference rooms, etc.

We hope you had an inspiring conference and an unforgettable stay in the beautiful city of Poznań. We also hope to meet you again for FedCSIS 2019 in Leipzig, Germany.

Co-Chairs of the FedCSIS Conference Series

Maria Ganzha, *Warsaw University of Technology, Poland and Systems Research Institute Polish Academy of Sciences, Warsaw, Poland*

Leszek Maciaszek, *Wrocław University of Economics, Wrocław, Poland and Macquarie University, Sydney, Australia*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, Warsaw Poland and Management Academy, Warsaw, Poland*

Annals of Computer Science and Information Systems, Volume 15

Proceedings of the Federated Conference on Computer Science and Information Systems

September 9–12, 2018. Poznań, Poland

TABLE OF CONTENTS

1ST INTERNATIONAL WORKSHOP ON AI METHODS IN DATA MINING CHALLENGES

6TH INTERNATIONAL WORKSHOP ON SMART ENERGY NETWORKS & MULTI-AGENT SYSTEMS

11TH INTERNATIONAL WORKSHOP ON COMPUTATIONAL OPTIMIZATION

COMPUTER SCIENCE & SYSTEMS

1ST INTERNATIONAL WORKSHOP ON BIOMEDICAL & HEALTH ENGINEERING AND DATA ANALYSIS

11TH WORKSHOP ON COMPUTER ASPECTS OF NUMERICAL ALGORITHMS

5TH INTERNATIONAL CONFERENCE ON CRYPTOGRAPHY AND SECURITY SYSTEMS

3RD WORKSHOP ON CONSTRAINT PROGRAMMING AND OPERATION RESEARCH APPLICATIONS

3RD INTERNATIONAL WORKSHOP ON LANGUAGE TECHNOLOGIES AND APPLICATIONS

11TH INTERNATIONAL SYMPOSIUM ON MULTIMEDIA APPLICATIONS AND PROCESSING

INTERNATIONAL CONFERENCE ON INNOVATIVE NETWORK SYSTEMS AND APPLICATIONS

2ND INTERNATIONAL CONFERENCE ON SECURITY, PRIVACY, AND TRUST

2ND WORKSHOP ON INTERNET OF THINGS - ENABLERS, CHALLENGES AND APPLICATIONS

INFORMATION TECHNOLOGY FOR MANAGEMENT, BUSINESS & SOCIETY

16TH CONFERENCE ON ADVANCED INFORMATION TECHNOLOGIES FOR MANAGEMENT

13TH CONFERENCE ON INFORMATION SYSTEMS MANAGEMENT

24TH CONFERENCE ON KNOWLEDGE ACQUISITION AND MANAGEMENT

SOFTWARE SYSTEMS DEVELOPMENT & APPLICATIONS

5TH WORKSHOP ON MODEL DRIVEN APPROACHES IN SYSTEM DEVELOPMENT

6TH CONFERENCE ON MULTIMEDIA, INTERACTION, DESIGN AND INNOVATION

2ND INTERNATIONAL CONFERENCE ON LEAN AND AGILE SOFTWARE DEVELOPMENT

JOINT 38TH IEEE SOFTWARE ENGINEERING WORKSHOP (SEW-38) AND 5TH INTERNATIONAL WORKSHOP ON CYBER-PHYSICAL SYSTEMS (IWCPS-5)

5TH DOCTORAL SYMPOSIUM ON RECENT ADVANCES IN INFORMATION TECHNOLOGY

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1–4
DOI: 10.15439/2018F001
ISSN 2300-5963 ACSIS, Vol. 15

Adopting a Digital Business Operating System

Jan Bosch
Chalmers University of Technology
Department of Computer Science and Engineering
Gothenburg, Sweden
Email: jan@janbosch.com

***Abstract*—The role of software in society and in industry in particular continues to grow exponentially. Most companies either have or are in the process of adoption continuous deployment of their software at products in the field and collect data concerning the performance of their systems. The continuous, fast feedback loops that companies now have available allow for a fundamentally different way of organizing. In fact, based on our work with dozens of companies, we have come to the conclusion that companies are moving towards a new, digital operating system. In this paper, we first present the key elements of the digital operating system and then discuss some of the challenges companies experience during the transformation.**

I. Introduction

AS A POPULAR QUOTE says, in the future, all companies will be software companies. It is safe to say that this is no longer the future, but reality today. Ranging from telecommunications to automotive and from banks to retail companies, the key differentiator for virtually any company these days is it's ability to create, deploy and evolve software better than its competitors. In fact, every company these days is a software-intensive business (SIB).

One of the main focus areas for SIBs is the adoption of continuous deployment, meaning that new software is deployed in systems on a frequent basis (at least every agile sprint). As part of continuous deployment, the company often also deploys instrumentation to ensure that the systems continue to function as desired. This instrumentation generates data that not only provides information about any quality issues, but also about the value that the software is generating for customers and for the company itself. This causes the adoption of data-driven practices.

Generalizing from these observations, and based on research that we have conducted with dozens of companies over the last decade, we have concluded that industry is moving towards an new, digital business operating system. This operating system consists of four dimensions:

- **Speed**: The history of SIBs is defined by constantly increasing speed. From yearly to quarterly to continuous releases, software is deployed more frequently. The primary driver for this speed is the shortening of feedback loops. The goal is to shorten the time from making a decision to observing or measuring the effect of the decision to the shortest possible.
- **Data**: When a company has a mechanism to deploy new software to its servers or products in the field, this implies that it also is possible to get data back. Although the notion of "Big Data" is prevalent and especially online companies can be very advanced in their use of data, such as through the application of A/B/n experimentation, our research shows that many companies still make quite limited use of the available data as a resource.
- **Ecosystems**: The third dimension is concerned with the ecosystems surrounding a SIB. Although traditional companies tend to aim at performing as much of the required activities internally, modern companies focus their own resources on the activities where the company is uniquely differentiating and partner with others for everything else.
- **Empowerment**: Finally, when a SIB has a solid understanding of what it seeks to do in-house and the data to track the creation and delivery of value, the need for the traditional hierarchical organization disappears or at least is diminished radically. Instead, individuals and teams can be empowered to deliver on defined output metrics without having to be managed in conventional ways. Teams that deliver continue to thrive and teams that fail to deliver on expectations receive help and support to improve or, failing that, are disbanded.

This work was supported by Software Center (www.software-center.se).

The remainder of this paper is organized as follows. In the next section, we present each of the dimensions of the digital business operating system framework in more detail. Subsequently, we highlight some of the challenges that companies experience when adopting aspects of the framework. Finally, we conclude the paper in section IV.

II. Digital Business Operating System

As we discussed in the introduction, software intensive businesses tend to adopt aspects of a new digital business operating system framework. This framework consists of four dimensions, i.e. speed, data, ecosystems and empowerment. In the subsequent sections, we present each dimension in a more detail.

A. Speed

Companies evolve through a number of typical steps when increasing the speed of their operations. As shown in figure 1, in our research, we have identified five key steps. These steps are traditional development, agile development, continuous integration, continuous deployment and, finally, R&D as an

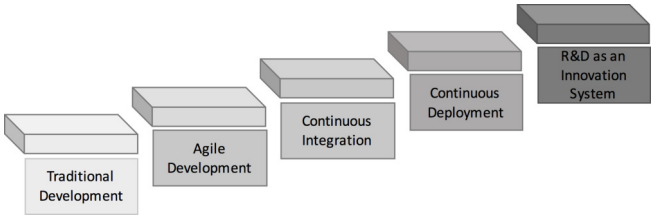


Fig. 1. The *speed* dimension

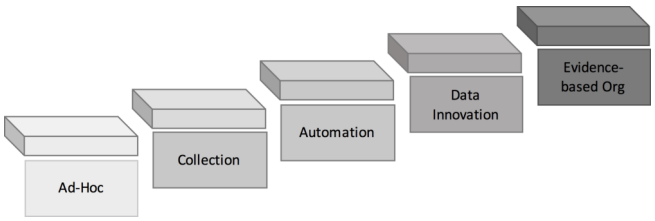


Fig. 2. The *data* dimension

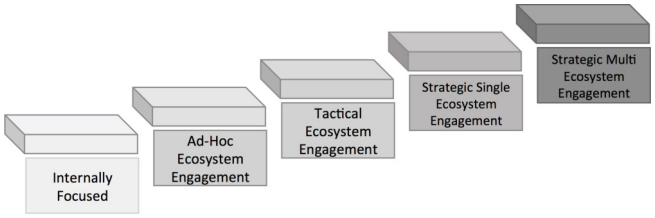


Fig. 3. The *ecosystems* dimension

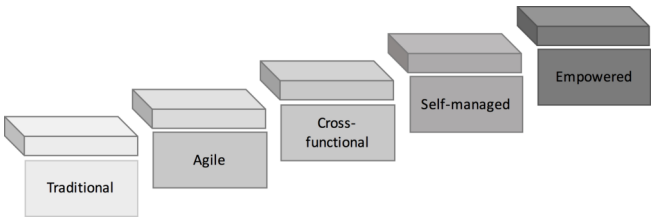


Fig. 4. The *empowerment* dimension

experiment system. All except the last term are well known. The final step is where the company starts to use its installed base and data capability to experiment with alternative ways of realizing functions and features.

For more information on the speed dimension, we refer to [5] and [4].

B. Data

The data dimension also presents the typical steps that a company evolves through when adopting data-driven practices. As shown in figure 2, companies move from working with data in an ad-hoc fashion to automating the collection to data. The next step is where the company introduces dashboards and other techniques to automatically collect, analyze and present data. The data innovation step is where the company continuously looks for novel insights in the available data in order to continuously update and improve dashboards and other outputs. The final stage is where all processes in the company are driven by data, including hiring, performance reviews and management.

For more information on the data dimension, we refer to [6], [7] and [8].

C. Ecosystems

The third dimension is concerned with the way the company works with the ecosystems surrounding it. No company is an island and by necessity interacts with suppliers, customers, complementers, competitors, regulators, governments and other stakeholders. However, many companies have a tendency to focus their energy internally and try to conduct as much as possible inside the walls of the company. Companies that have adopted the digital business operating system are much more concerned about focusing their own resources where they have the most differentiation and, consequently, can add the most value. For everything else, these companies engage their ecosystem partners.

In figure 3, the typical evolution path for a company is shown. Starting from an internal focus, the company initially engages with ecosystem partners in an ad-hoc fashion, then proceeds to a more tactical approach and over time becomes increasingly strategic about its engagements with the ecosystem.

For more information on the ecosystems dimension, we refer to [3], [10] and [2].

D. Empowerment

The final dimension of the digital business operating system addresses the organizational question. Traditionally, companies have been organized hierarchically, largely mirroring the military organizations that were one of the few examples of organizing large groups of individuals that existed at the beginning of the industrial revolution. Hierarchical organizations are effective for organizing repetitive and often manual work in low-complexity contexts, but fail in high complexity situations where creativity and non-standard solutions are required. As virtually all repetitive work these days is automated, the only work left for individuals is that which can not easily be standardized. Consequently, we need alternative ways of organizing and the common denominator is empowerment.

In figure 4, the typical evolution path for a company moving from traditional hierarchies to empowerment is shown. Starting from a hierarchical approach to organizing, agile software development provides a first step on the transition to empowerment as agile teams have significantly more freedom to operate than their counterparts working in a hierarchical waterfall approach. From there, we see companies adopt increasing cross-functional empowerment, fully self management and finally the fully empowered state.

For more information on the empowerment dimension, we refer to [9] and [1].

III. Challenges

Adopting the digital business operating system for a company that has worked traditionally is quite challenging as it affects all functions in the company. In addition, everything from business models to system architecture, from process to organization and from formal structures to company culture is affected. Consequently, the challenges that companies experience are significant and it is sometimes difficult to separate root causes from symptoms that are the consequence of root causes.

In order to provide a high-level overview of the key challenges that companies experience, we use the BAPO model [1]. The BAPO model considers business, architecture, process and organization. In the sections below, we briefly discuss the key challenges that companies experience when adopting the digital business operating system.

A. Business

Most product companies have traditionally maintained a transactional business model, meaning that customers engage with the company every couple of years or even decades. For instance, most people will buy a new car every five years and have little connection with the car company in the meantime. One of the consequences of digitalization is that it typically assumes a continuous deployment model for software. This has several consequences for the business.

The first consequence is that the business model changes from a product-centric to a service-centric model. In many industries, customers subscribe to a service in order to get access to the product, rather than buying it outright. That changes the customer relationship from a transactional to a continuous one. A second consequence is that it requires a different financial model because the company needs to finance the product for the customer and then earn back the cost through service.

B. Architecture

There are two primary changes required to the architecture when adopting a digital business operating system. First, because we're looking to support continuous deployment, it is typically required to deploy individual components independently in order to avoid significant downtime on systems in the field. This requires a significantly improved level of modularization.

Second, as we're looking to use the data from systems in the field for managing quality and for determining whether we're delivering value for customers, the architecture needs to be prepared for easy instrumentation of arbitrary parts of the system. Our research shows that companies often seek to add points of data collection and instrumentation continuously.

C. Process

Continuous deployment, as we saw in the speed dimension of the digital business operating system, requires continuous integration and test. Although many companies pride themselves on the high quality of their systems, experience shows that many companies require significant investment in build and test infrastructure before the quality of software reaches a point where it can be deployed continuously.

Second, although agile practices are widely adopted at the team level, true business agility where all functions in the company operate in short cycles and can rapidly respond to changes in the market is an elusive concept for many organizations. Product management, marketing, sales, systems engineering and general management often have a tendency to operate in yearly and quarterly cycles and convincing these functions to adopt a continuous, agile approach is a significant challenge.

D. Organization

As indicated in the discussion around the empowerment dimension, to successfully operate in this new approach, the traditional hierarchical and functionally structured company needs to be reorganized into smaller cross-functional teams. The amount of coordination overhead and delay in traditional organizations is simply too costly and slow.

An interesting observation from our research is that company culture often has significant number of implicit assumptions that are based on a waterfall style of doing business, meaning that the change towards digitalization requires a fundamental change of the company culture itself. As culture eats strategy for breakfast, as the saying goes, it requires significant effort and role modeling from leaders to instill the cultural changes required to successfully operate in this new model.

IV. Conclusion

Virtually all companies these days are software-intensive businesses. Across domains, such as telecommunications, automotive, finance, defense and retail, the ability to create, deploy and evolve better software faster than competitors is the key differentiator.

Our research that we have conducted with dozens of companies over the last decade has led us to the conclusion that industry is moving towards an new, digital business operating system. This operating system consists of four dimensions:

- **Speed**: The history of SIBs is defined by constantly increasing speed. From yearly to quarterly to continuous releases, software is deployed more frequently. The primary driver for this speed is the shortening of feedback loops. The goal is to shorten the time from making a decision to observing or measuring the effect of the decision to the shortest possible.
- **Data**: When a company has a mechanism to deploy new software to its servers or products in the field, this implies that it also is possible to get data back. Although the notion of "Big Data" is prevalent and especially online companies can be very advanced in their use of data, such as through the application of A/B/n experimentation, our research shows that many companies still make quite limited use of the available data as a resource.

- **Ecosystems**: The third dimension is concerned with the ecosystems surrounding a SIB. Although traditional companies tend to aim to performing as much of the required activities internally, modern companies focus their own resources on the activities where the company is uniquely differentiating and partner with others for everything else.
- **Empowerment**: Finally, when a SIB has a solid understanding of what it seeks to do in-house and the data to track the creation and delivery of value, the need for the traditional hierarchical organization disappears or at least is diminished radically. Instead individuals and teams can be empowered to deliver on defined output metrics without having to managed in a conventional way. Teams that deliver continue to thrive and teams that fail to deliver on expectations receive help and support to improve or, failing that, are disbanded.

The transition from a traditional mode of operating to this new operating system is far from trivial and in section III we discussed some of the key isseus that SIBs experience. These challenges range from business and business model to architecture, process, organization as well as the norms, beliefs and culture of the company.

In the context of Software Center, we continue to study the challenge of Digitalization with the partner companies and help accelerate the adoption of the digital business operating system at these and other companies.

ACKNOWLEDGMENT

I would like to thank the companies involved in the Software Center as well as other companies that I have interacted with during the last years.

REFERENCES

[1] Jan Bosch. *Speed, data, and ecosystems: Excelling in a software-driven world*. CRC Press, 2017.

[2] Jan Bosch and Helena Holmström Olsson. Ecosystem traps and where to find them. *Journal of Software: Evolution and Process*, page e1961, 2018.

[3] Petra M Bosch-Sijtsema and Jan Bosch. Plays nice with others? Multiple ecosystems, various roles and divergent engagement models. *Technology Analysis & Strategic Management*, 27(8):960–974, 2015.

[4] Helena Olsson, Anna Sandberg, Jan Bosch, and Hiva Alahyari. Scale and responsiveness in large-scale software development. *IEEE Software*, 31(5):87–93, 2014.

[5] Helena Holmström Olsson, Hiva Alahyari, and Jan Bosch. Climbing the" stairway to heaven"–a mulitiple-case study exploring barriers in the transition from agile development towards continuous deployment of software. In *2012 38th Euromicro Conference on Software Engineering and Advanced Applications*, pages 392–399. IEEE, 2012.

[6] Helena Holmström Olsson and Jan Bosch. From Opinions to Data-Driven Software R&D: A Multi-case Study on How to Close the "Open Loop" Problem. In *40th EUROMICRO Conference on Software Engineering and Advanced Applications*, pages 9–16. IEEE, 2014.

[7] Helena Holmström Olsson and Jan Bosch. The HYPEX Model: From Opinions to Data-Driven Software Development. In *Continuous Software Engineering*, pages 155–164. Springer, 2014.

[8] Helena Holmström Olsson and Jan Bosch. Towards continuous customer validation: A conceptual model for combining qualitative customer feedback with quantitative customer observation. In *International Conference of Software Business*, pages 154–166. Springer, 2015.

[9] Helena Holmström Olsson and Jan Bosch. No more bosses? In *International Conference on Product-Focused Software Process Improvement*, pages 86–101. Springer, 2016.

[10] Helena Holmström Olsson and Jan Bosch. From ad-hoc towards strategic ecosystem management: the three-layer ecosystem strategy model. *Journal of Software Evolution and Process*, 29(7):e1876, 2017.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 5–11
DOI: 10.15439/2018F002
ISSN 2300-5963 ACSIS, Vol. 15

The Role of Computer Science and Software Technology in Organizing Universities for Industry 4.0 and Beyond

Mehmet Akşit
Software Technology
Formal Methods and Tools Group
University of Twente, Enschede,
The Netherlands
Email: m.aksit@utwente.nl

***Abstract*—Based on intensive cooperation with four large companies, a comparative analysis of the recent developments in industry, university organizations, computer science and software technology is presented. Within this context, also the Industry 4.0 phenomena is discussed. This paper further identifies the necessary organizational structures of universities to assist companies in their transition processes, defines the relevant sub-disciplines in computer science and finally describes the software engineering and technology challenges in designing and implementing economical and robust industrial systems.**

I. Introduction

INDUSTRY 4.0 aims at increasing the efficiency and effectiveness of manufacturing processes with the help of large-scale computerization. It is stated that Industry 4.0 is the 4th industrial revolution in the history of manufacturing. Publications on Industry 4.0 generally try to set-up a conceptual framework to explain what the Industrial 4.0 phenomena is. There are also some ongoing efforts on structuring universities and knowledge institutes so that their research and education activities can seamlessly support this upcoming transition towards Industry 4.0. We consider Industry 4.0 as a natural development in continuous transition from traditional to modern manufacturing processes. Since transitions occur with the help of technology, organizational structure of universities are crucial in accomplishing the Industry 4.0 objectives. The role of computer science and software technology is undisputable, since the success of Industry 4.0 largely depends on effectiveness and efficiency of computing systems. Based on our experience with four large high-technology companies, this paper introduces the attributes of the phenomena Industry 4.0 and beyond, depicts the required organizational structures of universities to assist companies in their transition processes, identifies the relevant sub-disciplines in computer science and finally describes the software engineering and technology challenges in designing and implementing economical and robust high-quality Industry 4.0 systems.

II. Industry 4.0 and Beyond

The term Industry 4.0 refers to computerization of manufacturing processes. There are four principle scenarios in Industry 4.0 [1]:

- **Interoperability** meaning that sensors, devices, machines, and people can connect and exchange information with each other.
- **Information transparency** meaning that a rich set of data can be gathered from various sources.
- **Technical assistance** meaning that machines, systems, processes, human beings, etc. can be intelligently and effectively assisted to monitor, control and optimize the overall manufacturing process.
- **Decentralized decisions** meaning that subsystems can autonomously take decisions where possible.

It is claimed that Industry 4.0 is the 4th industrial revolution in the history of manufacturing [2]. With Industry 4.0, it is expected that machines and systems will become more self-aware and self-learning so that their effectiveness and maintenance can be improved. In addition, due to networked data gathering and intelligent and autonomous process control, the manufacturing processes will be much more efficient and effective than traditional manufacturing processes.

One criticism to these claims is that industrial innovation is continuous and as such one cannot talk about a revolution [3]. Moreover, although there are some attempts to define the technical implications of Industry 4.0, it seems that Industry 4.0 touches to a large set of disciplines from sensors, industrial manufacturing to computer science and software technology (CS-ST). In particular, almost all disciplines of CS-ST are relevant for Industry 4.0.

We think that the concepts relevant to Industry 4.0 must be defined and understood in the process of on-going transition from traditional to modern manufacturing processes. It is important to stress that such transitions are not abrupt in nature but gradual, depending on the characteristics of manufacturing, technological and societal progresses. In the following we will make an attempt to compare traditional and modern manufacturing processes with each other.

Traditionally, the term industrial manufacturing referred to labor-intensive factories with specialized product portfolio. Production processes and products to be manufactured had to be predefined precisely. Process and product control and optimization activities were carried out in each phase of production separately. Due to advent of new technologies and changes in social structures, however, there have been

continuous changes to the ways how industrial manufacturing is realized. For example, during the last two to three decades, the following transitions have been observed in highly industrialized countries:

1. Knowledge intensive manufacturing **instead of** labor/ resource-intensive manufacturing. There is a greater dependence on intellectual capabilities than on physical inputs or natural resources [4].
2. Focus on owning and managing knowledge and skills and intellectual property rights **instead of** focusing on labor/resource intensive manufacturing processes [5]. When planned carefully and in certain circumstances, it may be much more profitable to outsource some activities and manufacturing processes [6].
3. Dynamically managed and optimized, multi-asset portfolio **instead of** fixed/ad-hoc, single-asset portfolio. The advantages are risks reduction, controlled risk taking, capital preservation and enhanced returns [7][8].
4. Mass customization **instead of** mass production. Mass customization is the automated manufacturing of tailored products. It has the combined advantages of the low unit costs of mass production with the flexibility of building products for more customer satisfaction [9].
5. Proactive self-organizing companies **instead of** inflexible hierarchically-organized companies. Such new organizational structures aim at effectively responding to changing markets and business contexts [10].
6. End-to-end alignment and optimization of (manufacturing) processes **instead of** focusing only on the improvement of individual phases. This improves companies due to enhancement of the whole supply-chain [11].
7. Multi-disciplinary usage of teamed personnel **instead of** working with solely operating individuals. Teaming helps organizations in continuous improvement, understanding complex systems, and in successful innovation [12].
8. Organizing businesses/enterprises globally through networks **instead of** isolated and/or localized organizations. This is an increasing necessity for any entrepreneurship and value creation [13]-[15].
9. Improved time-to-market **instead of** long sequences of research, design, manufacturing and marketing phases [16].
10. Intensive use of state-of-art CS-ST as the "main enabler" of modern businesses **instead of** considering CS-ST just like any other technical skill. CS-ST is essential in fulfilling the requirements of modern businesses, such as described in [9][11][13]-[16].
11. Strong cooperation with universities for the purpose of innovation **instead of** considering universities mainly as theoretical institutions that educate people [17]-[19].

From a rather narrow perspective, the term Industry 4.0 refers to autonomous cyber physical systems. Transitions that we observe in industry and formulated in 11 items in this section give a more comprehensive picture of this phenomena. They also refer to changes in business, manufacturing, marketing, organizational and industry-university cooperation processes. The related technical challenges will be discussed in the following sections of this paper. It is clear from these items that Industry 4.0 refers to (or a new name of) a part of a natural transition in manufacturing processes which has been taking place since several decades. We therefore term this list as "a list of attributes of manufacturing processes for Industry 4.0 and beyond".

III. Organizing Universities for Industry 4.0 and Beyond

Traditional universities have contributed to industries by educating engineers, applied mathematicians, administrative personnel, managers, etc. Within this context, two main categories of activities have been considered essential:

- **Research**, where academic personnel of the university are expected to be expert in certain fields. The selection of the topic of a field is not necessarily derived from industrial and societal needs; it can be ad hoc. The expertise is quite specific and theoretical. The excellence is measured according to number of publications in certain pre-classified journals.
- **Education**, where academic personnel of the university are expected to give lectures in their fields of expertise and examine the students by appropriate tests. In addition, students are expected to be supervised in writing their theses.

We believe that traditional universities with these characteristics cannot fulfil the requirements as demanded by modern industrial manufacturing processes and businesses as formulated in the previous section. Moreover, classical education methods, like long lectures followed by classical examinations cannot give the necessary education baggage to the students as desired. As such, we think that the following strategical and tactical changes are required:

1. Excellence in knowledge and in-depth specialization for academic personnel are still required. However, **the specializations must be derived from the needs of the targeted society and industry** instead of ad-hoc selection of topics; along this line, the topics must be synthesized through the scope of industrial and societal mid-term and long-term objectives. Otherwise, universities cannot be equipped with the necessary expertise in supporting companies in their innovative processes.

2. **The academic personnel must learn to work together in multi-disciplinary teams**. For example, theory-oriented persons must be able to work with practically-oriented-persons, and vice versa, different experts in the same discipline or among different disciplines, must cooperate together in university-industry joint projects. Otherwise, the complex problems of industry and society cannot be addressed effectively.
3. **The academic personnel must be proactive in forming networks** to cooperate with national and international institutions and colleagues not only from his/her own discipline but also from other disciplines. This is necessary to share expertise, to jointly define the desired research agenda, and to find solutions to complex industrial and societal problems.
4. **The academic personnel must be flexible enough to adapt themselves** in changing demands from industry and society. Otherwise, in due time, the expertise of academic personnel can be outdated or become less relevant.
5. The **education process must be tailored** to answer the mid-term and long-term needs of industry and society:

    - It must focus on the core concepts **instead of** hypes.
    - It must focus on gaining analytical skills, critical thinking and reasoning **instead of** memorizing what are in the books.
    - It must aim at teaching problem solving/synthesis **instead of** gaining knowledge which cannot be utilized for solving actual problems.
    - It must emphasize working in multi-disciplinary projects **instead of** only focusing on mono-disciplinary exercises.
    - It must enhance communication skills, such as oral and written presentation and argumentation skills **instead of** educating students with non-communicative and introvert attitude.
    - It must aim at increasing consciousness of students in ethical concerns **instead of** educating students with irresponsible and/or indifferent attitude.

6. The university must create **suitable organizational structures** to enable the academic personnel efficiently and effectively fulfil the objectives listed above. These include:

    - **Proactive and self-adaptive organization** to support the objectives of the university in dynamically changing contexts.
    - Organization to set-up and carry-out **multi-disciplinary projects** for industry and society.
    - Organization with an **award system to motivate** the academic personnel and students along the objectives of the university.
    - Organization which **emphasizes CS-ST** since it is the “main enabler” of all disciplines at the university.

To derive the required specializations within universities the following activities can be carried out:

1. **Understanding the context** of the university.
2. **Defining the strategic needs** of companies and society.
3. **Analyzing the current structure** of the university.
4. **Identifying the strong and weak points** of the university.
5. Based on the observations of the future trends as stated in the paper, **formulate a transition plan**.

These steps look quite obvious but due to involvement of many stakeholders, they are harder to implement than one may expect. While realizing the transition, the following quality attributes can be considered:

- **Relevancy**: The university must be highly relevant in addressing technical and social needs. To this aim, research, education and organization activities must be defined in close cooperation with the relevant companies and societal organizations.
- **Alignment with the current state-of-the-art research**: The research and education activities to be carried out must advance the state-of-the-art so that the companies and businesses can be matured to be the leaders in their context.
- **Cross-fertilization**: Different university research and education activities can benefit from each other. To maximize the benefit, it is important to strongly coordinate the related activities with each other.
- **Industry-as-laboratory**: To identify the relevant problems and to test the proposed solutions, it is important that the principle investigators and the affiliated (Ph.D. and/or M.Sc., etc.) students visit the companies regularly and carry out experiments within industrial and societal context. To this aim, companies must provide personal assistance and industrial case studies.
- **Academic research steering committee**: To coordinate the activities effectively and efficiently, it is important to mentor the students and monitor the progress of research and education activities and evaluate them with respect to the desired objectives. To this aim, academic research steering committees can be established where all the relevant stakeholders participate.

There have been also some attempts to classify universities according to their contributions to industry and society. To this aim, the concept of University 4.0 has been introduced. In [20], University 4.0 is defined as "an university which is outward looking, deeply connected to industry and the communities around it, and committed to serving the needs of its students". The definition of University 4.0 is largely consistent with our observations about the necessary changes of university organizations as presented in this paper. However, we consider evolution of industries and as well as universities as continuous processes which influence each other. Although correlated, it looks like that the developments around Industry 4.0 and University 4.0 currently are not structurally related. That is, both developments can actually be viewed and realized independent of each other. Our focus in this paper is more from technological perspective, which needs to be supported by dedicated methods. We consider discussions around University 4.0 is useful, but the conceptualization of this terminology is still in a premature state.

IV. The Developments in Computer Science and Software Technology

CS-ST is the main force in almost all industries; it creates added value for products and businesses. There is almost no product in the market which does not contain software or is not produced by a process controlled by software. To accomplish the objectives of Industry 4.0 and beyond, advanced CS-ST is needed.

There are all kinds of hypes over CS-ST in the popular media. Nevertheless, the recent developments in CS-ST are more or less shaped around the following topics:

1. Large infrastructures, service-oriented architectures, cloud computing, systems-of-systems, ecosystems [21]-[23].
2. Sensors, Internet of Things (IOT), and pervasive computing [24]-[26].
3. Big data and big data analytics [27]-[29].
4. Security and cybersecurity [30][31].
5. Cyber-physical systems [32].
6. Artificial intelligence and related topics including computational intelligence, machine learning and multi-agent systems [33]-[36].
7. Graphical processing, visualization and human-machine interaction including virtual reality [37][38].
8. High performance, and/or multi-core/parallel architectures including parallel programming [39].
9. Theoretical and practical work on algorithms and/or constraint-based "solvers" [40][41] to address a large category of mathematical problems. In general algorithms/solvers are applied to every category of computer science specializations listed in this section.
10. Software (engineering) methods and techniques [42] to fulfil the functional and qualitative requirements of software systems. The concepts of software engineering can be applied to every computer science specialization listed in this section.

V. The Role of Software Engineering Methods and Techniques

Economical, sustainable and robust software systems which fulfil functional and qualitative requirements are essential for all software systems. To accomplish the requirements of Industry 4.0 and beyond, software engineering methods and techniques are crucial. **No matter how intelligent a software solution is, if it cannot be realized with the desired quality attributes, one cannot expect an economical value out of it**. As such software engineering methods and techniques can be defined as crosscutting (meta-level) concerns that relate to all developments within CS-ST. In addition, many recent developments in computer science are more and more utilized within software engineering methods and techniques. The trends in software engineering methods and techniques therefore cannot be considered separately from the recent developments in computer science. For example, big data analytics and machine learning techniques are increasingly used to tune and optimize software engineering models and methods, cloud architectures and ecosystems are becoming part of software development environments, visualization techniques are used to detect anomalies in software architecture, etc.

After 4 years of intensive cooperation with high-tech industry, for example, we have identified the following trends [19]:

1. Product-line **instead of** product design. Most products are developed and manufactured by specialized companies, which market families of products. It is not economical to develop each product from scratch [43][44].
2. Systems of systems **instead of** systems perspective. Software systems for Industry 4.0 are generally adopted in large distributed settings. Scale-ability and interoperability of systems are essential. Systems of systems architectures, are therefore the natural candidates of the platforms of Industry 4.0 architectures [23].
3. Ecosystem design **instead of** platform design. Software ecosystems are an effective and economical way to construct large software systems for Industry 4.0 on top of a software platform by adding up software modules developed by different actors. In ecosystem design software engineering is spread outside the traditional borders of software companies to a group of companies and private persons [45].
4. Auto-adaptive control architectures **instead of** architectures without any control mechanisms. To realize the monitoring and controlling activities in Industry 4.0 and to cope with the changing requirements and context, software systems are expected to be more re-

active and self-adaptive. This generally requires built-in feedback control mechanisms in software. Self-adaptation can be realized at system level, subsystem level and/or at component-level. In addition, different styles can be adopted, such as single, master-slave, hierarchical and/or peer-to-peer control architectures [46][47].

5. Distributed problem solving including distributed algorithms, coordinating systems and multi-agent architectures **instead of** centralized problem solving with monolithic and/or localized architectures. Since computer systems for Industry 4.0 are distributed, to reduce complexity and enhance reliability, algorithms and intelligence in systems must be distributed as well. Accordingly, programming languages and techniques must adequately support distributed programming efforts by offering expressive and flexible abstractions [48]-[52].
6. Model-based development **instead of** straight-forward programming. Since more and more companies are specialized in certain product categories and in manufacturing processes, deriving software architecture from relevant domain models can help in reducing complexity, enhancing reuse and testability/verifiability of software systems. Model-based development has been adopted in various approaches such as product-line engineering (SPLE), model-driven engineering (MDE), domain specific architectures (DSA) and domain-specific programming languages (DSL), model-based verification (MBV), etc. [53]-[56].
7. Multi-objective optimization **instead of** ad-hoc handcrafted and/or single objective optimization. Along the line of model-based development, various algorithmic techniques and search-based methods have been introduced to compute the "optimal" architectural decomposition with respect to certain quality attributes. In addition, various run-time optimization techniques can be adopted in computing optimal control strategies and scheduling processes [57]-[59].
8. Modularization of semantic concerns **instead of** traditional abstraction mechanisms based on implementation concerns such as data or function. As a consequence of model-based development, software abstractions more and more correspond to the concerns of models. The concerns of a model are naturally based on the semantics of the model, and these cannot always be effectively represented as a data or function. Moreover, concerns in Industry 4.0 systems can be emerging meaning that they may appear or disappear dynamically. As such, programming languages and techniques must adequately support programming efforts by offering expressive and flexible abstractions for emergent semantic concerns [60] [61].
9. A rich set of composition mechanisms **instead of** a fixed set of language constructs for hierarchical organization of programs (such as class-inheritance). To support flexibility in control strategies and to cope with various evolution schemes, languages must offer generic and/or domain specific composition mechanisms to express, for example, object, aspect and event compositions and transformational techniques in a uniform manner. The languages must maintain their closure property in compositions so that scaleability of systems can be provided [62]-[67].
10. Uniform integration of verification techniques **instead of** independent tool- and technique-specific verification approaches. There are various model-based verification and testing approaches available. Examples are model-checking, static and dynamic analysis, run-time verification, model-based testing, adopting model-specific verification (simultaneously) based on continuous and/or discrete models, etc. Most of these techniques are complementary and as such combined usage of these may help in finding faults with less false-positive and false-negative cases [62][67]-[69].

VI. Conclusions

Industrial manufacturing today differs considerably from the past and there is a strong evidence that this trend will also continue in the future. The Industry 4.0 phenomena should be considered in this context. The main force behind this chance is the continuous evolution in CS-ST. Naturally, universities are indispensable elements of this progress. It is therefore important to carefully monitor and comparatively understand the recent developments in industry, and accordingly understand the impact of the trends in university organizations, computer science and software technologies.

References

[1] M. Hermann et al. "Design Scenarios for Industrie 4.0 Scenarios," in *Proc. 49th Hawaii International Conference on System Sciences*, 2016, pp. 3928-3937.

[2] S. Vaidya et al. "Industry 4.0 – A Glimpse". in *Proc. of International Conference on Materials Manufacturing and Design Engineering*, 2018, pp. 233-238.

[3] E. Garbee. "This Is Not the Fourth Industrial Revolution", *–via Slate at http://www.slate.com/articles/technology/future_tense/2016/01/the_world_economic_forum_is_wrong_this_isn_t_the_fourth_industrial_revolution.html*.

[4] W. Powell and K. Snellman, "The Knowledge Economy", *Annu. Rev. Sociol. 2004. doi: 10.1146/annurev.soc.29.010202.100037*, 2004, pp: 199–220.

[5] D. Modic and N. Damij, *Towards Intellectual Property Rights Management: Back-office and Front-office Perspectives*, Springer International Publishing AG, 2018.

[6] S. Cullen and M. Lacity, *Outsourcing- All You Need To Know*, White Plume Publishing, 2014.

[7] P. Sironi, *Modern Portfolio Management: From Markowitz to Probabilistic Scenario Optimisation*, Risk Books, 2015.

[8] Y. Lustig, *Multi-Asset Investing: A practical Guide to Modern Portfolio Management*, Harriman House Ltd., 2013.

[9] H. Kull, *Mass Customization: Opportunities, Methods, and Challenges for Manufacturers*, Springer Science Business Media, 2015.

[10] B. Robertson, *Holacracy: The Revolutionary Management System that Abolishes Hierarchy*, Henry Holt and Company LCC, 2016.

[11] Oracle, *Why End-to-end Visibility is Key to a Modern manufacturing Process.*

[12] A. C. Edmondson, *Teaming: How Organizations Learn, Innovate, and Compete in the Knowledge Economy*, Wiley, 2014.

[13] D. Sherman, *Maximum Success with LinkedIn: Dominate Your Market, Build a Global Brand, and Create the Career of Your Dreams*, McGraw-Hill Education Books, 2014.

[14] S. Thomas, *Instant Networking: The Simple Way to Build Your Business Network and See Results in Just 6 Months*, John Willey and Sons Ltd., 2016.

[15] E. Kaynak, R. Ajami, and M. Bear (Eds.), *The Global Enterprise: Entrepreneurship and Value Creation*, International Business Press, 2012.

[16] P. Smith, "Accelerated Product Development: Techniques and Traps", *in the PMDA Handbook of New Product Development. Second Edition*, K. Kahn (Ed.), John Willey and Sons Inc., 2004, ch. 12.

[17] C. Mascarenhas, J. Ferreira, and C. Marques, *University–industry Cooperation: A Systematic Literature Review and Research Agenda*, Science and Public Policy, Oxford Academic, 2018, pp. 1–11.

[18] M. Akşit, B. Tekinerdogan, H. Sözer, H. F. Safi and M. Ayas, "The DESARC Method: An Effective Approach for University-Industry Cooperation", in *Proc. of the International Conference on Advances in Computing, Control and Networking, ACCN 2015*, 2015, pp. 51-53.

[19] M. Akşit, B. Tekinerdogan, H. Sözer, H. F. Safi and M. Ayas, "Identifying the Research Needs of Four Large High-Technology Companies", in *Proc. ACCN 2016*, 2016, pp. 21-24.

[20] J. Dewar, "University 4.0: Redefining the Role of Universities in the Modern Era", *in Higher Education Review Magazine*, August 2017.

[21] K. M. Dhara, M. Dharmala, and C. K. Sharma, *A Survey Paper on Service Oriented Architecture Approach and Modern Web Services*, All Capstone Projects, http://opus.govst.edu/capstones/157, 2015.

[22] D. C. Marinescu, *Cloud Computing: Theory and Practice*, Morgan Kaufmann, 2017.

[23] DoD, *Systems Engineering Guide for Systems of Systems*, version 1.0, 2008.

[24] G. Ferrari (Ed.), *Sensor Networks Where Theory Meets Practice*, Springer, 2010.

[25] R. Buyya, and A. V. Dastjerdi (Eds.), *Internet of Things Principles and Paradigms*, Elsevier, 2016.

[26] N. Silvis-Cividjian, *Pervasive Computing Engineering Smart Systems*, Springer, 2017.

[27] H. Mohanty et al., *Big Data A Primer.* Springer, 2015.

[28] S. Pyne et al. (Eds.), *Big Data Analytics Methods and Applications*, Springer, 2016.

[29] P. Pääkkönen, and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", *Big Data Research*, Elsevier, vol. 2, no. 4, December 2015, pp. 166-186.

[30] L. Thames and D. Schaefer, *Cybersecurity for Industry 4.0*, Springer, 2017.

[31] M. Lehto and P. Neittaanmäki (Eds.), *Cyber Security: Analytics, Technology and Automation*, Springer, 2015.

[32] S. C. Suh, et al. *Applied Cyber-Physical Systems*, Springer, 2014.

[33] Y. Laalaoui, and N. Bouguila (Eds.), *Artificial Intelligence Applications in Information and Communication Technologies*. Springer, (2015.

[34] R. Kruse et al. *Computational Intelligence A methodological Introduction*, Springer, 2016.

[35] I. Goodfellow et al., *Deep Learning*, MIT Press, 2016.

[36] M. Hadzic et al., *Ontology-Based Multi-Agent Systems*, Springer, 2009.

[37] C. Ware, *Information Visualization Perception for Design*, Elsevier ScienceDirect, 2012.

[38] F. R. Leta (Ed.), *Visual Computing*, Springer, 2014.

[39] T. Rauber and G. Rünger, *Parallel Programming for Multicore and Cluster Systems*, Springer, 2013.

[40] B. Vöcking et al., *Algorithms Unplugged*, Springer, 2011.

[41] F. Rossi et al., "Handbook of Constraint Programming", in *Foundations of Artificial Intelligence*, vol. 2, Elsevier, 2006.

[42] I. Sommerville, *Software Engineering*, 10th Edition, Pearson, 2016.

[43] F. van der Linden et al. *Software Product Lines in Action*, Springer, 2007.

[44] G. Orhan, M. Akşit and A. Rensink, "A Formal Product-Line Engineering Approach for Schedulers", in *Proc. 22nd International Conference on Emerging Trends and Technologies in Convergence Solutions*, L. Jololian, D. E. Robbins and S. L. Fernandes (Eds.), Nov 2017, pp. 15-30.

[45] K Manikas and K. M. Hansen, "Software Ecosystems – A Systematic Literature Review", *The Journal of Systems and Software,* 86, 2013, pp. 1294–1306.

[46] S. Kounev et al., *Self-Aware Computing Systems*, Springer, 2017.

[47] M. Akşit and Z. Choukair, "Dynamic Adaptive and Reconfigurable Systems Overview and Prospective Vision", in *Proc. Workshop on Distributed Auto-adaptive Reconfigurable Systems (DARES)* - International Conference on Distributed Computing Systems (ICDCS), May 2003, Rhode Island, Providence, USA, pp. 84-89.

[48] M. Raynal, *Distributed Algorithms for Message-Passing Systems*, Springer, 2013.

[49] M. Akşit and L. Bergmans, "Guidelines for Identifying Obstacles When Composing Distributed Systems from Components", *in Software Architectures and Component Technology*, The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Dordrecht, 2002, pp. 29-56.

[50] M. Akşit, K. Wakita, J. Bosch, L. Bergmans and A. Yonezawa, "Abstracting Object Interactions Using Composition Filters", *in Object-Based Distributed Processing*, Lecture Notes in Computer Science 791, Springer Verlag, 1993, pp. 152-184.

[51] D. Shuai and X. Feng, "Distributed Problem Solving in Multiagent Systems: a Spring Net Approach", *IEEE Intelligent Systems*, vol. 20, no. 4, 2005, pp. 66 – 74.

[52] M. Hadzij et al., *Ontology-Based Multi-Agent Systems*, Springer, 2009.

[53] M. Akşit, B. Tekinerdogan, F. Marcelloni, and L. Bergmans, "Deriving Object-Oriented Frameworks from Domain Knowledge", *in Building Application Frameworks: Object Oriented Foundations of Framework Design*, John Wiley & Sons, 1999, pp. 169-198.

[54] J. Z. Pan et al., *Ontology-Driven Software Development*, Springer, 2013.

[55] Reinhartz-Berger et al., *Domain Engineering Product Lines*, *Languages, and Conceptual Models,* Springer, 2013.

[56] A. R. Da Silva, "Model-Driven Engineering: A Survey Supported by the Unified Conceptual Model", *Elsevier Computer Languages, Systems and Structures,* 43, 2015, pp. 139-155.

[57] A.J. de Roo, H. Sözer, L. Bergmans, M. Akşit, "MOO: An Architectural Framework for Runtime Optimization of Multiple System Objectives in Embedded Control Software", *Journal of Systems and Software*, 86 (10), 2013, pp. 2502-2519.

[58] H. Sözer, B. Tekinerdogan, M. Akşit, "Optimizing Decomposition of Software Architecture for Local Recovery", *Software Quality Journal*, 21 (2) 2013, pp. 203-240.

[59] M. Harman et al., "Search-Based Software Engineering: Trends, Techniques and Applications", *ACM Computing Surveys*, vol. 45, Issue 1, Article no. 11, 2012.

[60] S. Malakuti Khah Olun Abadi, and M. Akşit, "On Liberating Programs from the Von Neumann Architecture via Event-based Modularization". in *Companion Proc. of the 14th International Conference on Modularity*, New York: Association for Computing Machinery (ACM), 2015, pp. 31-34.

[61] S. Malakuti Khah Olun Abadi and M. Akşit, "Emergent Gummy Modules: Modular Representation of Emergent Behavior", in *Proc. of the 2014 International Conference on Generative Programming: Concepts and Experiences (GPCE)*, 2014, pp. 15-24.

[62] de Roo, A.J. and Sözer, H. and Akşit, M. (2014) Composing domain-specific physical models with general-purpose software modules in embedded control software. Software and Systems Modeling, 13 (1). pp. 55-81.

[63] L. Bergmans, W. Havinga, and M. Akşit, "First-Class Compositions--Defining and Composing Object and Aspect Compositions with First-Class Operators", *Transactions on Aspect-Oriented Software Development*, IX. 2012, pp. 216-267.

[64] S. Malakuti Khah Olun Abadi and M. Akşit, "Evolution of Composition Filters to Event Composition", in *Proc. 27th ACM Symposium on Applied Computing (SAC 2012)*, 2012, pp. 26-30.

[65] L. Bergmans and M. Akşit, "Composing Crosscutting Concerns Using Composition Filters", *Communications of the ACM*, 44 (10), 2001, pp. 51-57.

[66] T. Elrad, M. Akşit, G. Kiczales, K. Lieberherr and H. Ossher, "Discussing Aspects of Aspect-oriented Programming", *Communications of the ACM*, 44 (10), 2001, pp. 33-38.

[67] M. Akşit, "The 7 C's for Creating Living Software: A Research Perspective for Quality-oriented Software Engineering", *Turkish Journal of Electrical Engineering & Computer Sciences*, 12 (2), 2004, pp. 61-95.

[68] S. Ciraci, S. Malakuti, S. Katz and M. Akşit, "Checking the Correspondence Between UML Models and Implementation", in *Proc. of the 1st International Conference on Runtime Verification*, 2010, pp. 198-213.

[69] B. Nielsen, "Towards a Method for Combined Model-based Testing and Analysis", in *Proc. of the 2nd International Conference on Model-Driven Engineering and Software Development,* 2014, pp. 609-618.

13th International Symposium Advances in Artificial Intelligence and Applications

AAIA'18 brings together scientists and practitioners to discuss their latest results and ideas in all areas of Artificial Intelligence. We hope that successful applications presented at AAIA'18 will be of interest to researchers who want to know about both theoretical advances and latest applied developments in AI.

Topics

Papers related to theories, methodologies, and applications in science and technology in the field of AI are especially solicited. Topics covering industrial applications and academic research are included, but not limited to:

- Decision Support
- Machine Learning
- Fuzzy Sets and Soft Computing
- Rough Sets and Approximate Reasoning
- Data Mining and Knowledge Discovery
- Data Modeling and Feature Engineering
- Data Integration and Information Fusion
- Hybrid and Hierarchical Intelligent Systems
- Neural Networks and Deep Learning
- Bayesian Networks and Bayesian Reasoning
- Case-based Reasoning and Similarity
- Web Mining and Social Networks
- Business Intelligence and Online Analytics
- Robotics and Cyber-Physical Systems
- AI-centered Systems and Large-Scale Applications

We also encourage researchers interested in the following topics to submit papers directly to the corresponding workshops, which are integral parts of AAIA'18:

- AI in Medical Applications (AIMA'18 workshop)
- AI in Machine Vision and Graphics (AIMaVIG'18 workshop)
- AI in Reasoning Foundations (AIRIM'18 workshop)
- AI in Information Retrieval (ASIR'18 workshop)
- AI in Data Mining Challenges (DMGATE'18 workshop)
- AI in Smart Energy Networks (SEN-MAS'18 workshop)
- AI in Computational Optimization (WCO'18 workshop)

All submissions accepted to the main track of AAIA'18 and to the above workshops are treated equally in the conference programme and are equally considered for the paper awards.

Professor Zdzisław Pawlak Best Paper Awards

We are proud to continue the tradition started at the AAIA'06 and grant two "Professor Zdzisław Pawlak Best Paper Awards" for contributions which are outstanding in their scientific quality. The two award categories are:

- Best Student Paper. Papers qualifying for this award must be marked as "Student full paper" to be eligible.
- Best Paper Award.

Each award carries a prize of 300 EUR funded by the Mazowsze Chapter of the Polish Information Processing Society.

Event Chairs

- **Kwaśnicka, Halina,** Wrocław University of Science and Technology, Poland
- **Markowska-Kaczmar, Urszula,** Wrocław University of Science and Technology, Poland

Advisory Board

- **Kacprzyk, Janusz,** Polish Academy of Sciences, Poland
- **Marek, Victor,** University of Kentucky, United States
- **Matwin, Stan,** Dalhousie University, Canada
- **Michalewicz, Zbigniew,** University of Adelaide, Australia
- **Skowron, Andrzej,** University of Warsaw, Poland
- **Ślęzak, Dominik,** University of Warsaw, Poland

Area Supervisory Committee

- **Derksen, Christian,** SEN-MAS'18
- **Janusz, Andrzej,** DMGATE'18
- **Lasek, Piotr,** AIMA'18
- **Loukanova, Roussanka,** AIRIM'18
- **Markowska-Kaczmar, Urszula,** AAIA'18
- **Mozgovoy, Maxim,** ASIR'18
- **Śluzek, Andrzej,** AIMaViG'18
- **Zaharie, Daniela,** WCO'18

Program Committee

- **Baron, Grzegorz**
- **Bartkowiak, Anna,** Wrocław University, Poland
- **Bazan, Jan,** University of Rzeszów, Poland
- **Bembenik, Robert**
- **Betliński, Paweł,** Security On Demand, Poland
- **Błaszczyński, Jerzy,** Poznań University of Technology, Poland
- **Chakraverty, Shampa,** Netaji Subhas Institute of Technology, India
- **do Carmo Nicoletti, Maria,** UFSCar & FACCAMP, Brazil
- **Franova, Marta,** CNRS, LRI & INRIA, France
- **Froelich, Wojciech,** University of Silesia, Poland
- **Gawrysiak, Piotr**
- **Girardi, Rosario,** UNIRIO, Brazil

- **Jaromczyk, Jerzy,** University of Kentucky, United States
- **Jatowt, Adam,** Kyoto University, Japan
- **Jin, Xiaolong,** Institute of Computing Technology, Chinese Academy of Sciences, China
- **Kasprzak, Włodzimierz,** Warsaw University of Technology, Poland
- **Korbicz, Józef,** University of Zielona Góra, Poland
- **Kryszkiewicz, Marzena,** Warsaw University of Technology, Poland
- **Kulikowski, Juliusz,** Institute of Biocybernetics and Biomedical Engineering, Poland
- **Lopes, Lucelene,** PUCRS, Brazil
- **Matson, Eric T.,** Purdue University, United States
- **Menasalvas, Ernestina,** Universidad Politécnica de Madrid, Spain
- **Miyamoto, Sadaaki,** University of Tsukuba, Japan
- **Moshkov, Mikhail,** King Abdullah University of Science and Technology, Saudi Arabia
- **Myszkowski, Paweł B.,** Wroclaw University of Technology, Poland
- **Nowostawski, Mariusz,** Norwegian University of Technology and Science (NTNU), Norway
- **Ohsawa, Yukio,** University of Tokyo, Japan
- **Peters, Georg,** Munich University of Applied Sciences, Germany
- **Po, Laura,** Università di Modena e Reggio Emilia, Italy
- **Porta, Marco,** University of Pavia, Italy
- **Przybyła-Kasperek, Małgorzata,** University of Silesia, Poland
- **Raś, Zbigniew,** University of North Carolina at Charlotte, United States
- **Rauch, Jan,** University of Economics, Prague, Czech Republic
- **Reformat, Marek,** University of Alberta, Canada
- **Schaefer, Gerald,** Loughborough University, United Kingdom
- **Sikora, Marek,** Silesian University of Technology, Poland
- **Sikos, Leslie F.,** University of South Australia, Australia
- **Skonieczny, Lukasz**
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates
- **Stańczyk, Urszula,** Silesian University of Technology, Poland
- **Subbotin, Sergey,** Zaporizhzhya National Technical University, Ukraine
- **Sydow, Marcin,** Polish Academy of Sciences & Polish-Japanese Academy of Information Technology, Poland
- **Szczęch, Izabela,** Poznań University of Technology, Poland
- **Szczuka, Marcin,** University of Warsaw, Poland
- **Szpakowicz, Stan,** University of Ottawa, Canada
- **Szwed, Piotr,** AGH University of Science and Technology, Poland
- **Tomczyk, Arkadiusz,** Łódź University of Technology, Poland
- **Unland, Rainer,** Universität Duisburg-Essen, Germany
- **Unold, Olgierd,** Wrocław University of Technology, Poland
- **Zakrzewska, Danuta,** Łódź University of Technology, Poland
- **Zielosko, Beata,** University of Silesia, Poland
- **Ziółko, Bartosz,** AGH University of Science and Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 15–20
DOI: 10.15439/2018F52
ISSN 2300-5963 ACSIS, Vol. 15

Kestrel-based Search Algorithm (KSA) for parameter tuning unto Long Short Term Memory (LSTM) Network for feature selection in classification of high-dimensional bioinformatics datasets.

Israel Edem Agbehadji
ICT and Society Research Group Department of Information Technology Durban University of Technology, Durban, South Africa. Email: *21648757@dut4life.ac.za*

Richard Millham
ICT and Society Research Group Department of Information Technology Durban University of Technology, Durban, South Africa. Email: *richardm1@dut.ac.za*

Simon James Fong
ICT and Society Research Group Department of Computer and Information Science University of Macau, Macau, SAR Email: *ccfong@umac.mo*

Hongji Yang
Department of Computer Science University of Leicester Leicester, UK Email: hongji.yang@gmail.com

***Abstract*—Although deep learning methods have been applied to the selection of features in the classification problem, current methods of learning parameters to be used in the classification approach can vary in terms of accuracy at each time interval, resulting in potentially inaccurate classification. To address this challenge, this study proposes an approach to learning these parameters by using two different aspects of Kestrel bird behavior to adjust the learning rate until the optimal value of the parameter is found: random encircling from a hovering position and learning through imitation from the well-adapted behaviour of other Kestrels. Additionally, deep learning method (that is, recurrent neural network with long short term memory network) was applied to select features and the accuracy of classification. A benchmark dataset (with continuous data attributes) was chosen to test the proposed search algorithm. The results showed that KSA is comparable to BAT, ACO and PSO as the test statistics (that is, Wilcoxon signed rank test) show no statistically significant differences between the mean of classification accuracy at level of significance of 0.05. However, KSA, when compared with WSA-MP, shows a statistically significant difference between the mean of classification accuracy.**

***Index Terms*—kestrel-based search algorithm, deep learning, random encircling, long short term memory network.**

I. Desire Models

THE CONCEPT of big data may be characterized by volume, velocity, value, veracity and variety. The volume relates to the amount of data that has to be processed within a given time; velocity relates to how fast incoming data need to be processed and how quickly the receiver of information needs the results from the processing system [1]; and the value is what a user will gain in terms of insight from the data analysis; the variety is the different structures that data may take such as text and images while the veracity is authenticity of the data source. In order to manage effectively these aspects of big data, an important step is to reduce the volume of dataset by selecting relevant features for classification. However, this may not be achieved without tuning different parameters that fit the data to select relevant features and ensure accurate classification. This paper proposes a search strategy for classification that is based on the behaviour of kestrel bird (to discover the optimal weight parameter) and deep learning network (for classification of features).

The related work is presented in Section II. Section III describes the behaviour of the Kestrel bird with its mathematical modeling and algorithm. Section IV outlines the experimental setup; and provides experimental results with comparative meta-heuristic algorithms. Conclusions and future work are given in Section V.

II. Related Work

Feature selection is the process of selecting relevant features from large number of features in a dataset, while ignoring the rest of features that have little value on the output feature set. The feature selection methods are categorized into the filter method (that is classifier-independent) [2], wrapper method (that is classifier-dependent) [2] and embedded method [3]. However, when big data is involved, it results in high computational cost in training and selection of features [4]. This challenge led to the concept of deep learning which historically originated from artificial neural network [5].

A. Deep learning network

Deep learning is a sub-field of machine learning that is based on learning several levels of representations, corresponding to a hierarchy of features where higher-level features are defined from lower-level ones, and the same lower level features can help to define many higher-level features [5, 6]. It has been indicated in [7] that deep learning is a statistical technique that help in classifying patterns based on sampled data using neural networks with multiple layers. In principle, deep learning uses multiple hidden layers of non-linear processing that is hierarchical; and uses different parameters to learn from hidden layers using algorithms (such as back-propagation algorithms) with large amounts of available training data [8]. Deep learning methods for classification are deep discriminative models/supervised-learning (e.g., deep neural networks (DNN), recurrent neural networks (RNN), etc.) and generative/unsupervised models (e.g., deep belief networks (DBN), etc.). Deep neural network (DNN) sometimes referred at as DBN is a multilayer network with many hidden layers, whose weights are fully connected and initialized (pre-trained) using stacked RBMs or DBN [9]. Recurrent neural networks (RNN) is a discriminative model but also has been used as generative model where output results from a model represents the predicted input data. When RNN is used as a discriminative model, the output results from the model is a label sequence associated with input data sequence [9]. The learning of parame-

ters in RNN are improved by information flow in bi-directional RNN and by a cell with LSTM (long short-term memory where cells are responsible for remembering parameters within a time interval) [6] which are the building units for layers of RNN. The RNN composed of LSTM units is often referred to as LSTM network. However, the challenge with the RNN is that when training neural network for deep learning classification problems, the back-propagated gradients approach that is often used either grows or shrinks at each time step, so over many time steps it typically explodes or vanishes [10]. Building a classification model from deep learning techniques integrated with metaheuristic search methods (also referred to as random search strategy as earlier mentioned) enhances accuracy/quality to select useful and relevant features [11] in a dataset. The advantage of meta-heuristic search method is the use of random search strategy to avoid being trapped in local optima when the search space grows exponentially.

B. *Meta-heuristic algorithms*

Among the random search/meta-heuristic algorithms for feature selection in classification problems are Genetic algorithm (GA) [12], Ant Colony Optimization (ACO) [13], Particle Swarm Optimization (PSO) [14], BAT [15] and Wolf Search Algorithm (WSA) [16].

Genetic algorithms is an evolutionary approach that is based on survival of the fittest. Genetic algorithm has the biological principle that species live in a competitive environment and their continuous survival depends on the mechanics of "natural selection" (Darwin, 1868 as cited by [12]) in which an element or chromosomes in the genetic structure is represented by a binary string. A genetic algorithm is an adaptive search procedure which involves the use of operators such as crossover, mutation and selection methods to find a global optimal results/solution by optimizing an objective function/fitness function.

The Ant Colony Optimization (ACO) [13] is a meta-heuristics search method that is inspired by the foraging behavior of real ants in their search for the shortest paths to food sources. When a source of food is found, ants deposit pheromone to mark their path for other ants to traverse. Pheromone is an odorous substance that is used as a medium for indirect communication among ants. The quantity of pheromone depends on the distance, quantity and quality of food source. However, pheromone substance tends to decay or evaporate with time. While a lost ant that moves at random detects a laid pheromone, it is likely that it will follow the path to reinforce the pheromone trails by further depositing some amount of the trail substances while this path leads to a desired outcome. If the path does not lead to a desired outcome, it is no longer followed and the pheromone evaporates in time until it is no longer detectable. Thus, ants make probabilistic decisions on updating their pheromone trail and local heuristic information in order to explore larger search areas. The ACO has been applied to solve many optimization related problems, including data mining, where it was shown to be efficient in finding best possible solutions. ACO, when applied to feature selection, improves on performance of feature selection by finding the best possible path.

The Wolf Search Algorithm (WSA) [16], is bio-inspired heuristic optimization algorithm which is based on wolf preying behavior. The behaviour of wolves includes the ability to hunt independently by remembering their own trait (meaning wolves have memory); ability to only merge with its peer when the peer is in a better position (meaning there is trust among wolves to never prey on each other); ability to escape randomly upon appearance of a hunter; and the use of scent marks as a way of demarcating its territory and communicating with other wolves of the pack [17].

The Bat algorithm [15] is a bio-inspired method based on the behaviour of micro-bats in their natural environment. The unique behaviour that characterize bats is their echolocation mechanism. This mechanism helps bats orient and find prey within their environment. The search strategy of bat is controlled by the pulse rate and loudness of their echolocation mechanism. Whilst the pulse rate changes to improve on better position that was previously found, the loudness indicates to each other bat that best position is accepted/found. The bat behaviour has been applied in several optimization problems to find the best optimal solution. The bat algorithm search process starts with random initialization of the population, evaluation of the new population using a fitness function and finding the best population. Unlike wolf algorithm that uses attractiveness of prey to govern its search, bat algorithm uses the pulse rate and loudness to control the search for the optimal solution.

The Particle swarm [14] is a bio-inspired method based on the swarm behaviour such as fish and bird schooling in nature. The swarm behaviour is expressed in terms of how particles adapt, exchange information and make decision on change of velocity and position within a space based on position of other neighboring particles. The advantage of swarm behaviour is that as individual particle makes a decision, it leads to an emergent behaviour. This emergent behaviour is as a result of local interaction among individual particles in a population of particles.

The novelty of this paper is the integration of RNN with LSTM, with the proposed bio-inspired/meta-heuristic search method for feature selection. The section III discusses the proposed bio-inspired search method that tune parameters unto an RNN with LSTM so as to select features.

III. Proposed Kestrel-Based Search Algorithm

The bio-inspired algorithm is based on the behaviour of Kestrel bird when hunting for a prey. The Kestrel is a kind of bird that hunts by hovering (that is flight-hunt) or from a perch. These birds are strongly territorial and hunt individually. Author of [18] has shown that during a hunt, Kestrels are imitative rather than cooperative. This suggests that Kestrels prefer not to communicate with each other but rather they imitate the behaviour of other Kestrels with better hunting technique. Authors of [19] have shown that hunting behaviour can change based on type of prey, prevailing weather conditions and energy requirements (for gliding or dive). Aside these behaviour, during hunt, Kestrels use their eyesight to watch small and agile prey within its circling radius or coverage area referred to as the visual circling radius. The minute air disturbance from flying preys, and trail

of urine and faeces from ground preys give an indication of the availability of prey. Once available prey is detected, the Kestrel positions itself to hunt. Kestrels are able to hover in changing airstream, maintain fixed forward looking position with its eye on a prey, and uses random bobbing of head to find the least distance between its position and the position of a prey. Also, the Kestrel possess an excellent ultraviolet sensitive eyesight characteristic to visually locate trails because these trails of urine and faeces of prey reflect ultra-violet light.

In hovering, Kestrel perform a wider search (global exploration) across territories within their visual circling radius, maintain a motionless position with its forward looking eye fixed on prey, detect minute air disturbance from flying prey (particularly flying insects) to best position themselves to hunt prey, and mostly move with precision through changing airstream. Kestrels are able to flap their wings and adjust their long tails to stay in a place that is referred to as a still position in changing airstream. While in perch, mostly from high fixed structures, Kestrel changes its perch every few minutes, performs a thorough search (a local exploitation using its individual hunt behaviour) of its local territory with less energy requirements than a hovering hunt, and uses its ultraviolet sensitive capabilities to detect mammals such as voles closer to a perched area. The characteristics of Kestrels are summarized as follows:

1) Soaring: gives a larger search space (global exploration) within visual coverage area.

a. Still (motionless) position with forward looking eyesight fixed on prey.

b. Encircles prey beneath with keen eyesight.

2) Perching: Each Kestrel does thorough search (local exploitation) within visual coverage area.

a. Frequent bobbing of head.

b. Attracted to prey using detected visible trail then glides to capture.

3) Imitates the behaviour of a well-adapted Kestrel.

The following assumptions are made on the characteristics of the Kestrel: the still position gives a near perfect circle, thus frequent change in a circle direction depends on position of a prey in shifting the center of its circling direction; Frequent bobbing of head gives a degree of magnified or binocular vision that helps in measuring the distance to a prey that then enables the Kestrel to move with a speed to strike; Attractiveness is proportional to light reflection; thus, the higher or longer a distance from Kestrel to the trail, the less bright a trail. This distance rule applies to both hovering height and distance away from the perch; New trails are more attractive and worth pursuing than an old trail. Thus, the trail decay or trail evaporation depends on the half-life of trail; and a Kestrel, which is not well adapted to an environment, imitates the behaviour of well-adapted kestrels.

A. Mathematical formulation on Kestrel behaviour

The proposed computational model for Kestrel's is based on the description of Kestrel's behaviour and characteristics. The following mathematical expressions depict characteristics of the Kestrel:

1) Random Encircling

Encircling is when Kestrel randomly shifts (or changes) the center of circling direction in order to recognize the current position of prey. As the prey changes its current position, Kestrel uses the encircling behaviour to randomly encircle its prey. This movement of prey determines the best possible position assumed by Kestrel. The encircling $\vec{D}$ [20] is expressed as:

$$\vec{D}=|\vec{C}*\vec{x_p}(t)-\vec{x}(t)| \tag{1}$$

Thus:
$$\vec{C}=2*\vec{r1} \tag{2}$$

Where $\vec{C}$ is the coefficient vector, $\vec{x_p}(t)$ is the position vector of the prey, and $\vec{x}(t)$ indicates the position vector of a Kestrel, *r1* and *r2* are random numbers generated between 0 and 1.

2) Current position

The current best position of Kestrel is expressed as:

$$\vec{x}(t+1)=\vec{x_p}(t)-\vec{A}*\vec{D} \tag{3}$$

Thus:
$$\vec{A}=2*\vec{z}*\vec{r2}-\vec{z} \tag{4}$$

Where $\vec{A}$ is coefficient vector, $\vec{D}$ is the encircling value obtained, $\vec{x_p}(t)$ is the position vector of the prey, $\vec{x}(t+1)$ represents the current best position of Kestrels. $\vec{z}$ represents a parameter to control the active mode with $\vec{z}_{hi}$ as the parameter for flight mode and $\vec{z}_{low}$ as the parameter for perched mode, which linearly decreases from 2 (high active mode value) to 0 (low active mode value) respectively during the iteration process. This is expressed as:

$$\vec{z}=\vec{z}_{hi}-(\vec{z}_{hi}-\vec{z}_{low})\frac{itr}{Max_{itr}} \tag{5}$$

Where *itr* is the current iteration, *Max_itr* is the total number of iterations which are performed during the search. Other Kestrels that are involved in the search update their position according to the best position of the leading Kestrel. Also, the change in position of a Kestrel in airstream depends on frequency of bobbing, attractiveness and trail evaporation. This is expressed as the following:

a) Frequency of bobbing

The frequency of bobbing *f* is used for sight distance measurement in the search space. This frequency is expressed as:

$$f^{k}_{t+1}=f_{min}+(f_{max}-f_{min})*\alpha \tag{6}$$

Where, $\alpha \in [0,1]$ is a random number to control the frequency of bobbing within a visual range. f_{max} represents the maximum frequency and f_{min} is the minimum frequency both between 1 and 0 respectively.

b) Attractiveness

Attractiveness β indicates the light reflected from a trail, which is defined by:

$$\beta(r)=\beta_o e^{-\gamma r^2} \tag{7}$$

Where $\beta_o$represents the attractiveness, γ represents variation of light intensity between [0, 1]. *r* represents the sight distance $s(x_i, x_c)$ measurement which is expressed using Minkowski distance formulation as:

$$s(x_i, x_c) = \left(\sum_{k=1}^{n} |x_{i,k} - x_{c,k}|^{\lambda}\right)^{\frac{1}{\lambda}} \quad (8)$$

Thus, $$V \leq s(x_i, x_c) \quad (9)$$

Where x_i is the current sight measurement, x_c are all potential neighboring sight measurement near x_i, n is the total number of neighboring sights, λ is the order of position being considered (that is, 2), and V is the visual range.

c) Trail evaporation

A definition of a trail is the formation and maintenance of a line [13]. In natural environment, ants use trail both to trace the path to a food source and to prevent themselves from getting stuck in a single food source. Thus, ants, using these trails, can search many food sources in a search space. As ants continue to search, trails are drawn and pheromones are deposited on a trail. This pheromone help ants to communicate with each other about the location of food sources. Therefore, other ants continuously follow this path and also deposit substances for the trail to remain fresh. Similar to ants, Kestrels use trails in search of food sources. However, these trails are rather deposited by preys which provides an indication to Kestrels on availability of food sources. The assumption is that the substances deposited by a prey is similar to pheromone deposited on ants' pheromone trail. Additionally, when the source of food depletes, Kestrels no longer follow this path that leads to the location of a prey. Consequently, the trail pheromone begins to diminish with time at an exponential rate causing trails to become old and not worth pursuing. This diminishment denotes the unstable nature of the trail substances which can be theoretically stated as: if there are *N* unstable substances in a trail with an exponential decay rate γ, then an equation can be formulated to describe how *N* substance decreases in time *t* [21]. This equation is expressed as follows:

$$\frac{dN}{dt} = -\gamma N \quad (10)$$

Since the substances are unstable, it introduces a degree of randomness in the decay process. Thus, decay rate (γ) with time (t) is re-expressed as:

$$\gamma_t = \gamma_o e^{-\varphi t} \quad (11)$$

Where γ_o is a random initial value of substance that is decreased at each iteration and where *t* is the number of iterations or time steps. $t \in [0, Max\_itr]$ where *Max_itr* is the maximum number of iterations. The decay rate $\gamma_t$at time *t* to indicate a new trail or old trail is expressed as:

$$if\ \gamma_t \rightarrow \begin{cases} \gamma_t > 1, & trail\ is\ \text{new} \\ 0, & otherwise \end{cases} \quad (12)$$

Thus, a γ_t value greater than 1 indicates that a trail is new and trail is not decayed therefore KSA explores the search area, while a $\gamma_t$value of 0 indicates that trail is old, unattractive and trail has decayed therefore KSA would not explore the search area. Again, the decay constant φ is expressed by:

$$\varphi = \frac{\phi_{max} - \phi_{min}}{t_{1/2}} \quad (13)$$

Where φ is the decay constant, $\phi_{max}$is the maximum number substances in trail, ϕ_{min} is the minimum number of substances in trail and $t_{\frac{1}{2}}$ is the half-life period of a trail. Finally, position of Kestrel is expressed by:

$$x_{i+1}^k = x_i^k + \beta_o e^{-\gamma r^2}\left(x_j - x_i^k\right) + f_i^k \quad (14)$$

Where x_{i+1}^k is the current best position of the Kestrel that represents candidate solution and $x_i^k$is the previous position of Kestrel. Where $\beta_o e^{-\gamma r^2}$represents the attractiveness as expressed in equation (7) where γ is equal to γ_t. x_j represents a Kestrel with a better position whilst f_i^k is the frequency of bobbing as expressed in equation (6).

d) Velocity

The velocity of Kestrel is updated using the expression:

$$v_{t+1}^k = v_t^k + x_t^k \quad (15)$$

Where v_{t+1}^k is the current best velocity, v_t^k represents the initial velocity, whilst x_t^k represents the current best position of Kestrel.

3) Imitative behaviour

Kestrel birds are territorial and hunt individually rather than hunt collectively. As a consequence, a model by [22] that depicts the collective behaviour of birds for feature similarity selection could not be applied. Since Kestrels are imitative, it implies that a well-adapted Kestrel would perform action appropriate to its environment, while other Kestrels that are not well-adapted imitate and remember the successful actions. The imitation behaviour reduces learning and improves upon the skills of less adapted Kestrels. The imitation behaviour is mathematically expressed and applied to select similar features into a subset. A similarity value $Sim_{value(O,T)}$ that helps with the selection of similar features is expressed by:

$$Sim_{value(O,T)} = e^{\left(\frac{-\sum |O_i - E_i|^2}{n}\right)} \quad (16)$$

Where *n* is the total number of features, $\|(O_i - E_i)\|$represents the deviation between two features where O is the observed, E_i is estimate that is the velocity of kestrel in (15). Since the deviation is calculated for each feature dimension and the possibility of large volume of features in dataset, each time a deviation is calculated only the minimum is selected (the rest of the dimension is discarded), thus, to allow the handling of different problem to different scale of dimension of data [23]. Moreover, in cases where features that were imitated are not similar (that is dissimilarity), this is calculated by:

$$dis\_sim_{value(O,T)} = 1 - Sim_{value(O,T)} \quad (17)$$

The fitness function, which is similar to fitness function formulation used by [24], to evaluate each solution is expressed in terms of classification error of the RNN and the similar value obtained from each solution. The fitness function is formulated as:

$$fitness = \rho * Sim_{value(O,T)} + dis\_sim_{value(O,T)} * \rho \quad (18)$$

Where $\rho \in (0,1)$ is a parameter that controls the chances of imitating features that are dissimilar, C_{error} is the classification error of a RNN classifier and $Sim_{value(O,T)}$ refers to the feature similarity value obtained in feature imitation.

Our method to select features uses the RNN with LSTM network (as discussed in section II) and to also make decision on classification accuracy. Authors of [24] has shown that, the less the number of features in a subset and the higher the classification accuracy, the better the solution. The proposed algorithm to implement feature selection is expressed in Table 1 as follows:

TABLE 1: PROPOSED ALGORITHMIC STRUCTURE

Set parameters Initialize population of n Kestrels using equation. Start iteration (loop until termination criterion is met) Generate new population using random encircling Compute the velocity of each kestrel using equation (15) Evaluate fitness of each solution (18) Update encircling position for each Kestrel for all i=1 to n Find the optimal features using RNN with LSTM End loop Output results

In Kestrel Search Algorithm, each kestrel referred as search agent checks the brightness of trail substances using the half-life period; random encircling of each position of a prey before moving with a velocity; imitates the velocity of another Kestrel so that each Kestrel will swarm to the best skilled search agent.

IV. EXPERIMENTAL RESULTS

A. Experimental setup

The proposed algorithmic structure was implemented in MATLAB 2018A. In each run, we performed 100 iterations to select the best/optimal parameter. The best parameter was fed into the LSTM network in which 100 epochs were performed as suggested by [25] that it guarantees optimum results on classification accuracy. To avoid the network instability, all neurons in the input to output layers on a network learned at the same rate (that is with smaller learning rate) [25]. The initial parameters for each meta-heuristic algorithm is defined as follows: KSA (Frequency of bobbing (fb=0.97); perched parameter (zmin=0.2); flight parameter (zmax=0.8); half-life parameter (half-life=0.5); dissimilarity = 0.2; similarity =0.8); PSO [14] (w=1;c1=2.5;c2=2.0); ACO [13] (α=1;ρ=0.05); BAT [15] (β=1; A=1; r=1); WSA-MP [16] v=1;pa = 0.25; α = 0.2, which were suggested by authors of the algorithms as the best parameter that guarantee an optimal solution. To test the robustness of our proposed algorithm, six benchmark datasets shown in Table 2 (from Arizona State University) were used as it represent a standard benchmark dataset with continuous data.

TABLE 2: DATASET FOR EXPERIMENT

Dataset	#of Instances	#of classes	#of features in original dataset
Carcinom	174	11	9182
Glioma	50	4	4434
Lung	203	5	3312
SMK_CAN_187	187	2	19,993
Tox_171	171	4	5748
CLL_SUB_111	111	3	11340

B. Experimental results

In order to select the best optimal solution, the study applied the concept that the higher the classification accuracy, the better the solution and hence, the less the number of features in a subset [24]. With this concept in mind, the study first applied KSA and comparative algorithms to find the best learning parameter presented in Table 3. There are ten separate runs performed on each algorithm and the best was recorded as shown in Table 3

It is observed from Table 3 that out of the six datasets, KSA has the best learning parameter (highlighted in bold) in three datasets. The learning parameter of each meta-heuristic algorithm are fed into LSTM and the classification accuracy are recorded in Table 4:

It is observed from Table 4 that the algorithm with the best parameter is not the best choice on some datasets. For instance, BAT produced the best parameter of **0.0002043** on Tox_171 dataset but produced a classification accuracy of 0.6925. It could be observed that KSA provided the highest classification accuracy on four out of six datasets. This shows that our proposed approach can explore and exploit search space efficiently and find the best results that guarantees higher classification accuracy. The results from this experiment also indicate that no single algorithm can perform better than any other. Moreover, the average classification accuracy for each algorithm when computed shows that KSA has the higher average classification accuracy of **0.7267** while PSO has least of **0.4793**. In order to select features, [24] indicated that the higher the classification accuracy, the better the solution and hence, the less the number of features in a subset. Table 5 shows the number of feature selected by each algorithm.

It is observed from Table 5 that KSA selected less number of features in **four** datasets namely **Carcinom, SMK_CAN_187, Tox_171** and **CLL_SUB_111**; PSO selected less feature in **two** datasets namely **Glioma** and **Lung**. Additionally, on average KSA selected 2422 (see table 5) features, with average accuracy of 0.7267 (see table

TABLE 3: LEARNING PARAMETERS OF META-HEURISTIC ALGORITHMS

Learning parameter	KSA	BAT	WSA-MP	ACO	PSO
Carcinom	**1.3557e-07**	1.0401e-07	3.0819e-05	8.7926e-04	0.5123
Glioma	**2.3177e-06**	3.0567e-05	1.9852e-05	9.9204e-04	0.3797
Lung	**5.1417e-06**	4.4197e-05	3.0857e-05	6.231e-04	0.3373
SMK_CAN_187	0.015064	1.338e-05	**4.7188e-05**	2.7294e-05	2.5311
Tox_171	0.16712	**0.0002043**	0.086214	0.0023152	2.2443
CLL_SUB_111	0.82116	0.075597	0.76001	**0.011556**	9.6956
Average	1.67E-01	1.26E-02	1.41E-01	2.73E-03	2.62E+00

TABLE 4: CLASSIFICATION ACCURACY OF META-HEURISTIC ALGORITHMS

Classification Accuracy	KSA	BAT	WSA -MP	ACO	PSO
Carcinom	**0.7847**	0.7806	0.6908	0.7721	0.7282
Glioma	0.7416	0.7548	0.5063	0.7484	**0.7941**
Lung	0.5754	0.5754	0.5754	0.5754	**0.7318**
SMK_CAN_187	**0.6828**	0.6759	0.6585	0.6111	0.2090
Tox_171	**0.7945**	0.6925	0.7880	0.5889	0.2127
CLL_SUB_111	**0.7811**	0.4553	0.7664	0.4259	0.2000
Average	**0.7267**	**0.6558**	**0.6642**	**0.6203**	**0.4793**

TABLE 5: FEATURE SELECTED BY EACH ALGORITHM.

Feature selected	KSA	BAT	WSA -MP	ACO	PSO
Carcinom	**1977**	2015	2839	2093	2496
Glioma	1146	1087	2189	1116	**913**
Lung	1406	1406	1406	1406	**888**
SMK_CAN_187	**6342**	6480	6828	7775	15814
Tox_171	**1181**	1768	1219	2363	4525
CLL_SUB_111	**2482**	6177	2649	6510	9072
Average	**2422**	**3156**	**2855**	**3544**	**5618**

4) and average parameter of 1.67E-01 (see table 3); while on average PSO selected 5618 (see table 5) features, with average accuracy of 0.4793 (see table 4) and average parameter of 2.62E+00 (see table 3).

The study conducted statistical test on classification accuracy to identity the best algorithm. In order not to prejudice which algorithm outperformed each other, the mean of all the algorithms were considered as equal for the statistical analysis. The Wilcoxon signed rank test which is a non-parametric statistical procedure was used because it does not make underlying assumption about the distribution of parameters and underlining dataset for the evolutionary algorithm. The advantage of Wilcoxon test is that it helps to perform pairwise comparison while not making any assumptions about the population used since Wilcoxon test can guarantee to about 95% (that is, 0.05 level of significance) of efficiency if the population is normally distributed. The results on the test statistic is shown in Table 6

TABLE 6: ALGORITHM AND P-VALUE

Algorithm	Asymp. Sig. (2-tailed) (that is, p-value)
BAT – KSA	0.225
WSAMP - KSA	0.043
ACO – KSA	0.080
PSO – KSA	0.173

Based on the results on test statistics ($p<0.05$), the following analysis can be drawn. In respect of KSA comparison with BAT, ACO and PSO, there is no statistically significant differences between the mean of classification accuracy at level of significance of 0.05. Thus, KSA is comparable to BAT, ACO and PSO algorithms. In contrast, the comparison between KSA and WSA-MP shows a statistically significant difference between the mean of classification accuracy, where $p< 0.05$ (that is, $0.043<0.05$). Thus, comparing algorithms (KSA and WSA-MP) using the Wilcoxon text show the classification accuracy of these algorithms are different.

V. CONCLUSION AND FUTURE WORK

Compared with meta-heuristic algorithms, the classification accuracy results on KSA is different from WSA-MP while the classification accuracy of KSA is comparable to ACO, BAT and PSO. The advantage of KSA is the ability to adapt to different datasets and guarantees good solutions that is comparable to other meta-heuristic search methods for feature selection.

REFERENCES

[1] Longbottom, C. and Bamforth, R., (2013), "Optimising the data warehouse." Dealing with large volumes of mixed data to give better business insights. Quocirca.

[2] Dash, M. and Liu, H. (1997), "Feature selection for classification, intelligent data analysis 1", pg 131-156.

[3] Kumar, V. and Minz, S. (2014), "Feature selection: A literature review." Smart Computing Review, vol. 4, No. 3

[4] Lin, C-J., Support vector machines: status and challenges. 2006. Available on: https://www.csie.ntu.edu.tw/~cjlin/talks/caltech.pdf

[5] Deng, Li and Yu, Dong (2013), Deep Learning: Methods and Applications. Vol. 7, Nos. 3–4 pages: 197–387.

[6] Deng, Li., Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey. research.microsoft.com. 2013.

[7] Marcus, G. Deep Learning: A Critical Appraisal. 2018 https://arxiv.org/abs/1801.00631

[8] Patel, A. B., Nguyen, T. and Baraniuk, R. G., A Probabilistic Theory of Deep Learning. 2015.

[9] Deng, L., Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey. 2012.

[10] LeCun, Y., Bengio, Y. and Hinton, G. 2015. Review: Deep learning. Nature. Vol. 521

[11] Li, J., Fong, S., Wong, R. K., Millham, R. and Wong, K. K. L., (2017), "Elitist binary wolf search algorithm for heuristic feature selection in high-dimensional bioinformatics datasets."

[12] Agbehadji, I. E. (2011), "Solution to the travel salesman problem, using omicron genetic algorithm. Case study: tour of national health insurance schemes in the Brong Ahafo region of Ghana." Online Master's Thesis from KNUST, Accra-Ghana.

[13] Dorigo M. and Cambardella, L. M. (1997), "Ant colony system: A cooperative learning approach to traveling salesman problem," IEEE Trans. Evol., Comput. 1 (1), pp. 53-66.

[14] Kennedy, J. and Eberhart, R. C. (1995), "Particle swarm optimization." Proc. of IEEE International Conference on Neural Networks, Piscataway, NJ. pp. 1942-1948.

[15] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pp. 65–74, 2010

[16] Tang, R., Fong, S., Yang, X-S and Deb, S. (2012), "Wolf search algorithm with ephemeral memory."

[17] Agbehadji, I. E., Millham, R. and Fong, S. (2016), "Wolf search algorithm for numeric association rule mining." 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA 2016). Chengdu, China.

[18] Varland, D. E. (1991), "Behavior and ecology of post-fledging American Kestrels." Retrospective Theses and Dissertations Paper 9784.

[19] Vlachos, C, Bakaloudis, D., Chatzinikos, E., Papadopoulos, T. and Tsalagas, D. (2003), "Aerial hunting behaviour of the lesser Kestrel falco naumanni during the breeding season in thessaly (Greece)."

[20] Kumar, R. (2015), "Grey wolf optimizer (GWO)".

[21] Spencer, R. L. (2002), "Introduction to Matlab."

[22] Cui, X., Gao, J. and Potok, T. E. (2006), "A flocking based algorithm for document clustering analysis." 2006.

[23] Blum, A. L. and Langley, P. (1997), "Selection of relevant features and examples in machine learning." Artificial Intelligence, vol. 97, pp. 245-271.

[24] Mafarja, M. and Mirjalili, S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing. 2018.

[25] Batres-Estrada, G. 2015, Deep Learning for Multivariate Financial Time Series

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 21–24
DOI: 10.15439/2018F359
ISSN 2300-5963 ACSIS, Vol. 15

News articles similarity for automatic media bias detection in Polish news portals

Katarzyna Baraniak
Polish-Japanese Academy of Information Technology
Warsaw, Poland
Email: katarzyna.baraniak1@pjwstk.edu.pl

Marcin Sydow
Institute of Computer Science, Polish Academy of Sciences
and Polish-Japanese Academy of Information Technology
Warsaw, Poland
Email: msyd@ipipan.waw.pl

***Abstract*—Digital media have enormous impact on the public opinion. In the ideal world the news in public media should be presented in a fair and impartial way. In practice the information presented in digital media is often biased and may distort the opinion on a given entity/event or concept. It is important to work on tools that could support the detection and analysis of the information bias. One of the first steps is to study the methods of automatic detection of the articles reporting on the same topic, event or entity to further use them in comparative analysis or building a test or training set.**

In this paper we report on the experimental results concerning the problem of automatic detection of articles reporting on the same events or entities. We also report some experiments on detecting the source of information based on the content.

I. Introduction

In the ideal world the news in the public media should be presented in a fair and impartial way so that the reader is provided with honest and high-quality unbiased information and can make his own opinion about the political, economical, social or historical events, entities or concepts, etc.

Impartiality of the media that present news to the citizens is a crucial property of democratic system and is what one would expect.

For several reasons the information presented in the media is usually far from being impartial. One of the reasons for this is that various people may see the world events differently what may influence the way they present them. More importantly, in some cases the authors of news articles can intentionally introduce some bias into their publications, e.g. for political reasons, etc.

The problem is even more important in cases when some media (web portals, magazines, etc.) *systematically* introduce consequent intentional bias into the published content in order to intentionally misinform the reader about the state of the world.

This problem is important especially for the digital media, since they have significant influence on public opinion. The way they work changes over the time but they still remain one of the main source of information about daily events. The problem of text bias is common. It happens in newspapers, blogs, social networks etc. Each source of media may represent different point of view. Even such media as news portals, that should present impartial information, can describe events or people framing it differently.

It would be very valuable to work on tools that could support the detection and analysis of such systematic or intentional information bias in digital media in order to contribute to improve the quality and fairness of the information provided to the citizens.

Such tools are very complex and involve interdisciplinary approach including the elements of artificial intelligence, text mining, statistical data analysis, psychology, sociology, etc. One of a basic modules in any bias-analysis tool is a module that makes it possible to automatically or semi-automatically detect *pairs* of news articles (or, more generally: text documents), that report on the same event, topic or entity. Such pairs of articles, where each article comes from a different *source* (e.g. web portal, particular author, etc.) can be further used to make comparison-based analysis towards detecting information bias. The pairs are also necessary to build a training, test or reference set in the case of machine-learning approach to the described research problem.

In this paper we present a method and experimental results of detecting pairs of news articles on the same (similar) topic or reporting the same (similar) event, etc. We focus here on the news articles in news portals.

II. Related work

There exist multiple approaches to identifying text bias. For most of them the first inevitable phase of bias identification is to find the pairs (or clusters) of similar articles, paragraphs or sentences.

A. News articles similarity

The approach of finding similar texts by using Siamese networks is described in [8]. Siamese networks describe how similar a pair of text documents are. This networks use the same architecture of network and feed two text documents as an input. Then, given such an input pair, an output in the form of the value representing the distance, for example Manhattan distance, between the two text documents from the output is calculated as a measure of (dis)similarity.

In the paper [3] the author describes document text representations and variety of similarity measures for text clustering. They include the measures like: cosine similarty, Euclidean distance, Jaccard coefficient, Pearson Correlation Coefficient and Averaged Kullback-Leibler Divergence. Then, based on

the results of a clustering algorithm, there is made a comparison of the results on datasets including variety of news articles, academic papers and web pages.

Authors of the article [12] describe a similarity measure for news recommender systems. In this work there is made a comparison-based analysis of human judgement of similarity and some other measures such as: Lin and WASP measures.

Another work [7] presents research concerning articles on events. In particular it concerns tracking similar articles and clustering them to summarise the events under interest.

B. Media bias

In work [11] the authors identify the news framing which is the way of presentation of news. They compare it to google trends and demonstrate how news framing influences the public attention. They used the concept of mean similarity of a corpus. The mean similarity is calculated on pairs of n articles by average cosine distance of DocVec representation of articles. They discovered that the public opinion change with the mean similarity.

In paper [10] the authors describe some linguistic features that reveal the bias in a text. They refer to the form of verbs, part of speech tags and subjective words. Instead of news articles they used data from Wikipedia, however their results may be used also for other types of texts.

A related work concerning the usability of linguistic features in the task of detecting special form of bias related to the phenomenon of Web Spam is presented in [9].

An interesting approach is presented in the very recent paper [2], where the authors identify three roles of entities framing people in news articles. These roles are hero, villain and victim. The results are presented in a visual form to compare how entities are described in different articles.

Article [4] presents an approach of identifying bias through analysis of mentions and quotations of politicians among different parties and in different periods of time.

Bias and its propagation is also investigated in social networks [6]. This work used twitter data to identify bias in short texts and to analyse its propagation among the users.

III. Problem Specification

In this paper we consider two research problems:

- news similarity detection problem
- information source recognition problem

A. News similarity detection

We proposed two approaches to news similarity detection: find all similar articles to the given one and given two articles decide if they are similar or not. The first approach to the problem is as follows. In a given collection of news articles from a given time window (e.g. particular day, etc.) detect the groups of articles that report on the same topic/event, etc. A manually labelled training set that is a collection of manually grouped news articles is prepared. We apply text mining techniques to identify the similar events. The models are evaluated using the metrics presented in the next section: averaged precision, averaged recall and averaged F-measure.

The second approach is to identify if 2 articles are similar. We apply the machine-learning approach to this problem. For each article there are computed several attributes based on the textual contents, keywords, etc. Then, the set is used to train some ML models. Finally, the models are used to automatically detect similar articles. The models are evaluated using the prepared group labels and some standard measures such as precision, recall or f-measure.

B. Towards news bias detection

The second research problem studied in this paper is the following. Given an article and the set of information sources (e.g. web news portals) is it possible to automatically recognise which source does this article comes from based only on the contents? This kind of experiment can be viewed as a one of simple tests of imparity of information sources. I.e. if it is possible to correctly predict the source of the news article based on its content then it is more likely that this information source has some information bias.

Of course some other reasons may make it possible to predict the source of the news article including the writing style, etc. However this simple test may serve as one of the multiple tools that could in ensemble help to detect information bias. More advanced bias-detection tools are envisaged in our ongoing research.

IV. Experimental Setup

A. Data Collection

We collect the data from two Polish news portals: 'dorzeczy.pl' and 'gazeta.pl'. These two are chosen from among the most popular Polish news portals. In addition, they are considered by many readers as examples of media having completely different views on the reality in Poland especially in the domain of social issues or politics and hence making it possible to build an interesting dataset with a potential of containing pairs (or clusters) of articles on the same/similar topic/event/entity but with potentially various forms of information bias. Articles are categorised by a predefined set of topic categories on each of the portals. The decision was made to focus on events connected to the politics in Poland or world. In this case we were looking for articles from category 'Information' ('Wiadomosci') in 'gazeta.pl' and in portal 'dorzeczy.pl' for categories 'Country' and 'World' ('Kraj' and 'Swiat').

In this work we decided to focus only on Polish media but it is possible to extend our research to other languages.

We have collected the articles from 01.01.2018 to 07.04.2018. Table I presents the number of articles.

1) Database: We store the data in MongoDB - a document database. We create article collection of news articles and their comments. The comments made by the users to the articles are not used in the experiments reported in this article but are kept for future, extended analyses. Each item in the collection

Table I
DATASET

news portal	number of articles
gazeta.pl	2623
dorzeczy.pl	4197

Table II
DISTRIBUTION OF ARTICLES AMONG GROUPS

group quantity	number of groups
1	145
2	36
3	17
4 or more	16

represents an article and contains the following fields: _id', 'article_id', title', 'date', 'lead_text', 'text', 'keywords', 'source', 'url' and 'comments'. Field comments contains 'author', date', 'comment_id', 'text'.

B. Data Annotation

For news similarity detection we needed to manually create an annotated data set. The common approach for text similarity recognition is to create a set of article pairs and annotate if they are similar or not. We realised that for news articles this approach may not be the best one. We want to find all articles that are similar and sometimes one news portal describes an event in one article and other news portal writes about this in a series of four articles, for example. Thus we define the task of annotating similar articles as follows.

For a given time window (e.g. a particular date) we collect all articles from the specified web news portals. Each article is assigned to a group with articles about similar event using some particular method. If there is no group with articles describing the event create there is created a new one. The group contains all articles about the same event.

We have annotated 385 articles from 6 randomly chosen days. Articles formed 213 groups. There are groups of consisting of one article or groups containing many articles. The distribution of articles among groups quantity is presented in table II Each record of annotated data contains (among others) the following attributes: 'date', 'article_id', 'group_id'.

C. Data Preprocessing

In the preprocessing phase we apply several operations including: removing stop-words, normalising- convert words to the base form using `Morfeusz` library [1].

D. Evaluation Measures

In order to evaluate experimental results we calculate the average precision, recall and f measure in each experiment.

The average precision is the average of precision of each group. That is given by the following expression:

$$ap = \frac{1}{N}\sum_{n=1}^{N} p_n = \frac{1}{N}\sum_{n=1}^{N} \frac{tp_n}{tp_n + fp_n} \tag{1}$$

Where N is the number of evaluated groups. Accordingly average of recall is given as:

Table III
EVALUATION OF ARTICLE'S SIMILARITY DETECTION

Algorithm	ap	ar	af1
Keywords similar.	0.60	0.57	0.42
Doc2Vec + cos sim	0.72	0.57	0.50
Doc2Vec+bigram+cos simililar.	**0.93**	0.60	0.64
Doc2Vec+trigram+cos similar.	0.92	0.63	0.66
TF-IDF +cos similar.	0.50	**0.69**	**0.53**

$$ar = \frac{1}{N}\sum_{n=1}^{N} r_n = \frac{1}{N}\sum_{n=1}^{N} \frac{tp_n}{tp_n + fn_n} \tag{2}$$

Finally, average F-measure is defined as follows:

$$af1 = \frac{1}{N}\sum_{n=1}^{N} \frac{2 * p_n * r_n}{p_n + r_n} \tag{3}$$

V. EXPERIMENTAL RESULTS ON ARTICLES SIMILARITY DETECTION

A. Group approach

In a group approach of finding similar articles we experimented with three methods for the news similarity detection problem:

- keyword set similarity - this is our simple baseline solution. We compared the number of similar keywords and find the most similar articles using predefined threshold based on the number of keywords.
- tf-idf with cosine similarity- after preprocessing of a textual data, we calculated tf-idf and cosine similarity between articles from a given data frame. Again we choose the most similar articles based on the predefined threshold.
- doc2vec[5] with cosine similarity in three variants: unigrams, bigram phrases, trigram phrases. For Each of these we choose doc2vec based on bag of words model. Similar preprocessing was done as for tf-idf.

The results of evaluation are presented in a table III. The best averaged results for a given measures were highlighted in bold. The best average precision is observed for two doc2vec models. That means for these models there are the least false positives. However the best f-measure and recall is observed for tf-idf algorithm. This algorithm is better choice if we want to find as many similar articles as possible without caring about dissimilar articles among them.

B. Pair approach

In this task we wished to identify if a pair of articles is similar or not. The data was split into test and train datasets as presented in IV. Similar articles was labelled as '1' and not similar articles as '0'. We have created the following features: cosine similarity on tf-idf vectors, number of similar keywords, normalized number of similar entities that is number of similar entities/sum of entities in both articles. In table V we present an evaluation of proposed algorithms.

Table IV
NUMBER OF ARTICLE PAIRS FOR SIMILARITY DETECTION

news portal	training set	test set
similar (1)	320	151
not similar (0)	7837	2624

Table V
EVALUATION OF ONE TO ONE ARTICLES PAIRS

Algorithm	Class	Precision	Recall	F1-score	Support
Siamese	0.0	0.95	0.86	0.90	2624
LSTM	1.0	0.07	0.19	0.10	151
	avg / total	0.90	0.82	0.86	2775
SVM	0.0	0.98	0.91	0.94	2624
polynomial	1.0	0.29	0.63	0.40	151
kernel	avg / total	0.94	0.90	0.91	2775
SVM	0.0	0.98	0.85	0.91	2624
linear	1.0	0.20	0.68	0.31	151
kernel	avg / total	0.94	0.84	0.88	2775
Logistic	0.0	0.98	0.89	0.93	2624
regression	1.0	0.24	0.63	0.35	151
	avg / total	0.94	0.87	0.90	2775
Gradient	0.0	0.96	0.95	0.96	2624
boosting	1.0	0.27	0.28	0.27	151
classifier	avg / total	0.92	0.92	0.92	2775

Except of Siamese LSTM all algorithms have quite good results. Support vector machines occur to be the best one. Siamese neural networks has high results in total but very low scores for similar pairs where the reason may be that it was not able to detect dependencies in long text.

VI. EXPERIMENTAL RESULTS ON NEWS ARTICLE SOURCE DETECTION

We experimented with three machine learning algorithms in the problem stated as prediction of the news article source based on its content. In all the experiments concerning this problem, the articles' attributes explicitly mentioning the actual source (e.g. the "source" attribute) were ignored in the prediction phase. Dataset for this task is presented in table VI. We have used the following algorithms: naive bayes, logistic regression, support vector machines.

The evaluation of proposed methods is presented in table VII. Support vector machines has the best score slightly outperforming logistic regression. These results show that based on simple approach, analysing the basics of used language we are able recognise the source.

VII. SUMMARY AND FUTURE WORK

Our ongoing research concerns the helper problem of recognizing news articles on (nearly) the same topic/event in order to find media bias.We have proposed 2 approaches and presented their advantages and disadvantages.

Table VI
NUMBER OF ARTICLES FOR MEDIA OUTLET DETECTION

news portal	training set	test set
gazeta.pl	2436	395
dorzeczy.pl	3591	606

Table VII
EVALUATION OF ARTICLE'S MEDIA OUTLETS DETECTION

algorithm	news portal	precision	recall	f1-score	support
Naive	dorzeczy.pl	0.69	0.98	0.81	606
Bayes	gazeta.pl	0.89	0.32	0.47	395
	avg / total	0.77	0.72	0.67	1001
Logistic	dorzeczy.pl	0.90	0.83	0.86	606
Regression	gazeta.pl	0.76	0.86	0.81	395
	avg / total	0.85	0.84	0.84	1001
SVM	dorzeczy.pl	0.88	0.86	0.87	606
	gazeta.pl	0.79	0.83	0.81	395
	avg / total	**0.85**	**0.85**	**0.85**	1001

We also presented preliminary results on predicting the source of news article based on the contents that seems to illustrate that bias might be present as one of the aspects making such prediction possible, however this needs deeper analysis.

Since some news articles often report multiple events, to improve our results, we plan to increase the granularity of recognition i.e. add recognising fragments of articles instead of the whole documents about similar events. That means that it is planned to detect fragments of text concerning similar events and detect bias in them. Also, we aim to extend research to other languages (e.g. Polish, English).

REFERENCES

[1] http://sgjp.pl/morfeusz/morfeusz-siat.html.

[2] D. Gomez-Zara, M. Boon, and L. Birnbaum. Who is the hero, the villain, and the victim?: Detection of roles in news articles using natural language techniques. In *23rd International Conference on Intelligent User Interfaces*, pages 311–315. ACM, 2018.

[3] A. Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[4] K. Lazaridou and R. Krestel. Identifying political bias in news articles. *Bulletin of the IEEE TCDL*, 12, 2016.

[5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.

[6] H. Lu, J. Caverlee, and W. Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222. ACM, 2015.

[7] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the second international conference on Human Language Technology Research*, pages 280–285. Morgan Kaufmann Publishers Inc., 2002.

[8] P. Neculoiu, M. Versteegh, and M. Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, 2016.

[9] J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.

[10] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1650–1659, 2013.

[11] K. Sheshadri, C.-W. Hang, and M. Singh. The causal link between news framing and legislation. *arXiv preprint arXiv:1802.05768*, 2018.

[12] N. Tintarev and J. Masthoff. Similarity for news recommender systems. In *Proceedings of the AHâĂŹ06 Workshop on Recommender Systems and Intelligent User Interfaces*. Citeseer, 2006.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 25–28
DOI: 10.15439/2018F87
ISSN 2300-5963 ACSIS, Vol. 15

Data Compression Measures for Meta-Learning Systems.

Marcin Blachnik
Silesian University of Technology
Department of Applied Informatics
Katowice, ul. Krasińskiego 8, Poland
Email: marcin.blachnik@polsl.pl

Mirosław Kordos
University of Bielsko-Biala
Department of Computer Science and Automatics
Bielsko-Biała, ul. Willowa 2, Poland
Email: mkordos@ath.bielsko.pl

Sławomir Golak
Silesian University of Technology
Department of Applied Informatics
Katowice, ul. Krasińskiego 8, Poland
Email: slawomir.golak@polsl.pl

***Abstract*—An important issue in predictive modeling is model selection. This process is time consuming and can be simplified with meta-learning. However, meta-learning systems need appropriate data descriptors for proper functioning. One of them are data compression measures which can be extracted out of the instance selection methods. When we only need to estimate the classification accuracy of the model, the compression obtained from instance selection is a good approximator, but when we need to estimate other performance measures such as the precision and sensitivity then the quality of the estimated performance drops. To overcome this issue we propose a new type of compression measure: the *balanced compression* which is sensitive to the class label distribution and shows high correlation with precision and sensitivity of the final classifiers. We also show that the application of the *balanced compression* as a meta-learning descriptor allows for precise assessment of the model performance, as proved by the presented experimental evaluation.**

I. Introduction

NOWADAYS, meta-learning [1], [2] is gaining more and more popularity. It is aimed at speeding up the prediction model construction which consists of model selection and model parameters optimization. The model selection process can be done without actually training the given model, by using other meta-model which assesses the quality of the data and estimates the performance of the desired classier or returns a ranking.

As shown in [3], [4], a good indicator that characterizes the dataset quality is the compression of the dataset obtained with instance selection algorithms [7]. It is defined as: $Cmp = 1 - \frac{\|\mathbf{P}\|}{\|\mathbf{T}\|}$ where $\mathbf{T} = [\{\mathbf{x}_1, y_1\}, \{\mathbf{x}_2, y_2\}, \ldots \{\mathbf{x}_n, y_n\}]$ is a training dataset that consists of n training instances $\{\mathbf{x}, y\}$, where $\mathbf{x} \in \Re^m$ and y is a label which takes one of l symbols, and the dataset $\mathbf{P}$ is a subset of the instances from $\mathbf{T}$ selected by the instance selection algorithm, so that $\mathbf{P} \subset \mathbf{T}$.

The main idea of using compression as a meta-learning descriptor (also called meta-attribute) is based on the observation that a dataset in which there is a lot of regularity can be compressed well, and thus high prediction accuracy should be achievable, while a dataset containing a lot of irregularities and a lot of noise will have a low compression ratio. Moreover, the instance selection methods are often used at the stage of data preprocessing, which means that the value of compression is obtained without additional computational cost. Some algorithms, including *CNN* and *ENN*, have been identified as the most useful for predicting the final model performance [4]. For example, the correlation between the compression ratio and the accuracy of the *k*NN, Gaussian SVM and Random Forest, obtained for *CNN* and *ENN* instance selection methods is above 0.9. However, the research carried out so far has focused only on the classical definition of the measure of prediction accuracy expressed as the ratio of the correctly classified examples to all evaluated examples.

It turns out that although correlation between compression and classification accuracy is very high, the correlation between compression and other measures of classifier performance is much weaker. We refer to such measures as the average precision (also known as the balanced accuracy), or the average sensitivity also called recall, which are especially important in the context of unbalanced classification problems.

This work addresses this problem by introducing a new type of compression, so-called balanced compression, which takes into account the number of rejected instances which belong to particular classes. The balanced compression linearizes the relationship between compression and precision and between compression and sensitivity. Implementing these measures allows to enhance the meta-learning system performance.

II. Instance Selection Algorithms

As it was mentioned, the compression achieved by instance selection can be used as a measure of the dataset quality. We presented also an intuitive dependence, which indicates that stronger compression is connected with greater regularity of the decision boundaries in the dataset, and at the same time it is easier for the classifier to reconstruct the desired decision boundary. In practice, however, this depends on the particular instance selection algorithm. These algorithms can be divided into three basic groups: condensing methods, noise filters and hybrid methods. Condensing methods are a set of algorithms used to reduce the dataset size, where the only criterion is maximization of compression while maintaining comparable prediction accuracy. A typical example is the *CNN* algorithm [8]. *CNN* was developed for use with the *k*NN classifier to reduce the computational complexity. The acceleration is accomplished by eliminating (compressing)

unnecessary instances in the dataset. However, this does not help increasing the prediction quality of this classifier. There are algorithms that allow for stronger compression at the same accuracy level, e.g. evolutionary based instance selection [6]. However, as the evolutionary approach belongs to the hybrid group, the correlation we observed although is still significant, is weaker then that obtained with *CNN*.

Noise filters, on the other hand, are a set of algorithms created with the purpose of finding and removing training instances that constitute noise in a dataset. An example and historically the first noise filter is the *ENN* algorithm designed to improve prediction accuracy of 1-NN [9]. *ENN* was also developed to work with the *k*NN algorithm. Its operation is based on the analysis of the closest neighborhood of a given instance and checking if the nearest neighbors will vote for the examined instance in accordance with its label. If not, then the instance is removed.

Also generalizations of these algorithms were proposed, where different classifiers, not only *k*NN can be embedded into the instance selection process [10]. However, in the experiments presented in this paper only instance selection based on 1-NN will be considered.

The third group of instance selection algorithms are hybrid methods. They combine the properties of the first two groups. They start by filtering out the noisy samples from the data and then condense the remaining dataset.

As it was shown in [4], each group of instance selection methods behaves differently with regard to the prediction accuracy. For the condensing methods, an increase in compression corresponds to an increase in prediction accuracy. In the case of noise filter methods, this relation is reversed, because the noise filters regularize and clean the datasets from noise. Thus more removed instances indicate here more noisy dataset, which means that with the increase of compression of the noise filters, the reduction of prediction accuracy is observed. The last group - hybrid methods combine both elements. This causes that the relationship between compression and prediction accuracy gets much weaker or totally disappear, because the properties of condensation methods are canceled out by the properties of noise filters. This causes that only the instance selection methods, which obtain different compression depending on noise in data find application in estimating the prediction accuracy.

III. Balanced Compression measure

As mentioned in the introduction, for unbalanced classification problems usually classical accuracy measure is not used and rather other performance measures are evaluated like average precision or average sensitivity. The purpose of these measures is to reflect the quality of the prediction model in the context of the number of instances in individual classes. A similar situation occurs in the case of compression measures. The commonly used compression measure ignores the number of rejected instanced within individual classes. It simply represents the ratio of the number of rejected samples to the size of the training set $\mathbf{T}$, thus, this measure is similar to the classical accuracy used in prediction systems. The natural conclusion from this is that we should adapt the measure of compression to data with unbalanced class distribution, so that the measure not only indicates the number of rejected samples but also the number of rejected samples within individual classes. It can bring tangible benefits in the form of additional information about the nature of the classification problem, in particular in the context of meta-learning systems.

An important difference between accuracy and compression is the fact, that in contrast to the evaluation of the accuracy of the classifier, in the case of compression we do not have the confusion matrix and the values resulting from it like *False Positives* or *False Negatives*. It is because instance selection methods do not perform prediction, instead we only have information which instances were selected and which rejected, so we do not know what type of error occurred. Therefore, the only factor possible to determine is the level of class c_i compression defined as $\frac{\|y_{\mathbf{T}}==c_i\|-\|y_{\mathbf{P}}==c_i\|}{\|y_{\mathbf{T}}==c_i\|}$, where $\|y_{\mathbf{T}} == c_i\|$ denotes the number of samples in the training set $\mathbf{T}$ which belong to class c_i and $\|y_{\mathbf{P}} == c_i\|$ denotes the number of instances in the dataset $\mathbf{P}$ (after instance selection) which belong to class c_i.

Based on this class compressions we define balanced compression as an average over all classes

$$Cmp_{Bal} = \frac{1}{l}\sum_{i=1}^{l}\frac{\|y_{\mathbf{T}} == c_i\| - \|y_{\mathbf{P}} == c_i\|}{\|y_{\mathbf{T}} == c_i\|} \tag{1}$$

where l denotes the number of classes. This measure can be also generalized by introducing class weights denoted as w_i which describes importance of particular class, so the balanced compression takes the form:

$$Cmp_{Bal} = \frac{1}{\sum w_i}\sum_{i=1}^{l} w_i\frac{\|y_{\mathbf{T}} == c_i\| - \|y_{\mathbf{P}} == c_i\|}{\|y_{\mathbf{T}} == c_i\|} \tag{2}$$

In the conducted experiments we assumed equal values of the weights $\underset{i=1\ldots l}{\forall} w_i = 1$.

IV. Experiments and Results

In order to verify the usefulness of the proposed *balanced compression* in the context of meta-learning systems, we carried out an experimental evaluation on 45 datasets obtained from Keel Project [11] using three popular classifiers: *k*NN, linear SVM and Random Forest. The experiments were performed with RapidMiner and the Information Selection package developed by the authors of this paper, which is available from the RapidMiner Marketplace and on the website www.prules.org [12]. The experiments were divided into two parts. In the first part the correlation measure was evaluated between compression measures and performance measures of the evaluated classifiers. It indicates how the new compression measure reflects the obtained classification performances. In the second part a real meta-learning system was constructed which is designed to predict performance of the three classifiers. The meta-learning system utilizes meta-attributes which are based on compressions obtained by both *CNN* and *ENN*.

A. Relationships Between Compression and Various Performance Measures

The first part of the experiments consists of two stages. In stage I, the three performance measures (accuracy, average precision and average sensitivity) for each of the 45 datasets were estimated using the cross-validation procedure. This stage also included parameter optimization for all evaluated classifiers (k for *k*NN, C- for linear SVM and the number of trees for Random Forest). In stage II, each dataset was compressed using the two previously described algorithms *ENN* and *CNN*, each time measuring both compression and balanced compression. The obtained results were then used to calculate Pearson's correlation coefficient between given type of compression and the type of classification performance measure independently for each classifier. Obtained correlations were collected in Tab. I for *CNN* instance selection, and in Tab. II for *ENN* instance selection algorithm.

Table I: Correlation between two types of compression obtained for *CNN* and the three performance measures for *k*NN, Linear-SVM and Random Forest

Compression type	Performance measure	*k*NN	SVM	Random Forest
Compression	Accuracy	0.937	0.902	0.900
Compression	Precision	0.783	0.662	0.767
Compression	Recall	0.738	0.640	0.767
Balanced compression	Precision	0.920	0.794	0.880
Balanced compression	Recall	0.932	0.808	0.880

Table II: Correlation between two types of compression obtained for *ENN* and the three performance measures for *k*NN, Linear-SVM and Random Forest

Compression type	Performance measure	*k*NN	SVM	Random Forest
Compression	Accuracy	-0.965	-0.924	-0.917
Compression	Precision	-0.774	-0.672	-0.745
Compression	Recall	-0.758	-0.661	-0.765
Balanced compression	Precision	-0.935	-0.844	-0.883
Balanced compression	Recall	-0.981	-0.863	-0.895

The results in the tables indicate that the correlation between classical compression and prediction accuracy is very high and ranges from 0.917 to 0.965 for the *ENN* algorithm and from 0.900 to 0.937 for the *CNN* (here for simplicity we evaluate absolute values of the correlation as the sign does not matter). However, changing the measure of the prediction quality to average precision or average sensitivity causes the correlation coefficient to drop rapidly to a level between 0.64 and 0.77 depending on the method of instance selection and on the classifier. Changing compression to the balanced compression results in a significant increase in the correlation coefficient, which for the *k*NN classifier again exceeds 0.9, and for the other classifiers varies between 0.8 and 0.88. This is a significant improvement over the correlation coefficients obtained with standard compression.

B. Meta-system - compression-based estimation of prediction quality

Meta-learning systems are used for the estimation of quality of predictive models [13], [1], [14], [15]. In these systems, for a known dataset repository, which consists of n_r datasets, the prediction performance of the selected classifier is estimated and the meta-attributes describing the properties of each of these datasets are extracted [16]. Next, a meta-set is created. The meta-set consists of the extracted meta-attributes (an input vector of the meta-learning system) and labels that express the accuracy of the given model, for which we would like to estimate the accuracy. Therefore, the meta-set consists of n_r samples, where a single instance describes one dataset from the repository. So we obtain a typical regression problem, because labels in the meta-set represent numerical values (accuracies). In the next step, the meta-set is used to build a meta-model, a model capable of estimating prediction accuracy for a given, previously unknown data set. When applying a meta-model to a new data set, it is necessary in the first step to determine the meta-attributes, create a record from them, and then pass them to the meta-model input. The meta-model then returns the estimated accuracy. Another commonly used solution is learning the meta-ranking model, where the meta-model returns the ranking of the best models or just the best prediction model [17].

It was shown in [4] that the use of compressions as meta-attributes lead to an improvement in the quality of the estimated accuracy in comparison to the classic meta-attributes used in the MLWizzard system [15]. Therefore, in the experiments a meta-system based only on the data set compression measures is constructed.

As a meta-model, Generalized Linear Model was used. In total we had 9 meta-models (for each of the three performance measures and for each of the tree classifiers). The meta-model was tested using the 5x10 cross-validatin procedure. The quality of the whole system was evaluated using *RMSE* calculated between predicted and real prediction performance. The obtained results are presented in Tab. III.

The results are placed in two columns. The first column contains the results obtained using classical compression of both *CNN* and *ENN* as meta-attributes, and the second column contains the results obtained with a balanced compression. For each of the tested classifiers, the three measures of accuracy (accuracy, average sensitivity and average precision) were estimated, and Welch's t-test [18] was used to determine if the results are statistically significantly different at $\alpha = 0.05$. The symbol (+) indicates results which are significantly better.

The obtained results clearly indicate that for meta-learning systems where the task is to estimate classical accuracy, the standard compression measure gives better results. However, when the aim of the process is to estimate average precision or average sensitivity, a much better solution is to use the balanced compression; each time the results obtained using balanced compression were statistically significantly better than those obtained with standard compression.

Table III: Results of the meta-learning system. The columns represent RMSE of the meta-model aimed at estimating classification performance of three classifiers: *k*NN, Linear SVM, and Random Forest using two meta-sets which consisted of: clasical compression based meta-attributes (column 1) and balanced compression - based meta-attributes (column 2)

		Compression RMSE±std	Balanced Compression RMSE±std
*k*NN	Accuracy	0.0328±0.0207(+)	0.0799±0.0400
	Recall	0.1336±0.0529	0.0703±0.0575(+)
	Precision	0.1265±0.0607	0.0884±0.0628(+)
SVM	Accuracy	0.0640±0.0296(+)	0.0910±0.0416
	Recall	0.1523±0.0610	0.1130±0.0600(+)
	Precision	0.1453±0.0711	0.1171±0.0622(+)
Random Forest	Accuracy	0.0437±0.0281(+)	0.0739±0.0408
	Recall	0.1276±0.0632	0.0954±0.0711(+)
	Precision	0.1284±0.0644	0.0979±0.0711(+)

The prediction quality of the *k*NN model can be estimated more precisely than those of SVM or Random Forest, which is natural, as the instance selection methods internally use the nearest neighbor mechanism to evaluate each of the instances. Random Forest ranked lower than *k*NN in terms of performance estimation but ranked higher than SVM. SVM's high performance estimation error was caused by the fact that the SVM considered in this study utilized a linear kernel, and thus it was a linear classifier, while Random Forest is a nonlinear classifier. By their very nature, methods of instance selection are nonlinear, and thus they can overestimate the results obtained by the linear model.

V. Conclusions

In this study we have shown that compression forms a strong linear relationship with the standard prediction accuracy. We have also shown that other measures of prediction quality do not correlate strongly with the standard compression obtained by instance selection.

To address this problem, we proposed a modified measure of compression called balanced compression. The purpose of balanced compression is to express the characteristics of the dataset preserving distribution of the class labels. This allowed to obtain almost linear relationship between the balanced compression and the accuracy measures such as average precision and average sensitivity. The importance of this linear relationship can be efficiently used in meta-learning systems, where the balanced compression allowed for a significant improvement in the estimation of average precision and average sensitivity compared to estimation performed using standard compression.

Acknowledgment

This research was supported by the Silesian University of Technology, Grant 11/040/RGJ17/0014.

References

[1] N. Jankowski, W. Duch, K. Grąbczewski, *Meta-learning in computational intelligence*. Springer Science & Business Media, vol. 358, 2011.
[2] L. Kotthoff, Ch. Thornton, H. Hoos, F. Hutter,K. Leyton-Brown, "Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA," *JMLR*, vol. 18, no. 1, pp. 826–830, 2017.
[3] M. Blachnik, "On the relation between knn accuracy and dataset compression level," *LNAI*, vol. 9692, pp. 541–551, 2016.
[4] M. Blachnik, "Instance selection for classifier performance estimation in meta learning," *Entropy*, vol. 19, no. 11, p. 583, 2017.
[5] M. Kordos, M. Blachnik, J. Kozłowski, M. Perzyk, O. Bystrzycki, M. Gródek, A. Byrdziak, Z. Motyka, "A Hybrid System with Regression Trees in Steelmaking Process," *LNAI*, vol. 6678, pp. 222-229, June 2011.
[6] M. Kordos, "Optimization of Evolutionary Instance Selection," *LNAI*, vol. 10245, pp. 359-369, ICAISC, June 2017
[7] S. García, J. Derrac, J. R. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Trans Pattern Anal and Mach Intell*, vol. 34, no. 3, pp. 417–435, 2012.
[8] P. Hart, "The condensed nearest neighbor rule." *IEEE Trans. on Information Theory*, vol. 16, pp. 515–516, 1968.
[9] D. Wilson, "Assymptotic properties of nearest neighbour rules using edited data." *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-2, pp. 408–421, 1972.
[10] M. Kordos, M. Blachnik, and S. Białka, "Instance selection in logical rule extraction for regression problems," *LNAI*, vol. 7895, pp. 167–175, 2013.
[11] F. Herrera, "Keel, knowledge extraction based on evolutionary learning," http://www.keel.es, 2005, spanish National Projects TIC2002-04036-C05, TIN2005-08386-C05 and TIN2008-06681-C06. [Online]. Available: http://www.keel.es
[12] M. Blachnik and M. Kordos, "Information selection and data compression rapidminer library," in *Machine Intelligence and Big Data in Industry*. Springer, 2016, pp. 135–145.
[13] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., 2015, pp. 2962–2970.
[14] M. Kozielski, "A meta-learning approach to methane concentration value prediction," in *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*. Springer, 2015, pp. 716–726.
[15] M. Reif, F. Shafait, and A. Dengel, "Meta-learning for evolutionary parameter optimization of classifiers," *Machine Learning*, vol. 87, no. 3, pp. 357–380, 2012.
[16] F. Pinto, C. Soares, and J. Mendes-Moreira, "Towards automatic generation of metafeatures," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2016, pp. 215–226.
[17] Q. Sun and B. Pfahringer, "Pairwise meta-rules for better meta-learning-based algorithm ranking," *Machine learning*, vol. 93, no. 1, pp. 141–161, 2013.
[18] B. L. Welch, "The generalization ofstudent's' problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1/2, pp. 28–35, 1947.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 29–33
DOI: 10.15439/2018F200
ISSN 2300-5963 ACSIS, Vol. 15

Deep Evolving Stacking Convex Cascade Neo-Fuzzy Network and its Rapid Learning

Galina Setlak
Rzeszow University of Technology
12 Al. Powstancow Warszawy,
35-959, Rzeszow, Poland
Email: gsetlak@prz.edu.pl

Yevgeniy Bodyanskiy, Iryna Pliss, Olena Boiko
Kharkiv National University of Radio Electronics,
14 Nauky ave., Kharkiv, Ukraine
Email: yevgeniy.bodyanskiy@nure.ua,
iryna.pliss@nure.ua,
olena.boiko@nure.ua

Olena Vynokurova
Kharkiv National University of Radio Electronics,
14 Nauky ave., Kharkiv, Ukraine

IT Step University,
83a Zamarstynivs'ka st., Lviv, Ukraine
Email: vynokurova@gmail.com,

***Abstract*—A deep evolving stacking convex neo-fuzzy network is proposed. It is a feedforward cascade hybrid system, the layers-stacks of which are formed by generalized neo-fuzzy neurons that implement Wang–Mendel fuzzy reasoning. The optimal in the sense of speed algorithms are proposed for its learning. Due to independent layer adjustment, parallelization of calculations in non-linear synapses and optimization of learning processes, the proposed network has high speed that allows to process information in online mode.**

I. Introduction

DEEP neural networks (DNNs) are currently the most intensively developing direction of Computational Intelligence due to their universal capabilities in solving a variety of information processing tasks. At the same time, DNNs are not without significant drawbacks, the main of which is the low speed of training due to the need to use error backpropagation across multiple layers. In this regard, increasing of the training speed of DNNs is a topical task.

It should be noted here that historically the first deep networks [2] were information processing systems based on the group method of data handling (GMDH) [5], [6], where training was conducted sequentially from input to output, all nodes of the system being independently tuned. Another advantage of the GMDH-networks is the possibility of increasing the number of layers to achieve the required accuracy of the resulting solution. Thus, this network evolves over time [7], [8], increasing the number of layers. It is important that the previously formed layers are not tuned anymore in the process of evolution, that significantly reduces the total training time. Deep neural networks based on GMDH were proposed in [9], [10] that exceeded the known DNNs in learning speed. However, in situations when data under processing are received online in the form of an information stream [11], [12], this learning speed may not be sufficient.

In such situations, it is more preferable to use the idea of cascaded neural networks [13], where each cascade is formed by a pool of neurons, and the input signal of each cascade is formed from the inputs of the network and the outputs of the previous cascades.

The usage of the traditional elementary perceptrons by F. Rosenblatt in the cascades leads to a significant increase in the number of these cascades, that again increases the learning time, although in principle the cascade network can operate in online mode. In connection with this, it was suggested in [14], [15] to optimize the output signal in each cascade, and instead of the usual neurons to use neo-fuzzy neurons (NFNs) [16]-[18], that have high approximating properties.

At the intersection of cascade neural networks and deep stacking neural networks [2] deep stacking hybrid networks have emerged [19], [20], where hybrid generalized additive wavelet-neuro-neo-fuzzy systems (HGAWNNFS) were used as stacks-cascades [21]-[25], synthesized on the basis of hybrid systems of computational intelligence and generalized additive models [26]. These systems showed high quality of information processing and high enough speed, although the computational bulkiness of stacks-HGAWNNFS reduces the speed of the network learning.

In this regard, it is interesting to introduce a deep evolving stacking cascade system, that has high learning speed, good approximating properties and that is simple in numerical implementation.

II. The Deep Evolving Stacking Cascade Network Architecture

In Fig. 1 the architecture of deep stacking cascade network is presented. It contains g layers-cascades-stacks [2], [27], [28], each of them is a hybrid system of computational intelligence with high approximating properties.

It can be seen that adding new stacks to the architecture does not require retraining of the already formed layers. Thus, this architecture evolves over time [7], [8], [15] by adding new stacks to achieve the required accuracy.

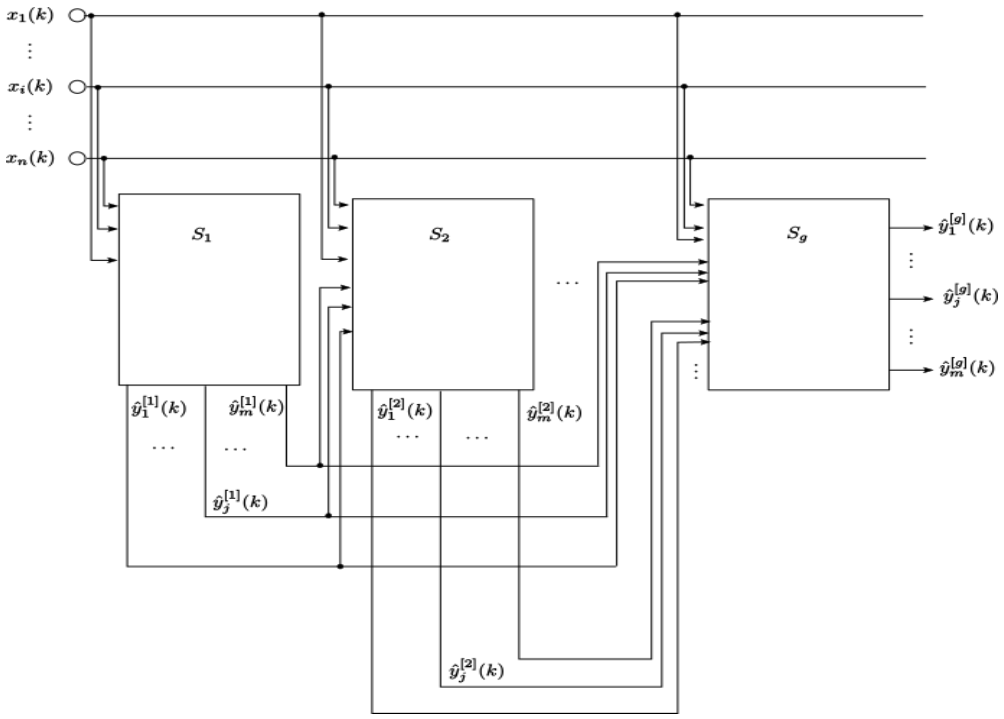


Fig. 1 Deep evolving stacking cascade network

To the input of the network's first layer S_1 an input vector $x(k)=(x_1(k),...,x_i(k),...,x_n(k))^{\mathrm{T}}\in R^n$ is fed (here $k=1,2,...$ is either the number of the observation in the training set, or the current discrete time index). On the output of this layer an output signal $\hat{y}^{[1]}(k)=(\hat{y}_1^{[1]}(k),...,\hat{y}_j^{[1]}(k),...,\hat{y}_m^{[1]}(k))^{\mathrm{T}}\in R^m$ is formed. In the situation when the signal $\hat{y}^{[1]}(k)$ at the output of the trained S_1 satisfies in accuracy all the requirements, i. e. the process of the network forming ends. Otherwise, the second layer S_2 is formed, the input of which is an extended vector $\left(x^{\mathrm{T}}(k),\hat{y}^{[1]\mathrm{T}}(k)\right)^{\mathrm{T}}\in R^{n+m}$ and the output of which is $\hat{y}^{[2]}(k)\in R^m$. To the third stack S_3 a signal $\left(x^{\mathrm{T}}(k),\hat{y}^{[1]\mathrm{T}}(k),\hat{y}^{[2]\mathrm{T}}(k)\right)^{\mathrm{T}}\in R^{n+2m}$ is fed.

And, finally, the input of the S_g is a vector $\left(x^{\mathrm{T}}(k),\hat{y}^{[1]\mathrm{T}}(k),...,\hat{y}^{[g-1]\mathrm{T}}(k)\right)^{\mathrm{T}}\in R^{n+(g-1)m}$, and the output of the whole network is $\hat{y}^{[g]}(k)\in R^m$.

Thus, the network provides a non-linear mapping $R^n\to R^m$, and the number of layers is limited only by the maximal permissible dimension of the input signal of the g th stack. At the same time, when the learning process is paralleled, this restriction is not essential.

It is important that the training of layers-stacks is realized practically independently of each other, and error backpropagation is not required in principle.

III. Generalized Neo-Fuzzy-Neuron as Stack of Proposed Network

As a "building block"-stack of the system under consideration, we propose to use the generalized neo-fuzzy-neuron (GNFN) [29], that is a generalization of the neo-fuzzy neuron (NFN) [16-18] for the multidimensional case. In Fig. 2 the architecture of the first S_1 GNFN-layer is presented. It contains n inputs and m outputs. All other GNFN-layers $S_2,...,S_g$ coincide in architecture with S_1 and differ only in the number of inputs. It should be also noted that GNFN has high approximating properties, simplicity of numerical implementation and parallelization of information processing.

A sequence of input signals $x(k)=(x_1(k),...,x_i(k),...,x_n(k))^{\mathrm{T}}\in R^n$ is fed to the input of a GNFN that is formed by the first layer-stack S_1. This stack consists of n multidimensional parallel non-linear synapses $MNS_i^{[1]}$, $i=1,2,...,n$, each of which has only one input, m outputs, h membership functions $\mu_{li}^{[1]}(x_i(k))$, $l=1,2,...,h$ and mh adjustable synaptic weights $w_{jli}^{[1]}$, $j=1,2,...,m$.

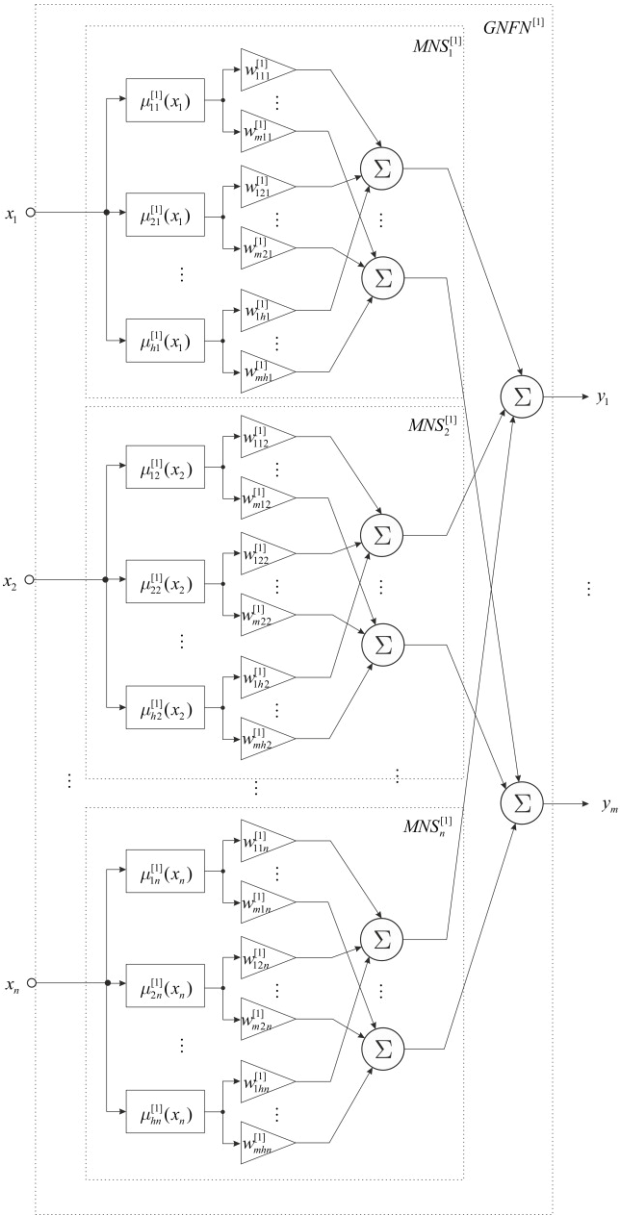


Fig. 2 Generalized neo-fuzzy neuron (GNFN)

The output of the first layer is a vector $\hat{y}^{[1]}(k)=\left(\hat{y}_1^{[1]}(k),...,\hat{y}_j^{[1]}(k),...,\hat{y}_m^{[1]}(k)\right)^{\mathrm{T}}$, that further together with the vector $x(k)$ is fed to the inputs of the layer S_2 in the form of $\left(x^{\mathrm{T}}(k),\hat{y}^{[1]\mathrm{T}}(k)\right)^{\mathrm{T}}$. Thus, S_1 contains nh membership functions and nhm synaptic weights.

Non-linear mapping, realized by this GNFN, in general case can be written in the form

$$\hat{y}_j^{[1]}(k)=\sum_{i=1}^{n}\sum_{l=1}^{h}w_{jli}^{[1]}\mu_{li}^{[1]}\left(x_i(k)\right)\ \ \forall j=1,2,...,m \qquad (1)$$

and it significantly depends both on the type of membership functions used and the algorithm for synaptic weights learning.

It should be also noted that multidimensional non-linear synapses $MNS_i^{[1]}$ in general case are zero-order Takagi–Sugeno–Kang (i.e. Wang–Mendel) neuro-fuzzy systems, that provide high approximating properties.

As the membership functions in the simplest case we can use triangular ones:

$$\mu_{li}^{[1]}(x_i)=\begin{cases}\dfrac{x_i-\bar{x}_{l-1,i}^{[1]}}{\bar{x}_{li}^{[1]}-\bar{x}_{l-1,i}^{[1]}}\ if\ x_i\in\left[\bar{x}_{l-1,i}^{[1]},\bar{x}_{li}^{[1]}\right],\\ \dfrac{\bar{x}_{l+1,i}^{[1]}-x_i}{\bar{x}_{l+1,i}^{[1]}-\bar{x}_{li}^{[1]}}\ if\ x_i\in\left[\bar{x}_{li}^{[1]},\bar{x}_{l+1,i}^{[1]}\right],\\ 0\ otherwise.\end{cases}$$

They satisfy the conditions of unity partition

$$\begin{cases}\mu_{l-1,i}^{[1]}(x_i)+\mu_{li}^{[1]}(x_i)=1\ if\ x_i\in\left[\bar{x}_{l-1,i}^{[1]},\bar{x}_{li}^{[1]}\right],\\ \mu_{li}^{[1]}(x_i)+\mu_{l+1,i}^{[1]}(x_i)=1\ if\ x_i\in\left[\bar{x}_{li}^{[1]},\bar{x}_{l+1,i}^{[1]}\right]\end{cases}$$

where $\bar{x}_{li}^{[1]}$, $l=1,2,...,h$ are membership functions' centers, that are in the simplest case evenly distributed on the x_i-axis.

The usage of triangular membership functions leads to the fact that at each instant of time k only two neighboring functions fire. This allows to adjust not all nhm synaptic weights on each iteration, but only $2nm$ of them. It is clear that the learning speed can be increased in this case.

IV. Deep Stacking Convex Neo-Fuzzy Network Learning

The learning process of the system under consideration is its synaptic weights' adjustment. Due to the cascade architecture of the system, each stack can be trained independently of the others. It is clear that in online mode the learning algorithms used must provide the maximum possible speed, i.e. they have to be based on the Gauss-Newton algorithms of second-order optimization for convex functions. In this case, the network itself is a convex one [30].

The learning process will be considered using the example of the first layer of the system S_1. For this, let's introduce a $(hn\times 1)$-vector of membership functions $\mu^{[1]}(x(k))=\left(\mu_{11}^{[1]}(x_1(k)),\mu_{21}^{[1]}(x_1(k)),...,\mu_{h1}^{[1]}(x_1(k)),\right.$ $\left.\mu_{12}^{[1]}(x_2(k)),...,\mu_{li}^{[1]}(x_i(k)),...,\mu_{hn}^{[1]}(x_n(k))\right)^{\mathrm{T}}$ and $(m\times hn)$-matrix of synaptic weights

$$W^{[1]}=\begin{pmatrix}w_{111}^{[1]} & w_{121}^{[1]} & \cdots & w_{1hn}^{[1]}\\ w_{211}^{[1]} & w_{221}^{[1]} & \cdots & w_{2hn}^{[1]}\\ \vdots & \vdots & \ddots & \vdots\\ w_{m11}^{[1]} & w_{m21}^{[1]} & \cdots & w_{mhn}^{[1]}\end{pmatrix}.$$

Thus, the mapping, realized in the first layer, can be written as

$$\hat{y}^{[1]}(k) = W^{[1]}\mu^{[1]}(x(k)).$$

Next, let's introduce the learning error of the j th component of $\hat{y}_j^{[1]}(k)$ of the output signal $\hat{y}^{[1]}(k)$:

$$e_j^{[1]}(k) = y_j(k) - \hat{y}_j^{[1]}(k) = y_j(k) - w_j^{[1]}\mu^{[1]}(x(k))$$

(here $w_j^{[1]}$ is the j th row of the weights matrix $W^{[1]}$, $y_j(k)$ is the j th component of the reference signal $y(k) = (y_1(k),...,y_j(k),...,y_m(k))^{\mathrm{T}}$ and the standard squared learning criterion of the j th output

$$E_j^{[1]}(k) = \sum_k \left(e_j^{[1]}(k)\right)^2 = \sum_k \left(y_j(k) - w_j^{[1]}\mu^{[1]}(x(k))\right)^2. \quad (2)$$

The gradient procedure for minimizing the criterion (2) has a general form

$$\begin{aligned} w_j^{[1]}(k) &= w_j^{[1]}(k-1) - \eta^{[1]}(k)\nabla_{w_j^{[1]}} E_j(k) = \\ &= w_j^{[1]}(k-1) - \eta^{[1]}(k)\nabla_{w_j^{[1]}}\left(e_j^{[1]}(k)\right)^2 = \\ &= w_j^{[1]}(k-1) + \eta^{[1]}(k)e_j^{[1]}(k)\mu^{[1]\mathrm{T}}(x(k)) = \\ &= w_j^{[1]}(k-1) + \eta^{[1]}(k)\times \\ &\times\left(y_j(k) - w_j^{[1]}(k-1)\mu^{[1]}(x(k))\right)\mu^{[1]\mathrm{T}}(x(k)) \end{aligned} \quad (3)$$

where $\eta^{[1]}(k)$ is learning rate parameter for S_1.

It is possible to increase the speed of the learning procedure (3) using either the standard recursive least-squares method (RLSM), that is a second-order optimization procedure:

$$\begin{cases} w_j^{[1]}(k) = w_j^{[1]}(k-1) + \dfrac{e_j^{[1]}(k)\mu^{[1]\mathrm{T}}(x(k))P^{[1]}(k-1)}{1+\mu^{[1]\mathrm{T}}(x(k))P^{[1]}(k-1)\mu^{[1]}(k)}, \\ P^{[1]}(k) = P^{[1]}(k-1) - \dfrac{P^{[1]}(k-1)\mu^{[1]}(x(k))\mu^{[1]\mathrm{T}}(x(k))}{1+\mu^{[1]\mathrm{T}}(x(k))P^{[1]}(k-1)\mu^{[1]}(k)}\times \\ \quad \times P^{[1]}(k-1), \end{cases} \quad (4)$$

or the optimized algorithm with tracking and filtering properties [31,32]:

$$\begin{cases} w_j^{[1]}(k) = w_j^{[1]}(k-1) + \left(r^{[1]}(k)\right)^{-1} e_j^{[1]}(k)\mu^{[1]\mathrm{T}}(x(k)), \\ r^{[1]}(k) = \alpha r^{[1]}(k-1) + \mu^{[1]\mathrm{T}}(x(k))\mu^{[1]}(x(k)) \end{cases} \quad (5)$$

where $0 \le \alpha \le 1$ is smoothing parameter.

The algorithm (5) can be rewritten in the matrix form

$$\begin{cases} W^{[1]}(k) = W^{[1]}(k-1) + \left(r^{[1]}(k)\right)^{-1} e^{[1]}(k)\mu^{[1]\mathrm{T}}(x(k)), \\ r^{[1]}(k) = \alpha r^{[1]}(k-1) + \left\|\mu^{[1]}(x(k))\right\|^2, \end{cases} \quad (6)$$

that with $\alpha = 1$ coincides with the multidimensional version [33] of the Kaczmarz – Widrow – Hoff learning algorithm:

$$\begin{aligned} W^{[1]}(k) &= W^{[1]}(k-1) + \frac{e^{[1]}(k)\mu^{[1]\mathrm{T}}(x(k))}{\left\|\mu^{[1]}(x(k))\right\|^2} = \\ &= W^{[1]}(k-1) + e^{[1]}(k)\mu^{[1]+}(x(k)), \end{aligned} \quad (7)$$

where $(\bullet)^+$ is pseudo-inversion symbol.

It should also be noted that the Kaczmarz algorithm is optimal by speed in the class of gradient adaptive learning procedures.

All other layers $S_2,...,S_g$ are adjusted in the same way, however with the increase in the dimensionality of the vector $\mu^{[g]}(x(k))$ defined as $(h(n+(g-1)m)\times 1)$, the advantage should be given to the procedures (6), (7), since RLSM (4) can be numerically unstable at high dimensions of the input space.

V. Experiments

To demonstrate the efficiency of the proposed system, we solved the classification task for the wine data set [34]. This data set has 13 attributes, 178 instances and 3 classes of wine. We used 80% of the data set to train the system and 20% for testing. For training the Kaczmarz – Widrow – Hoff algorithm (7) was used. The results of the experiment are shown in Table I. Classes predicted by the trained system on the test set are shown in Fig. 3 as a scatter plot of the first two principal components calculated using PCA.

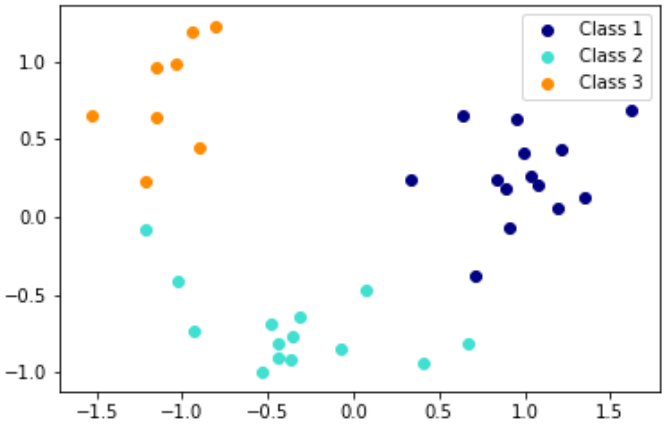


Fig. 3 Classes of wine predicted by the proposed system

VI. Conclusion

In the paper a deep evolving stacking convex neo-fuzzy network is proposed. It is a multi-layered hybrid system of computational intelligence. This network has a feedforward cascade architecture, the layers-stacks of which are formed

by generalized neo-fuzzy neurons that implement Wang–Mendel fuzzy reasoning. Since the output signals of the stacks depend linearly on the adjustable synaptic weights, the optimal in the sense of speed algorithms are used for their learning. Due to independent layer adjustment, parallelization of calculations in non-linear synapses and optimization of learning processes, the proposed network has high speed that allows to process information in online mode.

TABLE I.
RESULTS OF THE EXPERIMENTS

Number of member-ship functions	Number of cascades	Number of weights	Train accuracy by cascade		Test accuracy
5	3	720	1st	0.9648	0.9722
			2nd	0.9859	
			3rd	1.0	
7	2	609	1st	0.9859	0.9722
			2nd	1.0	
10	3	1440	1st	0.9859	0.9444
			2nd	0.9930	
			3rd	1.0	
25	1	975	1st	1.0	0.9167

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436-444, 2015.

[2] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85-117, 2015.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[4] D. Graupe, *Deep Learning Neural Networks. Design and Case Studies*. Singapore : World Scientific, 2016.

[5] A. Ivakhnenko, "The group method of data handling – a rival of the method of stochastic approximation," *Soviet Automatic Control*, vol. 13, no. 3, pp. 43-55, 1968.

[6] A. Ivakhnenko, "The group method of data handling – a rival of the method of stochastic approximation," *Automatica*, vol. 6, no. 2, pp. 207-219, 1970.

[7] N. Kasabov, *Evolving Connectionist Systems*. Springer-Verlag London, 2007.

[8] E. Lughofer, *Evolving Fuzzy Systems – Methodologies, Advanced Concepts and Applications*. Springer Berlin, 2011.

[9] G. Setlak, Ye. Bodyanskiy, O. Vynokurova, and I. Pliss, "Deep evolving GMDH-SVM-neural network and its learning for Data Mining tasks," in *Proc. 2016 Federated Conf. on Computer Science and Information Systems (FedCSIS),* Gdansk, Poland, pp. 141-145, 2016.

[10] Ye. Bodyanskiy, O. Vynokurova, I. Pliss, G. Setlak, and P. Mulesa, "Fast learning algorithm for deep evolving GMDH-SVM neural network in Data Stream Mining tasks," in *Proc. First IEEE Conf. on Data Stream Mining & Processing*, Lviv, Ukraine, pp. 318-321, 2016.

[11] A. Bifet, *Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams,* Amsterdam: IOS Press, 2010.

[12] C. C. Aggarwal, *Data Streams: Models and Algorithms (advances in database systems)*, New York: Springer, 2007.

[13] S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," in *Advances in Neural Information Processing Systems*, D. S. Touretzky Ed. San Mateo, CA : Morgan Kaufman, pp. 524–532, 1990.

[14] Y. Bodyanskiy, O. Tyshchenko, and D. Kopaliani, "A hybrid cascade neural network with an optimized pool in each cascade," *Soft Computing*, 19, №12, pp. 3445-3454, 2015.

[15] Y. Bodyanskiy, O. Tyshchenko, and D. Kopaliani, "Adaptive learning of an evolving cascade neo-fuzzy system in data stream mining tasks," *Evolving Systems*, 7, №2, pp. 107-116, 2016.

[16] T. Yamakawa, E. Uchino, T. Miki, and H. Kusanagi, "A neo-fuzzy neuron and its applications to system identification and prediction of the system behavior," in *Proc. 2nd Int. Conf. on Fuzzy Logic and Neural Networks*, pp. 477-483, 1992.

[17] E. Uchino and T. Yamakawa, "Soft computing based signal prediction, restoration and filtering," *Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms*, Boston: Kluwer Academic Publisher, pp. 331-349, 1997.

[18] T. Miki and T. Yamakawa, "Analog implementation of neo-fuzzy neuron and its on-board learning," *Computational Intelligence and Applications*, Piraeus: WSES Press, pp. 144-149, 1999.

[19] Ye. Bodyanskiy, I. Pliss, D. Peleshko, and O. Vynokurova, "Deep hybrid system of computational intelligence for time series prediction," Int. *J. "Information Theories and Applications"*, 24, №1, pp. 35-49, 2017.

[20] Ye. Bodyanskiy, O. Vynokurova, I. Pliss, D. Peleshko, and Yu. Rashkevych, "Deep stacking convex neuro-fuzzy system and its online learning," *Advances in "Intelligent Systems and Computing"*, vol. 582, Cham, Springer, pp. 49-59, 2018.

[21] Y. Bodyanskiy, G. Setlak, D. Peleshko, and O. Vynokurova, "Hybrid generalized additive neuro-fuzzy system and its adaptive learning algorithms," in *Proc. 2015 IEEE 8th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications "IDAACS 2015",* pp. 328-333, 2015.

[22] Y. Bodyanskiy, O. Vynokurova, G. Setlak, and I. Pliss, "Hybrid neuro-neo-fuzzy system and its adaptive learning algorithm," in *Proc. Int. Conf. on Computer Sciences and Information Technologies "CSIT 2015"*, pp. 111-114, 2015.

[23] Y. Bodyanskiy, O. Vynokurova, I. Pliss, D. Peleshko, and Y. Rashkevych, "Hybrid generalized additive wavelet-neuro-fuzzy-system and its adaptive learning," *Advances in Intelligent Systems and Computing*, vol. 470, Cham, Springer, pp. 51-61, 2016.

[24] Y. Bodyanskiy, O. Vynokurova, G. Setlak, D. Peleshko, and P. Mulesa, "Adaptive multivariate hybrid neuro-fuzzy system and its on-board fast learning," *Neurocomputing*, 230, pp. 409-416, 2017.

[25] Y. Bodyanskiy, O. Vynokurova, I. Pliss, and D. Peleshko, "Hybrid adaptive systems of computational intelligence and their on-line learning for green IT in energy management tasks," *Studies in Systems, Decision and Control*, vol. 74, pp. 229-244, 2017.

[26] T. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman and Hall / CRC, 1990.

[27] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, №2, pp. 241-259, 1992.

[28] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2133-2136, 2012.

[29] R. P. Landim, B. Rodrigues, S. R. Silva, and W. M. Caminhas, "A neo-fuzzy-neuron with real time training applied to flux observer for an induction motor," in *Proc. Vth Brazilian Symposium on Neural Networks*, pp. 67-72, 1998.

[30] L. Deng and D. Yu, "Deep convex net: a scalable architecture for speech pattern classification," in *Proc. of Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 2285-2288, 2011.

[31] Ye. Bodyanskiy, V. Kolodyazhniy, and A. Stephan, "An adaptive learning algorithm for a neuro-fuzzy network," Lecture Notes in Computer Science 2206, Berlin – Heidelberg – New York, Springer, pp. 68-75, 2001.

[32] P. Otto, Ye. Bodyanskiy, and V. Kolodyazhniy, "A new learning algorithm for a forecasting neuro-fuzzy network," *Integrated Computer-Aided Engineering*, vol. 10, №4, pp. 399-409, 2003.

[33] O. G. Rudenko, E. V. Bodyanskii, I. P. Pliss, "Adaptive algorithm for prediction of random sequences," *Soviet automatic control*, 12, №1, pp. 46-48, 1979.

[34] https://archive.ics.uci.edu/ml/datasets/wine

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 35–38
DOI: 10.15439/2018F160
ISSN 2300-5963 ACSIS, Vol. 15

Ranking Rough Sets in Pawlak Approximation Spaces

Zoltán Ernő Csajbók
Department of Health Informatics
Faculty of Health, University of Debrecen
Sóstói út 2-4, H-4406 Nyíregyháza, Hungary
Email: csajbok.zoltan@foh.unideb.hu

József Ködmön
Department of Health Informatics
Faculty of Health, University of Debrecen
Sóstói út 2-4, H-4406 Nyíregyháza, Hungary
Email: kodmon.jozsef@foh.unideb.hu

***Abstract*—Applying the cardinality of finite sets, interval numbers can be assigned to rough sets represented by nested sets. Borrowing two different comparison methods from Multiple Attribute Decision Making analysis, rough sets are compared and ranked on the model of interval numbers. Some special cases are investigated. Illustrative examples are presented relying on both methods. The calculated results are compared and interpreted.**

***Index Terms*—Rough sets, interval arithmetic, Possibility Degree Method, Midpoints Comparison Method.**

I. Introduction

ROUGH set theory (RST) was proposed by Pawlak in the early 1980's [1]. Information system in the Pawlak's sense can be viewed to some extent as a Multiple Attribute Decision Making (MADM) scheme (see, e.g., [2]).

In RST, rough sets represented by nested sets can be considered as an *interval set structure* to represent nonnumeric uncertainty on the model of interval numbers [3]. In our approach, however, by the cardinality of finite sets, *interval numbers* are assigned to rough sets represented by nested sets. Then, borrowing Possibility Degree Method and Midpoints Comparison Methods from MADM, rough sets can be compared and ranked numerically based on these interval numbers.

Section II presents some elementary notations for reasons of clarity. Section III and IV state fundamental knowledge about rough sets and interval arithmetic, respectively. Section V shows two comparison methods of interval numbers, namely, Possibility Degree Method and Midpoints Comparison Method. Then, it deals with the comparison and ranking of rough sets applying these two methods. It also contains simplified illustrative examples.

II. Basic Notations

Let U be a nonempty set, and $\mathcal{P}(U)$ denote the power set of U. Set operations union, intersection, difference, and complementation are denoted by $\cup$, $\cap$, $\setminus$, and ^{c}, respectively. Let $S \in P(U)$, and $\mathcal{S} \subseteq \mathcal{P}(U)$ be a nonempty family of sets. $|S|$ denotes the cardinality of S. $\cup\mathcal{S}$ and $\cap\mathcal{S}$ are defined by:

$$\cup\mathcal{S} = \{u \mid \exists S \in \mathcal{S}(u \in S)\},\ \cap\mathcal{S} = \{u \mid \forall S \in \mathcal{S}(u \in S)\}.$$

If $\mathcal{S}$ is empty, the conventions $\cup\emptyset = \emptyset$ and $\cap\emptyset = X$ are used.

The shorthand expression "iff" is used for "if and only if".

From now on, throughout the paper let U be a finite nonempty set of objects called the *universe*.

III. Rough Sets

Notions of rough set theory can be represented in many forms. For our purposes, their constructive granule based definitions [4] are formulated as follows.

Let E be an equivalence relation on U. The partition of U generated by E is denoted by U/E. The subset $[u]_E \in \mathcal{P}(U)$ is an equivalence class from U/E containing $u \in U$. The members of U/E are called *elementary sets* or simply *base sets*. Any union of base sets is referred to as *definable set*. By definition, $\emptyset$ is definable for any equivalence relation on U. Their collection is denoted by $\mathcal{D}_{U/E}(\subseteq \mathcal{P}(U))$.

The principal notions of RST are defined by:

$$\mathsf{l} : \mathcal{P}(U) \to \mathcal{D}_{U/E},\ S \mapsto \cup\{[u]_E \in U/E \mid [u]_E \subseteq S\},$$
$$\mathsf{u} : \mathcal{P}(U) \to \mathcal{D}_{U/E},\ S \mapsto \cup\{[u]_E \in U/E \mid [u]_E \cap S \neq \emptyset\}.$$

Values $\mathsf{l}(S)$ and $\mathsf{u}(S)$ are commonly called the *lower* and *upper approximations* of S. With the above notations, the ordered quintuple $PAS = \langle U, U/E, \mathcal{D}_{U/E}, \mathsf{l}, \mathsf{u}\rangle$ is called a finite Pawlak approximation space.

Having given an approximation pair, to identify and characterize the features of set approximations in RST, the following fundamental notions are defined:

- *boundary* of S is $\mathsf{bnd}(S) = \mathsf{u}(S) \setminus \mathsf{l}(S)$;
- S is *exact* (*crisp*), if $\mathsf{l}(S) = \mathsf{u}(S)$, i.e., $\mathsf{bnd}(S) = \emptyset$;
- S is *rough* (*inexact*), if it is not exact, i.e., $\mathsf{bnd}(S) \neq \emptyset$.

In RST the notions of exactness and definability coincide.

For any set S, an approximation pair divides the universe U into three mutual disjoint regions:

- $POS(S) = \mathsf{l}(S)$ — *positive region* of S;
- $NEG(S) = U \setminus \mathsf{u}(S) = \mathsf{u}^c(S)$ — *negative region* of S;
- $BN(S) = \mathsf{bnd}(S)$ — *borderline region* of S.

There are (at least) four equivalent definitions of rough sets, see, e.g., [5], [6]. In the following, the nested pair of sets $\langle \mathsf{l}(S), \mathsf{u}(S)\rangle$ will be used to represent rough sets. It is a family of inexact sets in such a way that for any $T \in \langle \mathsf{l}(S), \mathsf{u}(S)\rangle$, $\mathsf{l}(S) = \mathsf{l}(T)$, $\mathsf{u}(S) = \mathsf{u}(T)$ and $\mathsf{l}(S) \subseteq T \subseteq \mathsf{u}(S)$ hold.

Proposition III.1 ([7], Proposition 3.2) *Let $S_1 \subseteq S_2$. The pair $\langle S_1, S_2\rangle$ is a rough set of the form $\langle \mathsf{l}(S), \mathsf{u}(S)\rangle$ for a set S ($S_1 \subseteq S \subseteq S_2$) if and only if S_1 and S_2 are definable and $S_2 \setminus S_1$ does not contain any singleton base set.*

IV. Basics of Interval Arithmetic

An *interval number* or *interval* [2], [8] is a closed real interval of the form $a = [a^l, a^u] = \{x \in \mathbb{R} \mid a^l \leq x \leq a^u\}$. If $a^l = a^u$, $[a^l, a^u]$ contains a single real number $a = a^l = a^u$.

Two intervals $a = [a^l, a^u]$ and $b = [b^l, b^u]$ are said to be equal, in notation $a = b$, if $a^l = b^l$ and $a^u = b^u$.

The most common special terms for an interval a are:

- $m(a) = \frac{1}{2}(a^l + a^u)$ is the *midpoint* or *center* of a;
- $w(a) = a^u - a^l$ is the *width* or *diameter* of a.

Binary operations $+, -, \cdot, /$, addition, subtraction, multiplication, and division, respectively, can be defined on the set of intervals. Their endpoint formulae are the following [8]:

$$\begin{aligned} a + b &= [a^l + b^l, a^u + b^u];\\ a - b &= a + (-b) = [a^l - b^u, a^u - b^l], -b = [-b^u, -b^l];\\ a \cdot b &= [\min\{a^l b^l, a^l b^u, a^u b^l, a^u b^u\},\\ &\qquad \max\{a^l b^l, a^l b^u, a^u b^l, a^u b^u\}];\\ a/b &= a \cdot (1/b), \text{ where } 1/b = [1/b^u, 1/b^l] \ (0 \notin b). \end{aligned}$$

For nonnegative intervals a, b ($0 \leq a^l, b^l$), multiplication and division formulae are simplified to:

- $a \cdot b = [a^l b^l, a^u b^u]$;
- $a/b = [a^l/b^u, a^u/b^l]$, provided in addition that $0 < b^l$.

V. Comparing and Ranking Rough Sets

A. Possibility Degree Method

Many different equivalent methods have been proposed to compare two interval numbers [2], [9].

Definition V.1 ([2], Definition 4.5) *Let $a=[a^l,a^u]$, $b=[b^l,b^u]$ be two nonnegative intervals with $w(a) > 0$ or $w(b) > 0$. The* possibility degree *of $a \geq b$ is defined by*

$$p(a \geq b) = \max\left\{1 - \max\left\{\frac{b^u - a^l}{w(a) + w(b)}, 0\right\}, 0\right\}.$$

It is also said that $p(a \geq b)$ is the possibility degree of a over b.

Theorem V.2 ([2], Theorem 4.1) *Let $a = [a^l, a^u], b = [b^l, b^u]$ and $c = [c^l, c^u]$ be three nonnegative intervals. For their possibility degrees, the following properties hold:*

1) $0 \leq p(a \geq b) \leq 1$.
2) $p(a \geq b) + p(b \geq a) = 1$. *Especially,* $p(a \geq a) = \frac{1}{2}$.
3) $p(a \geq b) = 1$ *iff* $b^u \leq a^l$.
4) $p(a \geq b) = 0$ *iff* $a^u \leq b^l$.
5) $p(a \geq b) \geq \frac{1}{2}$ *iff* $a^u + a^l \geq b^u + b^l$.
   Especially, $p(a \geq b) = \frac{1}{2}$ *iff* $a^u + a^l = b^u + b^l$.
6) *If* $p(a \geq b) \geq \frac{1}{2}$ *and* $p(b \geq c) \geq \frac{1}{2}$, *then* $p(a \geq c) \geq \frac{1}{2}$.

It is said that

- a *superior to* b in the degree $p(a \geq b)$, in notation $a \succ b$, if $p(a \geq b) > p(b \geq a)$;
- a is *indifferent to* b, in notation $a \sim b$, if $p(a \geq b) = p(b \geq a) = \frac{1}{2}$;
- a is *inferior to* b in the degree $p(b \geq a)$, in notation $a \prec b$, if $p(b \geq a) > p(a \geq b)$.

Let $\{S_1, \ldots, S_n\} \subseteq \mathcal{P}(U)$ be a family of sets. Let us form the rough sets relating to them by their nested pair representations: $RS_i = \langle \mathsf{l}(S_i), \mathsf{u}(S_i)\rangle$ $(i = 1, 2, \ldots, n)$.

The cardinality of finite sets, as some sort of "size" of them, plays a key role in the rough set theory. Applying it, interval numbers can be assigned to the above rough sets:

$$RS_i \mapsto [RS_i] = [|\mathsf{l}(S_i)|, |\mathsf{u}(S_i)|] \ (i = 1, 2, \ldots, n).$$

To avoid heavy notations, the following simplified notations are introduced: $|\mathsf{l}(S_i)|$, $|\mathsf{u}(S_i)|$, $|\mathsf{bnd}(S_i)|$ are denoted by S_i^l, S_i^u, S_i^{bnd}, respectively.

By applying the method described by Xu in [2], ranking of rough sets can be carried out in the following steps:

Step 1. Provided that $w([RS_i]) > 0$ $(i = 1, \ldots, n)$, comparing each rough set with all rough sets as $(i, j = 1, 2, \ldots, n)$:

$$\begin{aligned} p_{ij} &= p([RS_i] \geq [RS_j])\\ &= \max\left\{1 - \max\left\{\frac{\mathsf{S}_j^u - \mathsf{S}_i^l}{w([RS_i]) + w([RS_j])}, 0\right\}, 0\right\}; \end{aligned}$$

arranging the numbers p_{ij}'s in a possibility degree matrix:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n}\\ p_{21} & p_{22} & \cdots & p_{2n}\\ & & \vdots & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}.$$

Of course, $p_{ij} \geq 0$, $p_{ij} + p_{ji} = 1$, $p_{ii} = \frac{1}{2}$ for $i, j = 1, \ldots, n$.

Step 2. Summing the numbers line by line:

$$p_i = \sum_{j=1}^{n} p_{ij} \ (i = 1, 2, \ldots, n).$$

Step 3. Ranking rough sets RS_i in descending (increasing) order in accordance with the values p_i's $(i = 1, 2, \ldots, n)$. The ith rough set is ranked higher (lower) than the jth rough set, if $p_i > p_j$ $(p_i < p_j)$.

B. Possibility Degree Method — A Special Case

The sets $S_1, S_2 \in \mathcal{P}(U)$ form an orthopair, if $S_1 \cap S_2 = \emptyset$. An orthopair is a reasonable means to represent bipolar information. Bipolarity arises in a natural way in RST as positive and negative regions. According to the Dubois and Prade typology [10], [11], orthopair models usually belong under the "Type II: Symmetric bivariate unipolarity". This bipolarity type well fits the nature of bipolarity representation in RST [12].

Let $\langle S_1, S_2\rangle$ be an orthopair. S_1 and S_2 are called the *positive* and *negative* reference set, respectively. Here, the positive and negative adjectives claim nothing else, only the sets S_1 and S_2 are well separated.

Let us form the rough sets relating to S_1, S_2 by their nested pair rough set representations:

$$RS_1 = \langle \mathsf{l}(S_1), \mathsf{u}(S_1)\rangle \text{ and } RS_2 = \langle \mathsf{l}(S_2), \mathsf{u}(S_2)\rangle.$$

By the above Steps 1–3, the following entities can be obtained with which the constituents of an orthopair can be ranked:

$$\begin{aligned} p_1 &= p_{11} + p_{12} = p([RS_1] \geq [RS_1]) + p([RS_1] \geq [RS_2]),\\ p_2 &= p_{21} + p_{22} = p([RS_2] \geq [RS_1]) + p([RS_2] \geq [RS_2]). \end{aligned}$$

Several interpretations of the obtained results can be stated:

- $p_1 > p_2$ ($p_1 < p_2$) means that the positive (negative) reference set is ranked higher than the negative (positive) reference set.
- $p([RS_1] \geq [RS_2]) = 1$ iff $\mathsf{S}_2^u \leq \mathsf{S}_1^l$.
  It means that the positive reference set is certainly superior to the negative reference iff the number of elements of U which can *possibly be classified as belonging to the negative reference set* is less than or equal to the number of elements of U which can *certainly be classified as belonging to the positive reference set*.
- $p([RS_1] \geq [RS_2]) = 0$ iff $p([RS_2] \geq [RS_1]) = 1$ iff $\mathsf{S}_1^u \leq \mathsf{S}_2^l$.
  It means that the negative reference set is certainly superior to the positive reference set iff the number of elements of U which can *possibly be classified as belonging to the positive reference set* is less than or equal to the number of elements of U which can *certainly be classified as belonging to the negative reference set*.
- $p([RS_1] \geq [RS_2]) = \frac{1}{2}$ iff $\mathsf{S}_1^u + \mathsf{S}_1^l = S_2^u + \mathsf{S}_2^l$ iff $\mathsf{S}_1^u - \mathsf{S}_2^u = \mathsf{S}_2^l - \mathsf{S}_1^l$. Let $\mathsf{S}_1^u - \mathsf{S}_2^u = \mathsf{S}_2^l - \mathsf{S}_1^l = K$.
  $K = 0$ means that the possibility degree of the positive reference set over the negative reference set is equal to $\frac{1}{2}$, iff the number of elements of U which can *possibly* be classified as belonging to the positive and negative reference sets, respectively, are equal, and, at the same time, the number of elements of U which can *certainly* be classified as belonging to the positive and negative reference sets, respectively, are also equal.
  Similar interpretations can be made for $K > 0$ and $K < 0$.

C. Possibility Degree Method — Illustrative Examples

These examples deal with studying the symptoms of thyroid dysfunctions. Although the problem emerged in Csajbók et al. [13], a substantially different solution is presented here.

Thyroid dysfunction diagnosis via clinical symptoms is an important problem [14]. We deal with only hypothyroidism and hyperthyroidism thyroid disorders [15]. The thyroid gland produces thyroid hormone. Hyperthyroidism occurs when the thyroid gland is "overactive", i.e., releases too much hormone, whereas hypothyroidism takes place when the thyroid gland is "underactive", i.e., does not produce enough hormone.

Let us consider a data table given in Table I, taken from [13]. It contains clinical symptoms which may indicate that someone, a patient, develops hypothyroidism or hyperthyroidism, perhaps neither of them. There are, of course, more symptoms of hypothyroidism and hyperthyroidism, but the example has been simplified here for illustrative purposes.

Clinical symptoms which are taken into account are the following: Weight change, Edema, Tachycardia, Increased sweating, Mood. Hypothyroidism and hyperthyroidism can accurately be diagnosed with laboratory tests. The last two columns in Table I are based on these results.

In the example, the universe U is a set of clinically observed patients: $U = \{P_1, P_2, P_3, P_4, P_5\}$. Let $S_1 = \{P_2, P_3\}$ and $S_2 = \{P_4, P_5\}$ be the sets of patients who demonstrably suffer from hypothyroidism and hyperthyroidism, respectively.

Example V.3 *If the column "Weight change" is chosen, the universe U can be partitioned into $\{P_1, P_5\}$, $\{P_2, P_3\}$, and $\{P_4\}$, reflecting the weight change being "no change", "gain", "loss", respectively. Then, based on this partition,*

$$\mathsf{l}(S_1) = \{P_2, P_3\},\ \mathsf{u}(S_1) = \{P_2, P_3\},\ \textit{i.e.,}\ [RS_1] = [2, 2];$$
$$\mathsf{l}(S_2) = \{P_4\},\ \mathsf{u}(S_2) = \{P_1, P_4, P_5\},\ \textit{i.e.,}\ [RS_2] = [1, 3].$$

Since $2+2 = 1+3$, $p([RS_1] \geq [RS_1]) = \frac{1}{2}$, by Theorem V.2, (5)*. That is $[RS_1]$ is indifferent to $[RS_2]$. It can be interpreted as follows: with respect to our knowledge represented in Table I and partitioning U by "Weight change", weight change does not contribute specifically to developing any of hypothyroidism and hyperthyroidism.*

Example V.4 *If the columns "Edema" and "Mood" are chosen, the universe U can be partitioned into $\{P_5\}$ and $\{P_1, P_2, P_3, P_4\}$, reflecting the edema and mood being "Edema = yes", "Mood = nervousness" and "Edema = no", "Mood = no", respectively. Then, based on this new partition,*

$$\mathsf{l}(S_1) = \emptyset,\ \mathsf{u}(S_1) = \{P_1, P_2, P_3, P_4\},\ \textit{i.e.,}\ [RS_1] = [0, 4];$$
$$\mathsf{l}(S_2) = \{P_5\},\ \mathsf{u}(S_2) = \{P_1, P_2, P_3, P_4, P_5\},\ \textit{i.e.,}\ [RS_2] = [1, 5].$$

With a simple calculation, we have

$$p([RS_1] \geq [RS_2]) =$$
$$= \max\left\{1 - \max\left\{\frac{\mathsf{S}_2^u - \mathsf{S}_1^l}{w([RS_1]) + w([RS_2])}, 0\right\}, 0\right\} = \frac{3}{8},$$
$$\textit{and } p([RS_2] \geq [RS_1]) = 1 - p([RS_1] \geq [RS_2]) = \frac{5}{8}.$$

These results can be interpreted as follows: with respect to our knowledge represented in Table I and partitioning U by "Edema" and "Mood", the overall contribution of the clinical symptoms edema and mood to the presence of

- *hypothyroidism has the possibility degree $\frac{3}{8}$,*
- *hyperthyroidism has the possibility degree $\frac{5}{8}$.*

D. Midpoints Comparison Method

In **Theorem V.2**, properties (3) and (4) mean that the possibility degree of a over b is equal to 0 or 1 iff they do not have a common area *regardless of the distance between a and b.*

To overcome this problem, Dymova et al. [16] proposed a method to measure the distance between intervals which, in addition, also indicates which interval is greater/lesser.

Let $a = [a^l, a^u]$, $b = [b^l, b^u]$ be two intervals and form their subtraction: $c = a - b = [c^l, c^u] = [a^l - b^u, a^u - b^l]$. Clearly, $c^l \leq 0$ and $c^u \geq 0$, if a and b overlap each other.

Then, the proposed distance measure between a and b is:

$$\Delta(a, b) = \frac{1}{2}\left(\left(a^l - b^u\right) + \left(a^u - b^l\right)\right) = m(a) - m(b).$$

That is, $\Delta(a, b)$ is simply the difference of the midpoints of a and b. This immediately implies that for intervals a and b with common midpoints, $\Delta(a, b) = 0$ holds.

Remark V.5 *It may seem that the measure $\Delta(a, b)$ is too simple. For its discussion, see [16]. In addition, on the important role of midpoints in comparison of intervals, see [17].*

TABLE I
CLINICAL SYMPTOMS OF THYROID DYSFUNCTION AND DIAGNOSIS BASED ON TEST RESULTS

No.	Weight change	Edema	Tachy-cardia	Increased sweating	Mood	Hypothy-roidism	Hyperthy-roidism
P_1	no change	no	no	no	normal	no	no
P_2	gain	no	no	no	normal	yes	no
P_3	gain	no	yes	no	normal	yes	no
P_4	loss	no	yes	yes	normal	no	yes
P_5	no change	yes	no	yes	nervousness	no	yes

E. Comparing the Two Methods

In [16], *experimental observations* show that the sign of $\Delta(a,b)$ is positive (negative), if $a \succ b$ ($a \prec b$). In addition, $abs(\Delta(a,b))$ is close to the Hamilton distance d_H and Euclidean distance d_E of the intervals a and b, where

$$d_H = \frac{1}{2}\left(abs\left(a^l - b^l\right) + abs\left(a^u - b^u\right)\right),$$
$$d_E = \frac{1}{2}\sqrt{\left(a^l - b^l\right)^2 + \left(a^u - b^u\right)^2}.$$

In regard to these experimental observations, let us compare our numerical results which were calculated with the help of the possibility degree method and midpoints comparison method.

S_1, S_2 are the sets of patients who demonstrably suffer from hypothyroidism and hyperthyroidism, respectively.

According to **Example V.3**, $[RS_1] = [2,2]$; $[RS_2] = [1,3]$, where RS_1, RS_2 are the rough sets concerning S_1, S_2 and based on the partition of U formed by "Weight change".

By applying the possibility degree method, $p([RS_1] \geq [RS_2]) = \frac{1}{2}$, i.e., $[RS_1]$ is indifferent to $[RS_2]$.

By applying the midpoints comparison method, the intervals $[RS_1]$, $[RS_2]$ are equal, i.e, $\Delta([RS_1],[RS_2]) = 0$, since their midpoints are equal. Of course, the sign rule does not work here.

The one interpretation is in accordance with the other.

According to **Example V.4** $[RS_1] = [0,4]$; $RS_2] = [1,5]$, where RS_1, RS_2 are the rough sets concerning S_1, S_2 and relying on the partition of U formed by "Edema" and "Mood".

By applying the possibility degree method, $p([RS_2] \geq [RS_1]) = \frac{5}{8}$, i.e., $[RS_1]$ is inferior to $[RS_2]$, $[RS_1] \prec [RS_2]$, in the degree $\frac{5}{8}$.

By applying the midpoints comparison method,

$$\Delta([RS_1],[RS_2]) = m([RS_1]) - m([RS_2]) = 2 - 3 = -1.$$

According to the sign rule of the midpoint comparison method, since the sign of $\Delta([RS_1],[RS_2])$ is negative, $[RS_1]$ is lesser than $[RS_2]$. This result coincides with the result $[RS_1] \prec [RS_2]$ obtained by the possibility degree method.

If the midpoints of two intervals are the same, there is no sense in comparing $abs(\Delta([RS_1],[RS_2]))$ with the Hamilton and Euclidean distances. This is the case in **Example V.3**.

In **Example V.4**, $abs(\Delta([RS_1],[RS_2])) = 1$. In this case, Hamilton and Euclidean distances can be calculated. For $[RS_1] = [0,4]$, $[RS_2] = [1,5]$, $d_H = 1$ and $d_E = \frac{\sqrt{2}}{2} \approx 0,71$. The Hamilton distance is the same as $abs(\Delta([RS_1],[RS_2]))$, and Euclidean distance estimates it to some extent.

VI. CONCLUSION

The paper has presented two comparison and ranking methods for rough sets in Pawlak approximation spaces. Although the two methods are borrowed from Multiple Attribute Decision Making analysis, their application to rough sets is a new approach. Based on the presented calculations and interpretations, it seems that this approach deserves attention.

Acknowledgement The authors would like to thank the anonymous referees for their useful comments and suggestions.

REFERENCES

[1] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.
[2] Z. Xu, *Uncertain Multi-attribute Decision Making. Methods and Applications*. Springer-Verlag, 01 2015. ISBN 978-3-662-45640-8
[3] S. K. M. Wong, L. Wang, and Y. Y. Yao, "Interval structure: A framework for representing uncertain information," *CoRR*, vol. abs/1303.5437, 2013.
[4] Y. Y. Yao, "On generalizing rough set theory," in *Proceedings of RSFDGrC 2003*, ser. LNAI, G. Wang, Q. Liu, Y. Yao, and A. Skowron, Eds., vol. 2639. Springer, 2003, pp. 44–51.
[5] M. Banerjee and M. Chakraborty, "Algebras from rough sets," in *Rough-Neuro Computing: Techniques for Computing with Words*, S. Pal, L. Polkowski, and A. Skowron, Eds. Springer, 2004, pp. 157–184.
[6] Z. Bonikowski, "A certain conception of the calculus of rough sets," *Notre Dame Journal of Formal Logic*, vol. 33, no. 3, pp. 412–421, 1992.
[7] V. W. Marek and M. Truszczyński, "Contributions to the theory of rough sets," *Fundam. Inf.*, vol. 39, no. 4, pp. 389–409, 1999.
[8] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 2009. ISBN 0-89871-669-1
[9] G. Facchinetti, R. Ghiselli Ricci, and S. Muzzioli, "Note on fuzzy triangular numbers," vol. 13, pp. 613 – 622, 07 1998.
[10] D. Dubois and H. Prade, "An introduction to bipolar representations of information and preference," *International Journal of Intelligent Systems*, vol. 23, no. 8, pp. 866–877, 2008. doi: 10.1002/int.20297
[11] ——, "An overview of the asymmetric bipolar representation of positive and negative information in possibility theory," *Fuzzy Sets and Systems*, vol. 160, no. 10, pp. 1355–1366, 2009. doi: 10.1016/j.fss.2008.11.006
[12] D. Ciucci, "Orthopairs: A simple and widely used way to model uncertainty," *Fundam. Inf.*, vol. 108, no. 3-4, pp. 287–304, 2011.
[13] Z. Csajbók, T. Mihálydeák, and J. Ködmön, "An adequate representation of medical data based on partial set approximation," in *Proceedings of CISIM 2013*, ser. LNCS, K. S. et al., Ed., vol. 8104. Berlin Heidelberg: Springer-Verlag, 2013, pp. 108–116.
[14] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 944–949, 2009.
[15] P. Ladenson and M. Kim, "Thyroid," in *Goldman's Cecil Medicine*, L. Goldman and A. I. Schafer, Eds. Philadelphia, Pa: Saunders Elsevier, 2011, p. Chap. 233.
[16] L. Dymova, P. Sevastjanov, and A. Tikhonenko, "A direct interval extension of topsis method," *Expert Systems with Applications*, vol. 40, no. 12, pp. 4841 – 4847, 2013. doi: https://doi.org/10.1016/j.eswa.2013.02.022
[17] Y.-M. Wang, J.-B. Yang, and D.-L. Xu, "A preference aggregation method through the estimation of utility intervals," *Comput. Oper. Res.*, vol. 32, no. 8, pp. 2027–2049, Aug. 2005. doi: 10.1016/j.cor.2004.01.005

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 39–45
DOI: 10.15439/2018F268
ISSN 2300-5963 ACSIS, Vol. 15

Automatic Assessment of Student Understanding Level using Virtual Reality

Shota Hashimura
Graduate School of Information Science and Engineering
Ritsumeikan University, Shiga, Japan
Email: hashimura@de.is.ritsumei.ac.jp

Hiromitsu Shimakawa, Yusuke Kajiwara
College of Information Science and Engineering
Ritsumeikan University, Shiga, Japan
Email: {simakawa, kajiwara}@de.is.ritsumei.ac.jp

***Abstract*—The improvement of the efficiency in teaching requires knowing the understanding level of each student. However, it is difficult due to limited time in a class. We propose a Virtual Reality (VR) space imposing assignments on students, to know their understanding level from their behavior which comes from cognitive loads during their answering. The VR space presents a student an assignment and a working space to answer it. In general, students solve assignments, using elements on their short term memory. When students solve same kind of assignments many times, they build generalized solution methods in their long term memory. When they engage in such assignments, their cognitive load is low enough to make them watch only the working spaces, keeping their hands working. On the other hand, when students have no solution pattern, their short term memory works hard. Their high cognitive load often stop their hands, because of confusion. They also look assignments and the working space many times, to reconsider solutions. Since answering behavior of students exposes their cognitive load, a VR space is ideal to estimate cognitive load. We conducted an experiment to evaluate the ability of the method to estimate the cognitive load. We examined the movement of the hand and the edit distance of student's answer from the correct sentence during their answering. We confirmed a fair correlation of the hands' stagnation with the confidence in students of good scores. We also found a relationship of eye movement with the change of the edit distance. The experiment result implies the possibility to estimate the cognitive load. The estimation would enable teachers to know students' understanding faults, which leads to education according to the understanding level.**

I. Introduction

IN EDUCATION, adjusting the difficulty of tasks in the class maximizes the learning effect[1][2]. Students cannot learn anything from too difficult tasks, nor anything from too easy tasks. Therefore, teachers try to adjust the progress of classes. For the adjustment, they need to know the cognitive load on students. In the context, the cognitive load in learning refers to the total amount of mental activity imposed on students. In face to face class, teaches can estimate the load, looking their behaviors. Teachers can also ask questions to students, to receive feedbacks from students. Such direct communications is the best way to know the cognitive load on students, but it takes so long time to communicate with large number of students. To avoid it, teachers assign students paper tests or e-learning tests instead of direct communication. Although these tests can check understanding level of many students at one time, teachers cannot know behavior of students from these tests. These tests cause miss-understanding of the cognitive load. One example is a correct answer by luck. In addition, it is a hard work for students and teachers to perform tests many times. It is troublesome to adjust the difficulty level of the class for students. We must find the easy way to estimate student understanding level correctly.

There are two types of memories in human brain: a working memory and a long term memory. Each of the memories has its own functions. We focus on difference of these functions to estimate understanding level. The function of the working memory is information processing to understand situations and carry out tasks. Its capacity is limited[4] and the memory is lost within about 20 s[5]. On the other hand, the long term memory has large size, to store patterns which are often used in the processing in the working memory. Each of the patterns is treated as one chunk, when it is restored from the long term memory to the working memory. The patterns are referred to as schemata[6].

To solve tasks which are not mastered well, students need to process information without schema. It is hard for students, because the working memory should store a lot of information at the same time[7]. Since the capacity of the working memory is small, the tasks make the cognitive load high. By contrast, when students master the tasks through repeated practices, they restore the schemata corresponding to them from the long term memory to the working memory. Since schemata combine several pieces of information into one chunk, they help students reduce the number of pieces on the working memory. Consequently, schemata reduce the cognitive load[8][9][10].

When students make mistakes and show hesitation in learning tasks, they seem not to have established schemata on the knowledge to achieve the tasks. They seem to have high cognitive load caused by a lot of information on the working memory. Based on the idea, this study proposes a method to estimate student understanding level correctly from their behaviors to answer tasks in learning using a VR space. In this study, we utilize tests to sort English words in a VR space. In a VR space, we can record detailed behavior such as gaze shifts and hand movements. We analyze the behavior along with test results, to estimate their cognitive load. The estimation reveals student understanding level. In addition, this paper discusses a way to examine what part of the learning task imposes the

high cognitive load on the students, which enables us to find what knowledge they lack.

In this paper, section Ⅱ explains the relationship of operation in a VR space with cognitive load. Section Ⅲ clarifies the method to figure out the understanding level from behavior. Section Ⅳ presents an asessment system of understanding level using a VR tool. Section V evaluates the method by an experiment. Section Ⅵ discusses the result of the experiment. Section Ⅶ concludes our works.

II. Relationship of operation in VR space with cognitive load

A. Cognitive load

When people understand matters, their brains memorize the information on the matter. It imposes the loads on their brain. The load is referred to as cognitive load. People cannot understand the matter without memorizing it. Some people take things as they are. Other people connect the related things as single facts before they remember them. It is good to group the related facts as single ones, in order to reduce the load of the memories. People would memorize the pattern of related things which are often used on the memories. The patterns of related things and processing results are treated as chunks[11] .

There are two types of human memory: the working memory and the long term memory. Each memory is specialized for their role.

The working memory stores the information temporarily to process it. People must store all information on tasks in their working memory to achieve them[2][3]. Nevertheless, the capacity of the working memory is small[4]. It is reported the working memory only can store around four pieces of information even in the case of young adults[12].

By contrast, the long term memory has large capacity. The long term memory stores the pattern of thinking and relationship of information. It is referred to as a schema. A schema can combine several chunks as a bigger chunk. The combination makes the load of the working memory smaller[8][9][10]. For example, let us assume to remember a sequence of six letters of “MEMORY”. If a child who does not know English tries to remember this alphabet sequence, the child has to memorize each character like ’M’, ’E’, ’M’, ’O’, ’R’, ’Y’. On the other hand, if you know English word “MEMORY”, you can combine that information as one chunk, which reduces the burden on the working memory. The difficulty of a learning task depends on the cognitive load, while the cognitive load is determined by knowledge of students. Students who have appropriate knowledge to solve questions can decrease the number of chunks in the working memory. Therefore, the cognitive load is also reduced.

Students can learn no knowledge from too difficult assignments, because they cannot proceed the task. In the same way, too easy assignments give no knowledges to students, because there are no new things for them. Estimation of the cognitive load can change the teaching, because we can adjust the difficulty for each student to maximize the effect of learning.

B. Human sense in VR space

In VR space, the movement of users is measured to make the users feel they move in the space, as if they move in the real space. The movement is measured with 2 wearable devices: a head mount display, and handy mortion controllers.

The head mount display, which is used to display the VR space to users, measures the position and the rotation of the user, to display the virtual space naturally to the user. The space presented inside the head mound display changes according to the head movement, so that the users take their views just like in their ordinal life. They can see anything in a VR space from any position and any direction in the way they want. For this reasons, users can take three important factors to feel reality, 3D spatiality, real time responsibility and self-projecting[13].

Recent VR can detect hands movement using handy motion controllers. Due to the motion controllers, users can interact with VR objects. We can know quickly the detailed position and rotation of the motion controllers. We can also detect the grasping of users. Therefore, we can reproduce their hands in VR space, through the projection of the virtual hand models on the position where the users feel their real hands are placed. In addition, the device enables us not only to rotate the virtual hands as the actual hands rotate, but also to bend the virtual fingers as the real fingers. Since the movement of virtual hands is identical with real ones, the reality is provided wiht users.

Suppose students engage in learning tasks in in a VR classroom. Using wearable devices, we can detect the movement of the gaze and the hand of the students. Users can interact with virtual objects without operative difficulties, because they can grasp virtual objects as they do in the real life. The records expressing their behaviors contain few noises. The movement shows their hesitation, confidence and cognitive load in the learning tasks. It shows purely their understanding level.

C. Related works

Many studies try educational data mining[14]. For example, Ivancevic, Celikovic & Lukovic find the seats selection of students in classroom, which is related to their assessment[15].

There are some studies to reveal the understanding level of e-learning students. As one example, Nakamura[16] used a camera to analyze the facial movement. His team succeeded estimating 75% of the subjective difficulty of the students from their facial behavior. However, facial behavior depends on individuals. The method requires a specialized estimator for each student. It has also a problem to record the behavior of the students by a camera, from the viewpoint of privacy.

Eye gaze is used to know focus and attention of users[17]. In the case of VR, eye gaze is one of the pointing way to interact virtual objects without hand interaction. Some of head-mount displays can detect gaze of users(e.g. FOVE[18]). However, they are expensive. Since most of them cannot detect the gaze, they use head-based interactions as a proxy of gaze pointing[19][20]. This pointing is an operation, which is a conscious behavior. To estimate the students understanding, it

is also important to analyze the unconscious behavior such as taking a look at hints.

We need a method to estimate each student understanding level. To do this, we should find predictors which explain the understanding level of individual students.

III. Clarifying Understanding Level from Behavior

A. Recording behavior consisting of small tasks

We propose the method which estimates student understanding level from their behavior. A task of a specific student consists of small tasks. When the student finishes each of the small tasks, the main task is over too. We focus on these small tasks such as looking at a question and picking up a word card. We record the order of these small tasks as behavior, which is the target data to be analyzed.

After teachers estimate the understanding level from behavior, they can improve their classes with the records so that the classes suit for the student understanding.(Fig. 1)

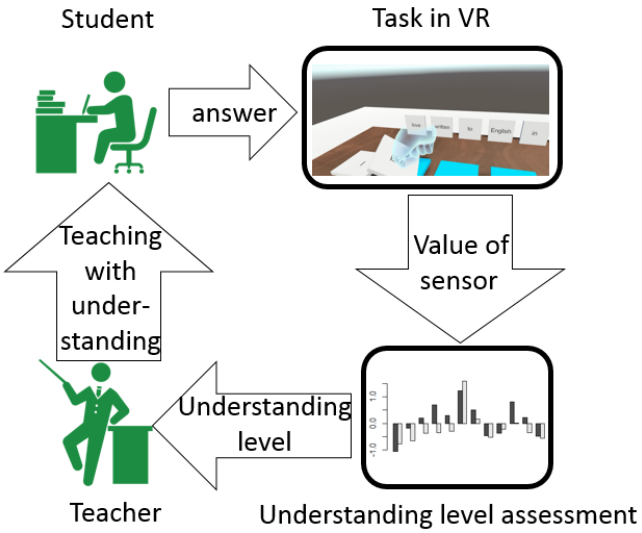


Fig. 1. Operation of proposal method

B. Estimating understanding level of each student

This paper takes an example which is a sort test of English words to explain the way to assess student understanding level. The tests are conducted in a VR space. Every student who tries the test wears a head mount display and hand gears specific to the VR space. The pair of the head mount display and the hand gears enables the student to experience the sort test provided in the VR space. At the same time, the pair is equipped functions to record the movement of the head and the both hands of the student.

We record movement of the hands and the head of the student in VR test (Fig. 2). When students have low cognitive load, the movement of their hands and head is smooth. On the other hand, when students have high cognitive load, the movement often stops. Our research estimates the cognitive load from the movement. Furthermore, we reveal what students do not understand through analysis of test results and the cognitive load, which is estimated from their behavior.

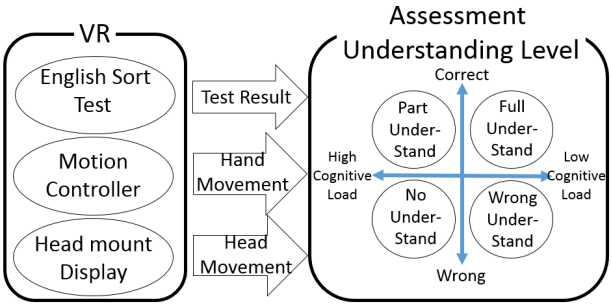


Fig. 2. Method overview

C. Assessment of cognitive load

We can estimate the student understanding level with the analysis of the cognitive load, deeper than that without behavior. We quantify the cognitive load based on machine learning. Since cognitive load is phenomenon in brain, it is not observable. It is difficult to quantify. However, students who solve questions with low cognitive load have high confidence, while high cognitive load makes confidence low. We can know the cognitive load from confidence of students appearing on their behavior. We examine the following two relationships to assess the cognitive load.

(1) Relationship between the hand movement and the confidence
As we mentioned, high confidence is assumed to make hands movement smooth. We tried to confirm this assumption.

(2) Relationship between the gaze movement and the edit distance
If students have high confidence, they do not look same place repeatedly. This gaze movement seems to be related to edit distance, which represents the number of how many correction is needed to finish the English sort test. In the test, we assume the more the edit distance is, the higher the cognitive load gets, because it is more difficult for the students to image correct sentences. To make the assumption confirmed, we examine the relationship between the gaze movement and the edit distance.

IV. Assessment of Understanding Level using VR

A. VR English sort test

We prepared a VR English sort test to know what behavior students show while they engage is the test. This test is constructed using Unity5.6.2 p2. Fig. 3 shows a snapshot of a participant sight when a student takes the test. The movie of experiment is located at www.de.is.ritsumei.ac.jp/publication/englishtest.mp4

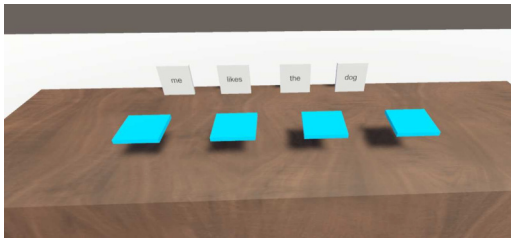


Fig. 3. Sight of students in VR test

In the experiment, participants move the words to correct order of sentence. The blue squares are columns for answer. Participants grab word cards by their hands to place them on those columns. As a hint, a Japanese sentence is presented above the word cards. The participants should look up the hint. When they look up and look around, the head mounted display they wear sense its position and direction. The data show the participant gaze movement which include the information of their understanding level.

B. Head movement and gaze movement

We used Oculus Rift, a head mount display. The participants engage in the test, wearing it on their heads. They usually look at a specific item in the VR space, to move the item from one place to another. For the participants, this behavior in the VR space is more natural than their behavior while they engage in conventional e-learning using a mouse to scroll pages. This head movement is related to the gaze movement. It implies the difficulty the participants have. Since students who have enough knowledge can imagine the correct English sentence consisting of the given words, they can solve the test without looking the hint many times. On the other hand, students fail to imagine the correct one try to make their answers, looking the hint many times and gazing many word cards repeatedly.

C. Hand movement

We used Oculus Touch, the motion controller. When users wear a pair of the motion controllers on their hands, virtual hands appear in VR space. The virtual hands move in the VR space according to that of the hands of users. Using the motion controllers, the users can grab and put things in the VR space by their hands. Since the movement of the virtual hands is coincident with that of the user hands, the VR test is better than conventional e-learning in terms of the ease of control. Because of the ease, the hand movement which students take while VR test contain information which achieves clear analysis of the student understanding level.

V. Experiment

A. Experiment contents and purposes

We experiment on the confidence of participants for a sort test of English words in the VR space.

There are two purposes in this experiment. One is to confirm the relationship between the movement of hands and the confidence. The other purpose is to examine the possibility to assess the confidence from their gaze movement.

In the experiments, 11 college students took assignments to sort English words in the VR space. We conducted 2 kinds of experiment. The first one provides 12 assignments taken by four students. The second one provides 15 assignments taken by 7 students. Some assignments are difficult, while others are easy. We recorded the movement of their heads and right hands every 20 milliseconds. Every time the students finish assignments, we asked their confidence for their answer. In the first experiment, they evaluate their confidence on a scale of 1 to 5. In the second one, they used 4 grade evaluation. No students took both assignments.

B. Relationship of hand movement with confidence

We calculated the ratio of the time in which the hand is stopped to the whole answering time, where the stop of the hand means the sum of the absolute values of the hand location change in x, y, and z directions in 20 milliseconds is less than 2 millimeters. We examined the correlation between the ratio and the confidence.

Unfortunately, we experienced data missing for 2 students. Excluding the two student data, we analyzed the correlation of 9 students. The results are shown in Table I.

In the results, the correlation coefficient goes below -0.7 for seven of the nine students. It means 78% students have low confidence when they hold their hand in assignments. We confirmed the correlation coefficient between their confidence and the hand movement, which means we can estimate the confidence of students from their hand movements.

The two students who were low in the correlation got poor scores in the assignments than others. These students may have answered in the assignments without deep considerations.

C. Relationship between gaze movement and edit distance

Students who have low confidence to solve the assignments would repeat to look at the hints and the word cards alternatively. They are likely to show many gaze shifts which come from the alternative looking. On the other hand, students who have enough knowledge can imagine the correct sentences. They can solve the test without such an unnecessary gaze shift. Unnecessary gaze shifts of students seem to be related their indecision. We detected the amount of unnecessary gaze shifts by the method explained in Fig 4.

During the assignment, every student would look at the hint to determine a specific card word to grasp. Every 20 milliseconds, the method identifies the direction the head faces. It is calculated from the rotation of the head mount display. The method also figures out the vector which corresponds to the direction to the word card the student grabbed next. This vector starts from the student head, and reaches to the word card. The method calculates the cosine similarity of two vectors. When the direction of the two vectors is identical, the cosine similarity takes 1.0, the highest value. It decreases, as the student gazes items located in other direction than that of the word card. We calculated the amount of unnecessary

TABLE I
RELATIONSHIP OF HAND MOVEMENT WITH CONFIDENCE

Students	Correlation Coefficient	Number of correct answers
A	-0.806	8/12
B	-0.91	8/12
C	-0.911	7/12
D	-0.568	5/12
E	-0.897	11/15
F	-0.704	10/15
G	-0.845	9/15
H	-0.765	6/15
I	-0.380	5/15
Mean	-0.754	

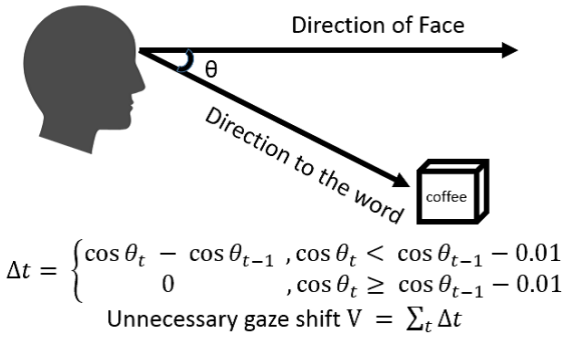


Fig. 4. How to detect a unnecessary gaze shift

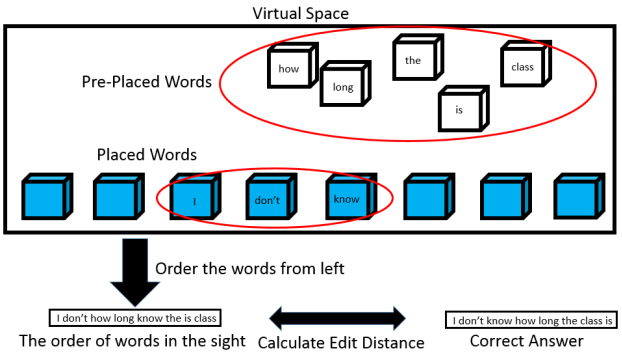


Fig. 5. Calculation of edit distance between correct answer and order of words in sight

gaze shifts as the sum of decline of the cosine similarity from the grasp of one word card to another. We ignore the difference of the cosine similarity more than -0.01 during 20 milliseconds, because it means the student gaze the same item in the duration.

When students lack enough knowledge, they would modify the word order in their answers many times, because they have poor confidence. The frequent modification increases the edit distance of their answers against the correct answers. We tried the multiple regression analysis, where the edit distance is the response variable, while the explanatory variables are the amount of gaze shifts and the time for the students use to answer. The result is showed in Table II. All students except the student E have higher P-value than 0.05. Moreover, the adjusted R-squared coefficients are very low. It means there are no relationship the gaze shift and the edit distances.

After the experiment above, we reconsider the estimation model. We review processes in which students answer assignments. Putting words one by one, students would make sentences. In the process, students would look at not only placed words but also pre-placed words the students are required to be placed in the right order. Therefore, when we estimate the cognitive load, we should consider all words, though we addressed only placed words in the previous experiment.

We re-calculate the edit distance with all words including the placed words and pre-placed words(Fig. 5). We tried the multiple regression analysis to fit this new edit distance with the amount of gaze shifts and the answer time. The result is showed in Table III. The P-value is improved to below 0.05 for about 5 of 7 students. Moreover, the adjusted R-squared coefficient is also improved, although its average is still around 0.04.

VI. DISCUSSION

A. Assessment the confidence from the behavior

The negative correlation became weaker in two of nine students. They answered fewer correct answers. After the test, we interviewed these students. This interview revealed that one of them is likely to express low confidence even for assignments he quickly gave correct answers. The student may lose confidence than usual, because of successive high-level assignments. However, the other student often expresses high confidence even for wrong answers. The mismatches of the confidence from the answers seem to be caused by lack of knowledge. That is called Dunning-Kruger effect[21]. When the test is too difficult for them, they cannot understand what is correct, which makes them misunderstand the difficulty of questions. Our method cannot estimate the confidence whose scores are low. However, we can find the confidence of students who scored well. In other words, we can choose students suitable to estimate their confidence from their scores.

The adjusted-R squared was around 0.04 in the result of multivariate regression to fit the edit distance of all words,

TABLE II
RESULT OF MULTIPLE REGRESSION ANALYSIS

students	Unnecessary Gaze Shift		Thinking time		Adjusted R-squared
	Coefficient	P-value	Coefficient	P-value	
E	0.09769	0.025	0.10198	0.019	0.03112
F	-0.01961	0.667	0.06203	0.174	0.00739
G	-0.00121	0.977	-0.0591	0.163	0.00331
H	-0.00632	0.855	0.00354	0.918	-0.00585
I	0.04637	0.364	-0.00804	0.875	-0.00427
J	0.07278	0.099	0.08668	0.05	0.01497
K	-0.00632	0.855	0.00354	0.918	-0.00585
Mean	0.02292	0.48	0.02382	0.389	0.0051

TABLE III
RESULT OF MULTIPLE REGRESSION ANALYSIS BY CONSIDERING ALL WORDS

Students	Unnecessary Gaze Shift		Thinking Time		Adjusted R-squared
	Coefficient	P-value	Coefficient	P-value	
E	2.5853	0.00003	0.7614	0.21	0.08793
F	1.7009	0.022027	1.2457	0.0922	0.02405
G	0.9263	0.117	0.3426	0.561	0.002795
H	0.6490	0.1487	0.4752	0.2900	0.000610
I	2.3578	0.000910	2.9850	0.00003	0.1043
J	2.758	0.000635	1.065	0.179976	0.06258
K	1.0168	0.03497	1.8697	0.00012	0.04157
Mean	1.4993	0.04632	1.0931	0.060032	0.040479

using the amount of gaze shifts and the answer time. It means there are still hidden explanation variables. However, the p-value of the gaze shifts goes below 0.05. Therefore, the amount of gaze shifts certainly is one of the explanation variables for the edit distances. Note that the p-value and adjusted-R squared for the edit distance calculated from all words are better than those for the edit distance calculated from only placed words. When students answer the sort test of English, they do not stop thinking about words which have already been placed to reduce cognitive load. They would be conscious of all words to find the best combination of them.

In this experiment, we focused on movement students took to put each words. However, it seems students answers are influenced by the flow of sentences. It is expected to improve the cognitive load estimation with consideration of each chunk which is formed based on English grammar.

B. Lack of knowledge revealed with cognitive load

The result of this study shows that we can estimate the confidence of students from analysis of their behavior. Since the confidence is influenced by the cognitive load, we can estimate the cognitive load by looking behavior of students.

As we discussed in section V.A, when students answer the word sort test of English, they do not stop thinking of words they have already placed to reduce their cognitive load. They keep being conscious of all words, to find the best combination of them. Considering all words is hard work, which arises high cognitive load. Therefore, they try to combine words as chunks. The chunks correspond to confidential parts in their answer. Since each chunk occupies only one working memory, they can decrease the load of memory. Students can store all words in the assignment on the working memory, utilizing chunks. At that time, they can solve it.

Students seem to divide the whole task into small tasks. The process of their card placement appears, according to the order in which the students solve the small tasks. If one small task imposes high cognitive load on a student, the student takes either of a long time to solve it or a miss operation. Our method finds these tasks causing high cognitive load. These tasks tell what kinds of lacks in knowledges students have. Students can reduce the cognitive load when they store appropriate knowledges in their long term memory. On the other hand, if they do not have these knowledges, they have operate lots of information in their working memory. In this case, the cognitive load is high. Therefore, we can find the lack of knowledges.

Improvement of our method would enable teachers to know the lack in knowledge of each student. The method would make it easy for teachers to take care of students in the best way for each.

VII. CONCLUSION

We proposed a method which contributes to estimating students understanding level. In this method, we analyze behaviors of students in the English word sort assignment in a VR space. The analysis reveals students understanding level from their cognitive load. Students who have low confidence due to their lack of knowledge hesitate about answering the questions. Such students often stop their hands. They also show more gaze shift which come from comparison of their answers with hints. We had an experiment to confirm the relationship of the hands movements with the confidence. We examined the edit distances which represents how many correction is needed to finish English word sort assignment, We confirmed it influences the cognitive load. We also checked the relationship of the gaze shifts with the edit distances. According to the result of the experiment, we can estimate

the confidence of students from their hands movements of students except those who got poor scores. Moreover, the gaze shifts can be one explanatory variables to explain the edit distances. However, the gaze shifts cannot fully explain the edit distances. We need to find more explanatory variables.

If we can estimate the confidence of students by analyzing the observable value, we can know the cognitive load of students, while they are answering the tests. The cognitive load tells us where their weak points stay, which enables teachers to improve their teaching for each student. Consequently, this method promotes students understanding.

REFERENCES

[1] Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. Journal of Educational Psychology, 93, 579-588

[2] Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. Educational Psychology Review, 19(4), 469-508.

[3] Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251-296.

[4] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

[5] Peterson, L., & Peterson, M. (1959). Short-term retention of individual verbal items. Journal of Experimental Psychology, 58, 193-198.

[6] Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. Psychological Review, 102, 211-245

[7] Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (pp. 19-30). New York: Cambridge University Press

[8] Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. Cognition and Instruction, 12(3), 185-233.

[9] Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10(3), 251-296.

[10] Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. Journal of Educational Psychology, 88, 49-63.

[11] Fred Paas , Alexander Renkl & John Sweller (2003) Cognitive Load Theory and Instructional Design: Recent Developments, Educational Psychologist, 38:1, 1-4, DOI: 10.1207/S15326985EP3801_1

[12] Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24, 87-185

[13] S. Tachi, M. Sato, M. Hirose(2010),*"Science of Virtual Reality (バーチャルリアリティ学)"*, The virtual reality society of japan

[14] Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. "A systematic review on educational data mining." IEEE Access 5 (2017): 15991-16005.

[15] V. Ivancevic, M. Celikovic, I. Lukovic, "The individual stability of student spatial deployment and its implications", Int. Symp. Comput. Edu. (SIIE), pp. 1-4, Oct. 2012.

[16] K. Nakamura, K. Kakusho, M. Murakami, and M. Minoh(2010). "Estimating Learners' Subjective Impressions of the Difficulty of Course Materials by Observing Their Faces in e-Learning" The IEICE Transactions on Information and Systems(Japanese Edition) Vol.J93-D No.5 pp.568-578

[17] Richard A. Monty and John W. Senders. 1976. Eye Movements and Psychological Processes. Lawrence Erlbaum Associates, Hillsdale, NJ.

[18] FOVE. FOVE Eye Tracking Virtual Reality Headset. Retrieved September 19, 2017 from https://www.getfove.com/

[19] Mathieu Nancel, Olivier Chapuis, Emmanuel Pietriga, Xing-Dong Yang, Pourang P. Irani, and Michel Beaudouin-Lafon. 2013. High-precision pointing on large wall displays using small handheld devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'13, 831. https://doi.org/10.1145/2470654.2470773

[20] Marcos Serrano, Barrett Ens, Xing-Dong Yang, and Pourang Irani. 2015. Gluey: Developing a Head-Worn Display Interface to Unify the Interaction Experience in Distributed Display Environments. In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI'15, 161–171. https://doi.org/10.1145/2785830.2785838

[21] Kruger, Justin; Dunning, David (1999). "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments". Journal of Personality and Social Psychology 77 (6): 1121-34

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 47–56
DOI: 10.15439/2018F19
ISSN 2300-5963 ACSIS, Vol. 15

Efficient Discovery of Top-K Sequential Patterns in Event-Based Spatio-Temporal Data

Piotr S. Maciąg
Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19,
00-665, Warsaw, Poland,

***Abstract*—We consider the problem of discovering sequential patterns from event-based spatio-temporal data. The dataset is described by a set of event types and their instances. Based on the given dataset, the task is to discover all significant sequential patterns denoting the attraction relation between event types occurring in a pattern. Already proposed algorithms discover all significant sequential patterns based on the significance threshold, which minimal value is given by an expert. Due to the nature of described data and complexity of discovered patterns, it may be very difficult to provide reasonable value of significance threshold. We consider the problem of effective discovering K most important patterns in a given dataset (that is, discovering top-K patterns). We propose algorithms for unlimited memory environments. Developed algorithms have been verified using synthetic and real datasets.**

I. Introduction

DISCOVERING knowledge from spatio-temporal data is gaining attention of researchers nowadays. Based on literature, we can distinguish two basic types of spatio-temporal data: event-based and trajectory-based [1]. Event-based spatio-temporal data is described by a set of event types $F = \{f_1, f_2, \ldots, f_n\}$ and a set of instances D. Each instance $e \in D$ denotes an occurrence of a particular event type from F and is associated with instance identifier, location in spatial dimension and occurrence time. Fig. 1 provides possible sets $D = \{a1, a2, \ldots, d10\}$ and $F = \{A, B, C, D\}$. The same datasets are presented in Table I. Event-based spatio-temporal data and the problem of discovering frequent sequential patterns in this type of data have been introduced in [2].

The task of mining spatio-temporal sequential patterns in given datasets F and D may be defined as follows. We assume that the *following* relation (or attraction relation) $f_{i_1} \rightarrow f_{i_2}$ between any two event types $f_{i_1}, f_{i_2} \in F$ denotes the fact, that instances of event type f_{i_1} attract in their spatial and temporal neighborhoods occurrences of instances of event type f_{i_2}. The strength of the following relation $f_{i_1} \rightarrow f_{i_2}$ is investigated by dividing the density of instances of type f_{i_2} in spatio-temporal neighborhoods of instances of type f_{i_1} and density of instances of type f_{i_2} in the whole spatio-temporal embedding space V. If obtained ratio is greater than 1, then it is possible that $f_{i_1} \rightarrow f_{i_2}$ constitute a pattern. We provide the strict definition of density in Section III. The problem introduced in [2] is to discover all significant sequential patterns defined in the form $f_{i_1} \rightarrow f_{i_2} \rightarrow \cdots \rightarrow f_{i_m}$, where the significance threshold is given by an expert. In contrary to this approach, we consider the problem of discovering K most significant patterns in the given dataset. Providing significance threshold for discovering patterns may be difficult due to the complex nature of considered task.

TABLE I
An example of a spatio-temporal event-based dataset

Identifier	Event type	Spatial location	Occurrence time
a1	A	19	1
a2	A	83	1
⋮	⋮	⋮	⋮
b1	B	25	3
b2	B	1	3
⋮	⋮	⋮	⋮
c1	C	25	7
c2	C	15	7
⋮	⋮	⋮	⋮
d1	D	21	11
d2	D	13	12
⋮	⋮	⋮	⋮

The rest of the paper is organized as follows. Related work is described in Section II. In Section III, we provide elementary notions. Our algorithms and main results are presented in Section IV. In Section V, we provide experimental results for both real and synthetic data. In Section VI, we give conclusions and future problems. The main results of the paper are:

1) We introduce the notion of top-K patterns in event-based spatio-temporal data, namely we define the ranking of top-K sequential patterns with minimal length given by parameter *min_len* and point out the efficient pruning strategy for creating the top sequences set.
2) We formulate the algorithm discovering such top sequential patterns in event-based spatio-temporal data.
3) Proposed algorithm has been verified using both synthetic and real datasets. For experiments on synthetic data we used the same types of datasets as used in [2]. As a real datasets, we used the two types of datasets containing event instances related to air pollution data.

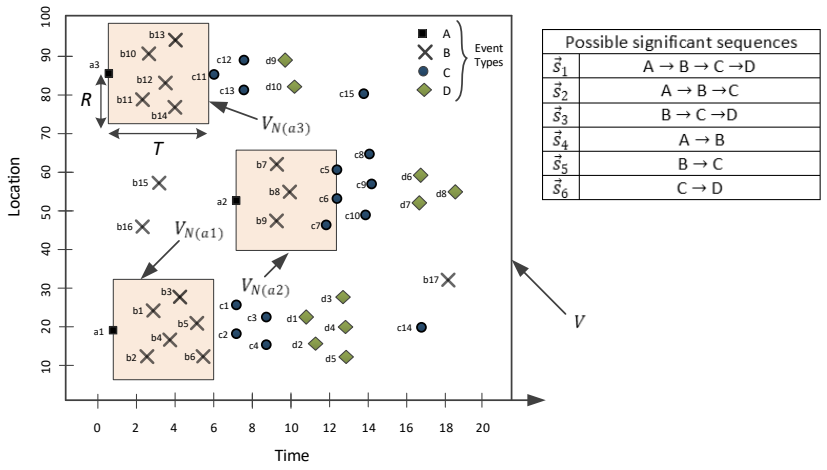


Fig. 1. Visualization the spatio-temporal event-based dataset from Table I and a set of possible significant sequences

II. Related Work

The problem of discovering top-K most important frequent patterns in various types of data has been well investigated in literature. TFP algorithm for discovering top-K closed frequent patterns for transaction databases has been given in [3]. In this approach, the user has a possibility to provide a parameter min_len specifying minimal length of discovered patterns (that is, minimal number of items occurring in a pattern). TFP discovers top-K closed frequent patterns by means of the FP-growth algorithm (proposed in [4]) for frequent patterns mining. The authors of [5] extends approach proposed in [3], by considering the problem of effective discovering top-K closed sequential patterns for transaction databases (the notion of closed sequential patterns has been introduced in [6]) and giving algorithm TSP for that purpose. In [7], the authors provide an algorithm discovering top-K jumping emerging patterns.

The problem of discovering sequential patterns in databases containing transactions records has been well investigated. The reader may refer to [8], [9], [10], [3] for the fundamental notions in this topic. More recently, surveys on methods for mining sequential patterns are given in [11], [12]. More recent papers in the area of mining top important frequent itemsets are [13], [14], [15].

Various types of methods have been developed for discovering patterns in event-based spatio-temporal data. The authors of [2] introduce the notion of sequential pattern for event-based spatio-temporal data and provide algorithms for both limited and unlimited memory environments. Obtained results show usefulness of proposed approach, however experiments (i.e. computation time) obtained for large datasets seem to be unsatisfactory. On the other hand, results presented in [2] are not well verified using real datasets. The additional drawback of algorithms proposed in [2] is large number of noise and redundant patterns obtained during mining process. The method of discovering top-K introduced in our article eliminate these deficiencies. A survey of methods for discovering patterns in spatio-temporal data is given in [16], [17]. The problem of discovering hierarchical spatio-temporal patterns has been considered in [18]. The problem of discovering spatio-temporal patterns from trajectory data and objects movements data has been considered in [19], [20], [1], [21], [22], [23].

III. Basic Notions

The dataset given in Fig. 1 is contained in the spatio-temporal space V, which temporal dimension is of size 20 and spatial location is provided by numbers between 0 and 100. For simplicity in Fig. 1 we denote spatial location in only one dimension. Usually, spatial location is defined by two dimensions (f.e. geographical coordinates). By $|V|$ we denote the volume of space V, calculated as the product of spatial area and size of time dimension. Spatial and temporal sizes of spatio-temporal space are usually given by an expert. For example, for Fig. 1 $|V| = 20 * 100 = 2000$. In the following definitions and notions we use terms sequential patterns and sequence interchangeably.

Definition 1. Neighborhood space. By $V_{N(e)}$ we denote the neighborhood space of instance e. For $V_{N(e)}$ having cylindrical shape, R denotes the spatial radius and T temporal interval of that space. The volume $|V_{N(e)}|$ of neighborhood space is equal to $\pi * R^2 * T$.

The shape of $V_{N(e)}$ is given by an expert and may be adjusted to particular dataset. Consider example given in Fig. 1 where we denote neighborhood spaces $V_{N(a1)}, V_{N(a2)}, V_{N(a3)}$. In Fig. 2, we provide an example of cylindrical neighborhood space $V_{N(a1)}$ with spatial location specified by two coordinates. The volume of that space is $|V_{N(a1)}| \approx 384.65$.

The reader may refer to [2] for other possible definitions of neighborhood spaces.

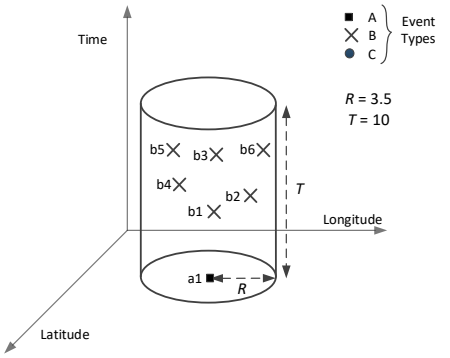


Fig. 2. Possible shape of neighborhood space $V_{N(a1)}$

Definition 2. Neighborhood [2]. For a given event type f and an occurrence of event instance e of that type, the neighborhood of e is defined as follows:

$$\begin{aligned} N(e) = &\{p | p \in D \\ &\wedge distance(p.location, e.location) \leq R \\ &\wedge (p.time - e.time) \in [0, T]\} \end{aligned} \quad (1)$$

where R denotes the spatial radius and T temporal interval of the neighborhood space $V_{N(e)}$.

As the neighborhood $N(e)$ of instance e, we denote the set of instances contained inside the neighborhood space $V_{N(e)}$.The neighborhood of instance $a1$ (shown in Fig. 2) with respect to event type B is $N(a1) = \{b1, b2, b3, b4, b5, b6\}$.

Definition 3. Density [2]. For a given spatiotemporal space V, event type f and its events instances in D, density is defined as follows:

$$Density(f, V) = \frac{|\{e | e.type = f \wedge \text{e is inside V}\}|}{|V|} \quad (2)$$

that is, density is the number of instances of type f occurring inside space V divided by the volume of that space.

Definition 4. Density ratio [2]. Density ratio for two event types f_{i_1}, f_{i_2} and their instances in D is defined as follows:

$$DR(f_{i_1} \rightarrow f_{i_2}) = \frac{avg_{e.type=f_{i_1}}(Density(f_{i_2}, V_{N(e)}))}{Density(f_{i_2}, V)} \quad (3)$$

where $\rightarrow$ denotes the *following* relation between event types f_{i_1}, f_{i_2}.
$avg_{e \in f_{i_1}}(Density(f_{i_2}, V_{N(e)}))$ specifies the average density of instances of type f_{i_2} occurring inside the neighborhood spaces $V_{N(e)}$ created for instances $e \in f_{i_1}$. V denotes the whole considered spatio-temporal space and $Density(f_{i_2}, V)$ specifies density of instances of type f_{i_2} inside that space.

If the value of density ratio for event types f_{i_1} and f_{i_2} is greater than one, then instances of type f_{i_1} attract in their spatio-temporal neighborhood spaces occurrences of instances of type f_{i_2}. If the value is below one, then they repel occurrences of instances of type f_{i_2}. If the value is equal to one, then there is no correlation between these two event types.

Definition 5. Sequence (sequential pattern) $\overrightarrow{s}$ and tailEventSet($\overrightarrow{s}$) [2]. $\overrightarrow{s}$ denotes a m-length sequence of event types: $s[1] \rightarrow s[2] \rightarrow \cdots \rightarrow s[m-1] \rightarrow s[m]$. tailEventSet($\overrightarrow{s}$) denotes the set of instances of type $\overrightarrow{s}[m]$ participating in the sequence $\overrightarrow{s}$.

Consider sequence $\overrightarrow{s_4} = A \rightarrow B$ given in Fig. 1. The length of the sequence is 2 and tailEventSet($\overrightarrow{s_4}$) = $\{b1, b2, \ldots, b14\}$ contains instances of event type B, which are in neighborhoods of instances of event type A.

Definition 6. Sequence index [2]. For a given m-length sequence $\overrightarrow{s}$, sequence index is defined as follows:

1) When $m = 2$ then:

$$SI(\overrightarrow{s}) = DR(\overrightarrow{s}[1] \rightarrow \overrightarrow{s}[2]) \quad (4)$$

2) When $m > 2$ then:

$$SI(\overrightarrow{s}) = \min \begin{cases} SI(\overrightarrow{s}[1:m-1]), \\ DR(\overrightarrow{s}[m-1] \rightarrow \overrightarrow{s}[m]) \end{cases} \quad (5)$$

where sequence $\overrightarrow{s}$ is constituted of event types $\overrightarrow{s}[1] \rightarrow \overrightarrow{s}[2] \rightarrow \cdots \rightarrow \overrightarrow{s}[m]$.

Example 1. Consider the dataset given in Fig .1. As an example let us consider the process of expanding sequence $\overrightarrow{s_1}$. One may notice that density of instances of type B is significant in the neighborhood spaces created for instances of type A. 1-length sequence $\overrightarrow{s_1} = A$ will be expanded to $\overrightarrow{s_1} = A \rightarrow B$ and as the tail event set of $\overrightarrow{s_1}$, the set of instances of type B contained in $N(a1)$ or $N(a2)$ or $N(a3)$ will be remembered (that is, tailEventSet($\overrightarrow{s_1}$) = $\{b1, b2, \ldots, b14\}$). The neighborhood spaces will be created for each instance contained in tailEventSet($\overrightarrow{s_1}$) and $\overrightarrow{s_1}$ will be expanded with event type C, to create $\overrightarrow{s_1} = A \rightarrow B \rightarrow C$. Actual will be tailEventSet($\overrightarrow{s_1}$) = $\{c1, c2, \ldots, c13\}$. In the same manner, the sequence will be expanded with event type D.

The sketch of the ST-Miner algorithm provided in [2] is as follows. First, for each event type in a dataset F, a 1-length sequence is created. Then, in a depth-first manner, each sequence is expanded with each event type in F, if the value of density ratio between the last event type in the sequence and event type considered to be appended is greater than the predefined threshold. The value of density ratio between these two event types is calculated by taking all instances from the tail event set of the sequence, creating their neighborhood spaces and verifying the ratio of the average density of instances of event type considered to be appended in these neighborhood spaces and the total density of instances of that type in the embedding space. If the value of density ratio is below given threshold, then the sequence is not expanded

any more. If the opposite is true, the sequence is expanded in the recursive way. The minimal value of density ratio between any two consecutive event types participating in the sequence is the sequence index (SI($\overrightarrow{s}$)).

IV. Discovering Top-K Patterns

In this section, we provide our algorithms discovering top-K sequential patterns.

Definition 7. For a sequence $\overrightarrow{s} \rightarrow f$ of length $m+1$, we say that f follows event type $\overrightarrow{s}[m]$. tailEventSet($\overrightarrow{s} \rightarrow f$) contains all instances of type f contained in the neighborhoods created for instances from tailEventSet($\overrightarrow{s}$).

Definition 8. Supersequence and subsequence. For two sequences $\overrightarrow{s_i} = \overrightarrow{s_i}[1] \rightarrow \overrightarrow{s_i}[2] \rightarrow \cdots \rightarrow \overrightarrow{s_i}[m_i]$ and $\overrightarrow{s_j} = \overrightarrow{s_j}[1] \rightarrow \overrightarrow{s_j}[2] \rightarrow \cdots \rightarrow \overrightarrow{s_j}[m_j]$, where $m_j > m_i$, $\overrightarrow{s_j}$ is supersequence of $\overrightarrow{s_i}$ ($\overrightarrow{s_i}$ is subsequence of $\overrightarrow{s_j}$) if only $\overrightarrow{s_i}[1] = \overrightarrow{s_j}[1] \wedge \overrightarrow{s_i}[2] = \overrightarrow{s_j}[2] \wedge \cdots \wedge \overrightarrow{s_i}[m_i] = \overrightarrow{s_j}[m_i]$.

In Fig. 1, $\overrightarrow{s_1}$ is supersequence of $\overrightarrow{s_2}$ ($\overrightarrow{s_2}$ is subsequence of $\overrightarrow{s_1}$). Please note however, that for example $\overrightarrow{s_1}$ is not supersequence of $\overrightarrow{s_3}$ (and $\overrightarrow{s_3}$ is not subsequence of $\overrightarrow{s_1}$).

Definition 9. Top-K sequence (sequential pattern). We say that sequence $\overrightarrow{s}$ of length *min_len* is the K-th top sequence (sequential pattern), if there exist K-1 sequences in the top sequences set with length *min_len* and the sequence index of each is equal or greater than $SI(\overrightarrow{s})$.

Definition 10. Pruning threshold θ. Actual pruning threshold θ for sequences considered to be in the top sequences set is equal to the sequence index of any K-th top already discovered sequence.

Lemma 1. For a given sequence $\overrightarrow{s}$ of a minimal length *min_len*, if the sequence index $SI(\overrightarrow{s})$ is below the actual pruning threshold θ, then $\overrightarrow{s}$ and any of its supersequences do not belong to top sequences and $\overrightarrow{s}$ should not be expanded with new event types any more.

Proof: If the sequence index of considered sequence $\overrightarrow{s}$ is below pruning threshold θ, then $\overrightarrow{s}$ does not belong to already discovered top sequences. By means of Definition 6 and Definition 8 any supersequence of $\overrightarrow{s}$ also does not belong to top-K sequences set, so $\overrightarrow{s}$ should not be expanded with new event types. ■

Informally the approach discovering top-K sequences is as follows: starting with 1-length sequences (that is, sequences containing singular event types) expand each sequence in a depth-first manner up to the moment when its length is at least *min_len*. We start discovering sequences with the basic value of pruning threshold θ equal to 1. At the same time we maintain the set of top-K already discovered patterns. By $D(f)$ we denote set of instances of type f in D.

In Algorithm 2, if the sequence index of considered sequence $\overrightarrow{s}$ is greater than pruning threshold θ then $\overrightarrow{s}$ will be expanded with new event types. Additionally, considering $\overrightarrow{s}$ to be inserted into top-K sequences ranking, three scenarios are possible:

Algorithm 1 Procedure for discovering top-K sequential patterns

Require: D - dataset containing event types and their instances, F - set of event types.
Ensure: A set of top-K sequential patterns.
1: **for** each event type $f \in F$ **do**
2: Create 1-length sequence $\overrightarrow{s}$ from f.
3: TailEventSet($\overrightarrow{s}$) := $D(f)$.
4: *ExpandSequence($\overrightarrow{s}$).*
5: **end for**

1) If the length of the sequence $\overrightarrow{s}$ is at least *min_len*, and if there are few than K - 1 patterns in the top-K set, then $\overrightarrow{s}$ is inserted into the set (case 1 in Fig. 3).
2) If the length of the sequence $\overrightarrow{s}$ is at least *min_len* and there are K - 1 patterns in the top-K set, then $\overrightarrow{s}$ is inserted into the set and pruning threshold θ is set to sequence index of already K-th sequence in the set (case 2 in Fig. 3).
3) If the length of the sequence $\overrightarrow{s}$ is at least *min_len* and there are K patterns in the top set, then if the sequence index of $\overrightarrow{s}$ is equal to threshold theta θ, then $\overrightarrow{s}$ is inserted into top set (cases 5, 6 in Fig. 3).
4) If the length of the sequence $\overrightarrow{s}$ is at least *min_len* and there are K patterns in the top set, then if the sequence index of $\overrightarrow{s}$ is less than threshold theta θ, then $\overrightarrow{s}$ is not inserted into the set (case 3).
5) If the length of the sequence $\overrightarrow{s}$ is at least *min_len* and there are K patterns in the top set, then if the sequence index of $\overrightarrow{s}$ is greater than threshold theta θ, then $\overrightarrow{s}$ is inserted into the set, θ is set to the value of any K-th sequences' sequence index and all the sequences with sequence indexes less than θ are deleted from the top set (case 4).

In Fig. 3, we show possible scenarios where $\overrightarrow{s}$ is considered to be inserted into top-K sequences set.

In Algorithm 2, Spatial Join procedure performed in step 2 calculates a join set between tail event set of $\overrightarrow{s}$ and set of instances $D(f)$ (that is, calculates tailEventSet($\overrightarrow{s} \rightarrow f$)). Spatial join may be performed using the *plane sweep* algorithm proposed in [24]. Algorithm 3 calculates actual sequence index of sequence $\overrightarrow{s} \rightarrow f$. DR($\overrightarrow{s}[m] \rightarrow f$) in step 1 of Algorithm 3 is calculated as follows. The nominator of ratio in Definition 4 is the average density of instances from tailEventSet($\overrightarrow{s} \rightarrow f$)) inside the neighborhood spaces created for instances from tailEventSet($\overrightarrow{s}$). The denominator is the density of instances of type f (that is, $D(f)$) inside embedding space V.

V. Experiments

We performed experiments on both generated (synthetic) and real datasets. Our experiments have been conducted using machine with Intel Core i7-6700HQ CPU, each 2.6GHz and 16GB of RAM.

Algorithm 2 ExpandSequence($\overrightarrow{s}$)

Require: $\overrightarrow{s}$ - sequence to be expanded, K - number of top sequences to discover, min_len - minimal length of discovered sequences, θ - pruning threshold for top sequences.

for each event type $f \in F$ **do**
  TailEventSet($\overrightarrow{s} \rightarrow f$) := SpatialJoin(TailEventSet($\overrightarrow{s}$), $D(f)$).
  Calculate SequenceIndex($\overrightarrow{s} \rightarrow f$).
  if SI($\overrightarrow{s} \rightarrow f$) $\geq \theta$ **then**
    if length($\overrightarrow{s} \rightarrow f$) $\geq min\_len$ **then**
      if Number of already discovered sequences $<$ K - 1 **then**
        Insert $\overrightarrow{s}$ into the top sequences set.
      else if Number of already discovered sequences $=$ K - 1 **then**
        Insert $\overrightarrow{s}$ into the top sequences set.
        θ := sequence index of the actual K-th sequence in the top-K set.
      else
        Insert $\overrightarrow{s}$ into the top sequences set.
        if $SI(\overrightarrow{s}) > \theta$ **then**
          θ := sequence index of the actual K-th sequence in the top-K set.
          Delete all sequences from the top sequences set with the sequence indexes less than θ.
        end if
      end if
    end if
    ExpandSequence($\overrightarrow{s} \rightarrow f$).
  end if
end for

Algorithm 3 Calculate SequenceIndex($\overrightarrow{s} \rightarrow f$)

Require: $\overrightarrow{s} \rightarrow f$ - a sequence of event types; $\overrightarrow{s}[m]$ - the last event type participating in $\overrightarrow{s}$.
Ensure: Actual sequence index $SI(\overrightarrow{s} \rightarrow f)$.
return min(SI($\overrightarrow{s}$), DR($\overrightarrow{s}[m] \rightarrow f$)).

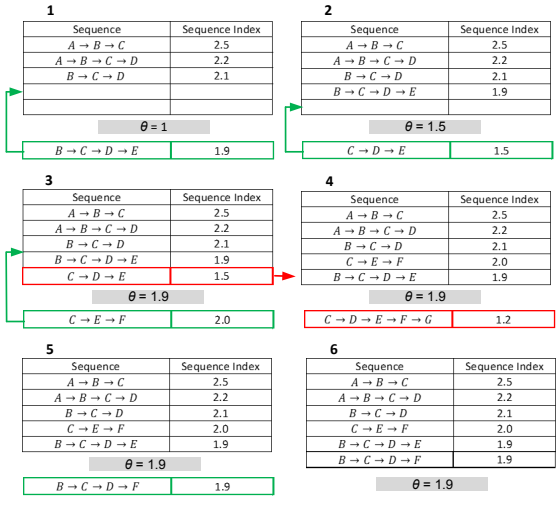


Fig. 3. Possibilities when $\overrightarrow{s}$ is considered to be inserted into top-K set with parameters min_len = 3 and K = 5

A. Experimental Results using Generated Data

We used the similar generator and notation of datasets names as proposed in [2]. In Table II, we recall parameters of data generator. In our experiments, we use cylindrical spatio-temporal neighborhood spaces $V_{N(e)}$ with parameters $R = 10$ (size of spatial dimension) and $T = 10$ (size of temporal window), similar to this one shown in Fig. 2. The whole spatio-temporal space V is given by parameters $DSize = 1000$ and $TSize = 1200$ (that is, both spatial dimensions are of size 1000 and temporal dimension is of size 1200). The total number of event instances in the dataset may be calculates as follows: $Pn * Ps * Ni * 2$, as in addition to patterns placed in a dataset we generate the same number of noise events.

We generated the same types of datasets as used in [2]. In Fig. 5, we show average computation times (we generated each dataset five times and averaged results) for three different types of datasets. In each case, computation time increases with increasing size of the dataset. We executed our algorithm for five values of K parameter (equal to 20, 40, 60, 80 and 100) and constant parameter min_len equal to 3. In Fig. 5, we are showing comparison of calculation time and the average number of discovered sequences for both STMiner proposed in [2] and our modification discovering top sequences set. The

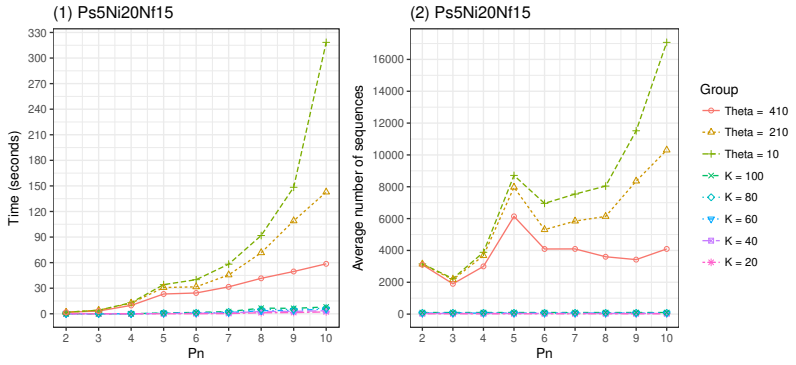


Fig. 4. The average computation times (plot 1) and average number of discovered sequences (plot 2) for both original STMiner algorithm proposed in [2] and our algorithm discovering top sequences set. The threshold $theta$ for STMiner has been set to three values: $10, 210, 410$

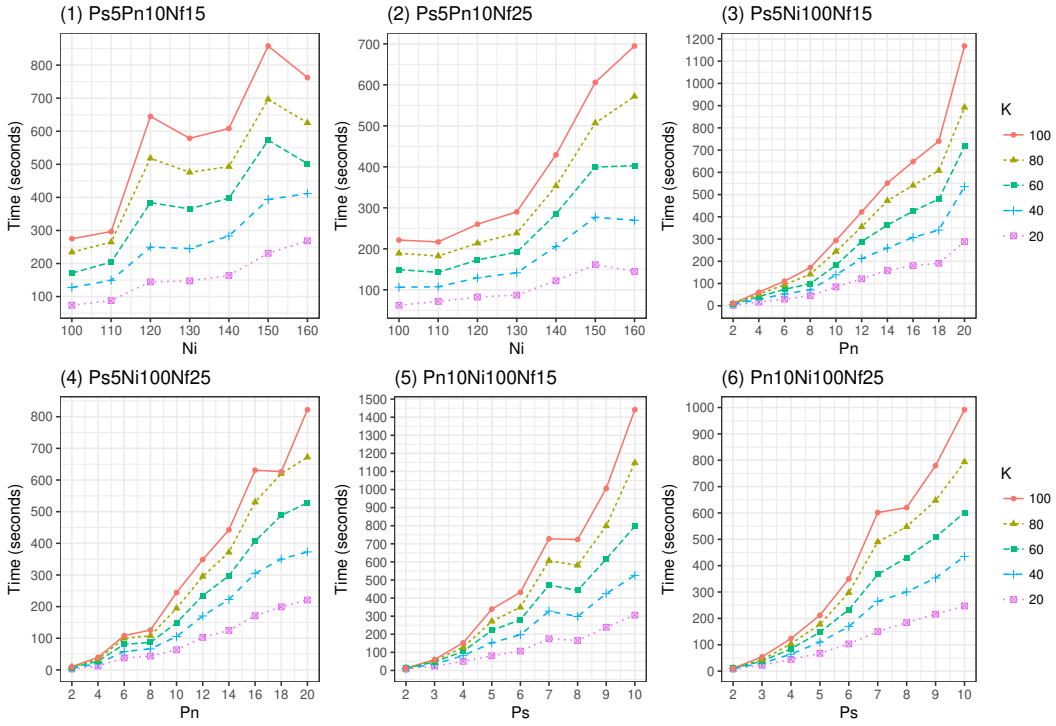


Fig. 5. Average computation times for randomly generated datasets with different number of instances per event type (diagrams (1), (2)), number of patterns (diagrams (3), (4)) and patterns lengths (diagrams (5), (6))

TABLE II
DESCRIPTION OF DATA GENERATOR PARAMETERS (ACCORDING TO [2])

Name	Description
Ps	Length (number of event types) of generated sequence
Pn	Number of sequences in generated data
$DSize$	Size of spatial dimensions of embedding space V
$TSize$	Size of temporal dimension of embedding space V
Nf	Total number of event types occurring in dataset
Ni	Number of instances per event type per sequence
R	Size of spatial dim. of neighborhood space $V_{N(e)}$
T	Size of temporal dim. of neighborhood space $V_{N(e)}$

size of the dataset for parameters $Pn = 10, Ps = 5, Ni = 20Nf = 15$ is 2000 event instances. As we may infer from Fig. 4, STMiner is impractical for even small datasets as it has a tendency to generate a huge number of redundant patterns. In Fig. 6, we show average calculation times for both STMiner and TopSTMiner when calculating exactly top 100 sequences set. To discover such sequences in STMiner algorithm we started with rather small θ threshold for sequence indexes and by its iterative increasing we obtained the set of 100 sequences.

B. Experimental Results using Real Data

For the first experiment on real data, we used the dataset of 14 types of pollutants available on the Internet repository

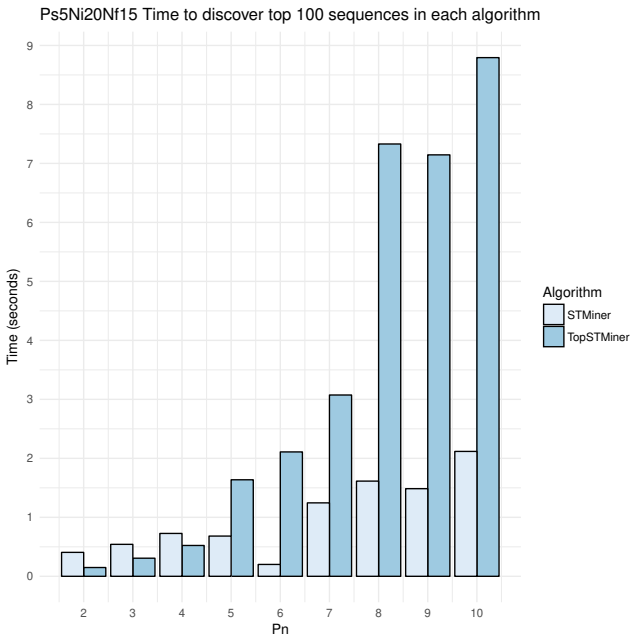


Fig. 6. The average computation times for both original STMiner algorithm proposed in [2] and our algorithm discovering top sequences set to calculate exactly top 100 sequences.

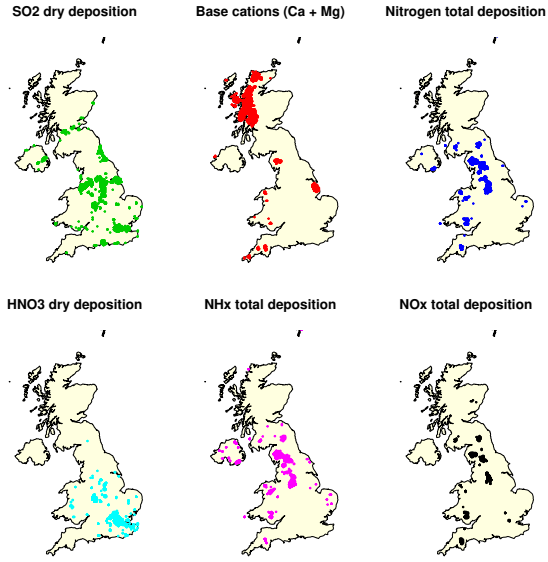


Fig. 7. Examples of extracted event types (SO2 dry deposition, base cations and total deposition of nitrogen, HNO3 dry deposition, NHx total deposition, NOx total deposition)

[25]. For each type of pollutant, the grids of resolution 5km^2 are available for years 2004-2014. Each grid contains numerical values of pollutant for United Kingdom region. For each grid and year, we calculated average value and standard deviation of pollutant. As abnormally high values of pollutants, we extracted events with values greater than three standard deviations from average. The task will be to investigate dependencies between these abnormal occurrences of pollutants. In Fig. 7, we show three types of events of pollutants extracted from the original dataset. We executed our algorithm with parameters $min\_len = 2$ and K $= 200$ and using cylindrical neighborhood space with parameters $R = 10$ km and $T = 1$ year. The types of pollutants available in the dataset and the number of abnormally high instances of each pollutant type in the final dataset are shown in Table III. In Table IV, we listed potentially interesting sequences form the top-100 set.

For the second experiment on real data we used the dataset of 6 pollutants obtained from 7 monitoring sites located in London Central: London Bloomsbury, London Eltham, London Haringey Priory Park South, London Harlington, London Hillingdon, London Marylebone, London Kensington. The name of pollutants and their numbers of instance in the extracted dataset are shown in Table V. The data have been obtained from the source [26]. Not each type of the pollutant is available for all of the stations. In Table VI, we show the name of each monitoring site, its location in the Northing, Easting system and available pollutants.

The original dataset contains hourly observations of pollutants shown in Table V for each day of 2015 for the stations mentioned above. For each type pollutant and for each station separately we extracted daily observations of such pollutant in the form of time series (that is, for each day we extracted 24 four observations respective to each hour). Then we clustered daily observations into four clusters to

TABLE III
TYPES OF POLLUTANTS USED IN THE FIRST EXPERIMENT (INST. - NUMBER OF INSTANCES)

Abbreviation	Pollutant type	Inst.
SOx-nss	Total deposition of oxidised sulphur	1989
SO4-nss	Wet deposition of sulphate	2256
SO2	Dry deposition of sulphur dioxide	1822
N	Total deposition of nitrogen	1339
NHx	Total deposition of reduced nitrogen	1252
NOx	Total deposition of oxidised nitrogen	579
NH3	Dry deposition of ammonia	1162
NH3-c	Concentration of ammonia	1292
NH4	Wet deposition of ammonium	2162
NO2	Dry deposition of nitrogen dioxide	698
HNO3	Dry deposition of nitric acid	1406
HNO3-c	Concentration of nitric acid	32
NO3	Wet deposition of nitrate	2021
Ca+Mg	Total deposition of base cations	2670
Ac	Total deposition of acidity	1406

TABLE IV
EXAMPLES OF PATTERNS DISCOVERED IN TOP-100 SET FOR REAL DATA FOR THE FIRST EXPERIMENT

Sequence	Sequence index
HNO3 $\rightarrow$ NO2	68.02
NO2 $\rightarrow$ HNO3	65.849
N $\rightarrow$ NOx $\rightarrow$ Ac $\rightarrow$ NHx $\rightarrow$ SOx	49.057
NOx $\rightarrow$ Ac $\rightarrow$ NHx $\rightarrow$ SO4	49.0339
NOx $\rightarrow$ Ac $\rightarrow$ NHx $\rightarrow$ NO3	47.8773
NOx $\rightarrow$ Ac $\rightarrow$ N $\rightarrow$ NH4	47.6831

TABLE V
TYPES OF POLLUTANTS USED IN THE SECOND EXPERIMENT

Pollutant type	Number of instances in dataset
Carbon Monoxide	52
Nitric Oxide	197
Nitrogen Dioxide	490
Ozone	534
PM10 particle deposition	161
PM2.5 particle deposition	73

obtain days with high concentration of the pollutant. For the clustering process we used R software, dtwclust package and distance time warping similarity between time series measure. The example of discovered clusters for Nitric Oxide pollutant for the London Eltham Station is shown in Fig. 8 and PM2.5 pollutant for the London Marylebone Station in Fig. 9. As the days with high concentration of pollutant we extracted these from cluster 2 for the former and cluster 4 for the latter. Each day with high pollutants' concentration has been marked as an event instance with event type corresponding to the pollutant type. The spatial location of the event instance is the location of respective monitoring station and the occurrence time is the corresponding day of occurrence.

We employed our algorithm to such dataset with parameters: $K = 100$, $min\_len = 2$, $R = 200$ meters and $T = 10$ days. The sizes of spatiotemoral space are as follows: $DSize1 = 37040$ meters, $DSize2 = 14262$ meters and $TSize$ = 364 days and are bounded by the locations of monitoring site and period of observation. The coordinates of stations are given in the Northing, Easting system. The parameter R specified as above means, that the algorithm will be looking for the interesting sequences considering events in each station separately. The set of top-15 sequences discovered from such dataset is shown in Table VII.

C. Results Discussion

For the experiments on synthetic data we show that even for small datasets our improvement discovering top sequences is more effective than the original algorithm STMiner proposed in [27]. As it has been explained, for many datasets and specific applications it may be difficult to provide a minimal sequence index threshold for discovered sequences. The algorithm proposed in the paper allows to eliminate this drawback by specifying the number of top sequences to discover. For the experimental results on real data we used two datasets, which have been preprocessed to obtain a set of event instances. For each of these datasets we obtain some potentially interesting sequences, however the additional usefulness of the proposed algorithm may be verified in the future experiments.

VI. CONCLUSIONS

In the paper, we consider the problem of effective discovering of top-K sequential patterns in event-based spatio-temporal data. In particular, we introduced the notion of top-K sequence (sequential pattern), we proposed the method creating set of top-K sequences and dynamically updating the set based on

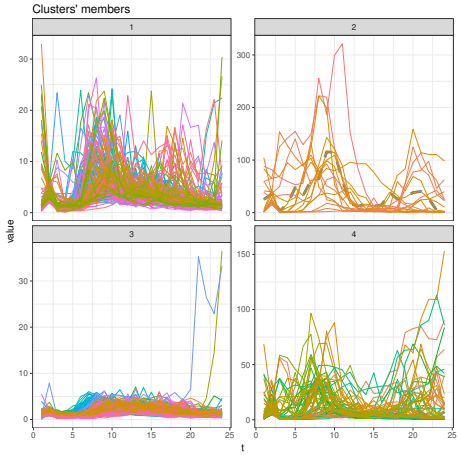


Fig. 8. Discovered clusters for Nitric Oxide pollutant for the London Eltham Station (cluster 2 contains days with high concentration of the pollutant)

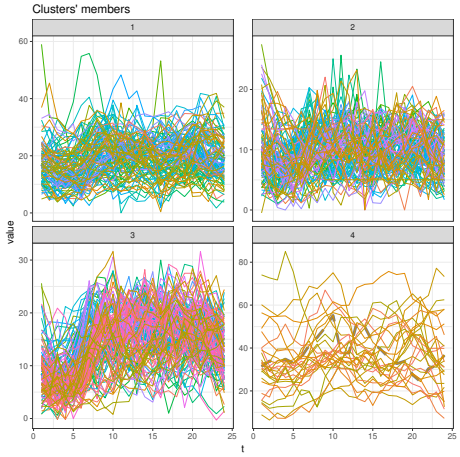


Fig. 9. Discovered clusters for PM2.5 pollutant for the London Marylebone Station (cluster 4 contains days with high concentration of the pollutant)

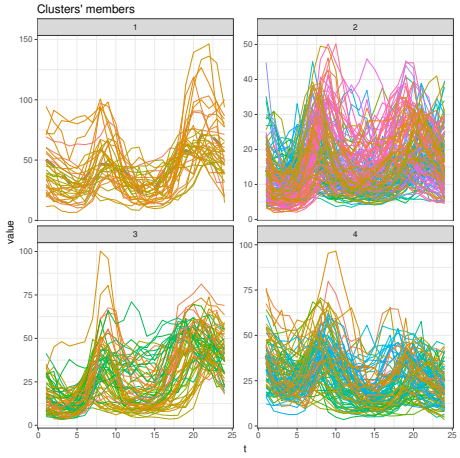


Fig. 10. Discovered clusters for Nitrogen Dioxide pollutant for the London Haringey Station (cluster 1 contains days with high concentration of the pollutant)

the rank of already expanded pattern. The approach allows to immediately prune patterns which for sure will not be

TABLE VI
MONITORING STATIONS, THEIR LOCATIONS (IN THE NORTHING, EASTING SYSTEM) AND AVAILABLE POLLUTANTS

Monitoring station	Location	Available pollutants
London Bloomsbury	530119, 182039	Nitric Oxide, Nitrogen Dioxide, Ozone, PM10, PM2.5
London Eltham	543981, 174655	Nitric Oxide, Nitrogen Dioxide, Ozone
London Haringey Priory Park South	529987, 188917	Nitric Oxide, Nitrogen Dioxide, Ozone, PM10, PM2.5
London Harlington	508295, 177800	Nitric Oxide, Nitrogen Dioxide, Ozone, PM10, PM2.5
London Hillingdon	506941, 178610	Nitric Oxide, Nitrogen Dioxide, Ozone
London Marylebone	528126, 182015	Carbon Monoxide, Nitric Oxide, Nitrogen Dioxide, Ozone, PM10, PM2.5
London Kensington	524045, 181749	Carbon Monoxide, Nitric Oxide, Nitrogen Dioxide, Ozone PM10, PM2.5

TABLE VII
EXAMPLES OF PATTERNS DISCOVERED IN TOP-100 SET FOR REAL DATA FOR THE SECOND EXPERIMENT

Sequence	Sequence index
PM10 → PM25	1000
PM25→ CarbonMonoxide	1000
PM25→ PM10	1000
PM25→ CarbonMonoxide→ NitrogenDioxide	974.079
CarbonMonoxide→ PM25	927.609
CarbonMonoxide→ PM25→ NitrogenDioxide	865.219
NitricOxide→ PM25	841.01
NitricOxide→ PM25→ PM10	841.01
CarbonMonoxide→ PM10	804.612
CarbonMonoxide → PM10 →PM25	804.612
CarbonMonoxide → PM10 → PM25 → Ni.Di.	804.612
NitrogenDioxide→ PM25	800.361
NitrogenDioxide→ PM25→ CarbonMonoxide	800.361
NitrogenDioxide→ PM25→ PM10	800.361
NitricOxide → PM25 → CarbonMonoxide	785.106

among the top-K sequences with length defined by *min_len* parameter. In the experiments, we show the efficiency of proposed approach. We also presented experimental results for real datasets. Obtained results are encouraging to investigate the topic in future research.

ACKNOWLEDGMENT

We acknowledge use of the dataset of UK Pollutants [25] available on the webpages http://www.pollutantdeposition.ceh.ac.uk/data and https://uk-air.defra.gov.uk/data/

REFERENCES

[1] Z. Li, *Spatiotemporal Pattern Mining: Algorithms and Applications*. Cham: Springer International Publishing, 2014, pp. 283–306.

[2] Y. Huang, L. Zhang, and P. Zhang, "A framework for mining sequential patterns from spatio-temporal event data sets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 433–448, April 2008.

[3] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequent closed patterns without minimum support," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 2002, pp. 211–218.

[4] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, Jan 2004. [Online]. Available: http://dx.doi.org/10.1023/B:DAMI.0000005258.31418.83

[5] P. Tzvetkov, X. Yan, and J. Han, "Tsp: mining top-k closed sequential patterns," in *Third IEEE International Conference on Data Mining*, Nov 2003, pp. 347–354.

[6] X. Yan, J. Han, and R. Afshar, *CloSpan: Mining: Closed Sequential Patterns in Large Datasets*, 2003, pp. 166–177. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9781611972733.15

[7] P. Terlecki and K. Walczak, "Efficient discovery of top-k minimal jumping emerging patterns," in *Rough Sets and Current Trends in Computing: 6th International Conference, RSCTC 2008 Akron, OH, USA, October 23-25, Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 438–447.

[8] R. Srikant and R. Agrawal, *Mining sequential patterns: Generalizations and performance improvements*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 1–17. [Online]. Available: http://dx.doi.org/10.1007/BFb0014140

[9] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings of the Eleventh International Conference on Data Engineering*, Mar 1995, pp. 3–14.

[10] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine Learning*, vol. 42, no. 1, pp. 31–60, Jan 2001. [Online]. Available: https://doi.org/10.1023/A:1007652502315

[11] P. Fournier-Viger, J. C.-W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A survey of sequential pattern mining," *Data Science and Pattern Recognition*, vol. 1, no. 1, pp. 54–77, 2017.

[12] C. H. Mooney and J. F. Roddick, "Sequential pattern mining – approaches and algorithms," *ACM Comput. Surv.*, vol. 45, no. 2, pp. 19:1–19:39, Mar. 2013. [Online]. Available: http://doi.acm.org/10.1145/2431211.2431218

[13] C. W. Wu, B.-E. Shie, V. S. Tseng, and P. S. Yu, "Mining top-k high utility itemsets," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 78–86. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339546

[14] J. Yin, Z. Zheng, L. Cao, Y. Song, and W. Wei, "Efficiently mining top-k high utility sequential patterns," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 1259–1264.

[15] F. Petitjean, T. Li, N. Tatti, and G. I. Webb, "Skopus: Mining top-k sequential patterns under leverage," *Data Min. Knowl. Discov.*, vol. 30, no. 5, pp. 1086–1111, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1007/s10618-016-0467-9

[16] S. Shekhar, M. R. Evans, J. M. Kang, and P. Mohan, "Identifying patterns in spatial information: A survey of methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 193–214, 6 2011.

[17] P. Mohan, S. Shekhar, J. A. Shine, and J. P. Rogers, "Cascading spatio-temporal pattern discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 1977–1992, Nov 2012.

[18] C. H. Yu, W. Ding, M. Morabito, and P. Chen, "Hierarchical spatio-temporal pattern discovery and predictive modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 979–993, April 2016.

[19] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung, "Mining, indexing, and querying historical spatiotemporal data," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 236–245. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014080

[20] Z. Li, B. Ding, J. Han, and R. Kays, "Swarm: Mining relaxed temporal moving object clusters," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 723–734, Sep. 2010. [Online]. Available: http://dx.doi.org/10.14778/1920841.1920934

[21] Z. Li, J. Wang, and J. Han, "eperiodicity: Mining event periodicity from incomplete observations," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1219–1232, May 2015.

[22] P. Yin, M. Ye, W.-C. Lee, and Z. Li, "Mining gps data for trajectory

recommendation," in *Advances in Knowledge Discovery and Data Mining*, V. S. Tseng, T. B. Ho, Z.-H. Zhou, A. L. P. Chen, and H.-Y. Kao, Eds. Cham: Springer International Publishing, 2014, pp. 50–61.

[23] Y. Li, J. Bailey, L. Kulik, and J. Pei, "Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases," in *2013 IEEE 13th International Conference on Data Mining*, Dec 2013, pp. 448–457.

[24] L. Arge, O. Procopiuc, S. Ramaswamy, T. Suel, and J. S. Vitter, "Scalable sweeping-based spatial join," in *Proceedings of the 24rd International Conference on Very Large Data Bases*, ser. VLDB '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 570–581. [Online]. Available: http://dl.acm.org/citation.cfm?id=645924.671340

[25] Uk pollutant deposition data. [Online]. Available: http://www.pollutantdeposition.ceh.ac.uk/data

[26] Department for environment food and rural affairs archive data. [Online]. Available: https://uk-air.defra.gov.uk/data/

[27] M. R. Haylock, N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, "A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006," *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D20, pp. n/a–n/a, 2008, d20119. [Online]. Available: http://dx.doi.org/10.1029/2008JD010201

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 57–65
DOI: 10.15439/2018F331
ISSN 2300-5963 ACSIS, Vol. 15

The Practical Use of Problem Encoding Allowing Cheap Fitness Computation of Mutated Individuals

Michal Przewozniczek
Department of Computational Intelligence
Faculty of Computer Science and Management
Wroclaw University of Science and Technology
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
Email: michal.przewozniczek@pwr.edu.pl

Marcin Komarnicki
Department of Computational Intelligence
Faculty of Computer Science and Management
Wroclaw University of Science and Technology
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
Email: marcin.komarcniki@pwr.edu.pl

***Abstract*—The usual assumption in the Evolutionary Computation field is that a cost of computing single fitness function evaluation is at last similar for all cases. Such assumption does not have to be true. In this paper we consider the recently proposed Problem Encoding Allowing Cheap Fitness Computation of Mutated Individuals (PEACh) effect that allows to significantly reduce the computation load of some of the fitness computations that occur during the evolutionary method run. To the best of our knowledge, it is the first experimental analysis that investigates the results of PEACh application to methods solving NP-hard practical problems.**

I. Introduction

THE OPTIMIZATION of computation load consumption by Evolutionary Algorithms (EAs) is a valid and important topic since these are the tools applied to solve hard computational problems. Many techniques were proposed to minimize the expenses for computing the fitness [7], [12], [6]. In this paper, we consider the Problem Encoding Allowing Cheap Fitness Computation of Mutated Individuals (PEACh) effect [17]. PEACh is a recently proposed technique that allows to significantly reduce the computation load spent on a single fitness computation without losing the precision of a result. To the best of our knowledge, except some theoretical experiments proposed in [17], there are no studies that would show the PEACh benefits obtained by applying it to Genetic Algorithms (GAs) solving a hard, practical problem. Therefore, in this paper, we apply PEACh to evolutionary methods applied to solve NP-hard flow optimization problem. The main objective of this paper is to check how significant (if any) is the PEACh influence on results quality and the method efficiency.

The other objective of this paper is to investigate which methods are more suitable to use PEACh benefits. In [17] the standard GA was pointed as a method that is capable of using PEACh optimization only for mutation operator. On the other hand, the multi-population methods employing so-called messy-coding [2], [8], [16], [15] were pointed out as those that should improve their speed more significantly.

This work was supported by the Polish National Science Centre (NCN) under Grant 2015/19/D/ST6/03115

Another important issue that was raised in [17] is the fairness of computation load measurement by using the Fitness Function Evaluation number (FFE). The application of PEACh does not change the method run, except some of the fitness value computations are performed significantly faster. Therefore, in such situations, the use of FFE as a fair computation load measure may be questioned. In this paper, by showing the speed-up, scale we verify if FFE is truly unfair computation load measure for methods using PEACh.

The rest of this paper is organized as follows. In the second section we present the related work, Section 3 contains a definition of the considered practical problem. The fourth section defines PEACh and analyses its possible applications to the considered practical problem. The research results and analysis is presented in Section 5. Finally, the last section summarizes this work and points on most promising future work directions.

II. Related work

In this section, we will present the different propositions of computation load optimization techniques employed in the Evolutionary Computation field. We will also discuss some related issues, eg. the fairness of the computation load measurement with the use of Fitness Function Evaluation number (FFE). Finally, we will briefly present the benefits of using multi-population approaches with a dynamically changed number of subpopulations since one of the methods considered in this paper is employing such techniques.

In some of the papers concerning Evolutionary Algorithms (EAs) applied to solve practical problems it is pointed that the computation load necessary to compute particular individual's fitness may be significantly decreased if the fitness value of similar (but not necessarily the same) individual is known. In [9], [10] different Resource-Constrained Scheduling Problem (RCSP) version are considered. To compute fitness of any individual, the problem solution that is represented by this individual must be constructed first. Then, the constructed solution is rated, and the fitness is computed. Some of the presented methods introduce small changes to already known and already rated individuals. Thus, the fitness of new individual that is a result of small genotype modification may be

computed in two different ways. It may be computed in a usual way by constructing and rating the solution. However, it is also possible to copy the solution of an old, already rated individual, modify it and rate it. In RCSP problem the computation load necessary for solution construction is significantly higher than the computation load required for rating the constructed solution. Thus, in NEH2 heuristic (Nawaz, Enscore, Ham) proposed in [9] the fitness of new solutions is computed by modifying the already known solutions rather than by building new solutions from scratch. In result, NEH2 can use a higher FFE number than its predecessor – NEH heuristic.

Surrogate model [6] is an interesting technique of computation load usage optimization that is useful when single fitness evaluation takes significantly long time (e.g. minutes, hours, or even days). In such situation, the use of evolutionary methods may be limited. The idea is to propose a new problem model that would mimic the real model but would be significantly cheaper to evaluate. Therefore, the use of surrogate model enables the use of EAs for problems to which they would be otherwise inapplicable due to practical reasons. The surrogate model is similar to PEACh effect considered in this paper because it optimizes the cost of fitness value computation. However, the difference is that PEACh leads to an exact fitness value, while surrogate model offers only the approximated fitness value.

Another technique that allows decreasing the computation costs is fitness caching [7], [12]. Its idea is based on storing the information about fitness values computed for the particular genotypes. When such knowledge is available instead of computing fitness the method checks if the fitness for the particular genotype was not already computed. If so, then instead of computing fitness the stored fitness value is returned. Such technique may significantly decrease FFE. However, the drawback is that checking if the fitness for the particular genotype was not computed before becomes more and more expensive in time as the list gets longer. At some point, the benefits brought by omitting the fitness computation may be exceeded by costs generated by the list search. Another issue is that the list of already rated genotypes may consume high amounts of memory. Therefore in [7] two different fitness caching techniques are described - *brutal fitness caching* (that is the technique described above) and *population fitness caching*. The population fitness caching works in the same way as its brutal version but instead using the genotypes list, the optimization is limited to the search through the current population. The research presented in [7] proposes an analysis of benefits brought by both fitness caching techniques. The research is based on modern evolutionary methods: Linkage Learning Genetic Algorithm (LTGA) [18], Dependency Structure Matrix Genetic Algorithm II (DSMGA-II) [5] and Parameter-less Population Pyramid (P3) [3]. Another important issue is shown in [7] is that when any fitness caching is used the FFE is not a reliable computation load measure when a method gets stuck. The reason for this situation is that once a method gets stuck it simply loses the capability of proposing new solutions, which have not been investigated yet. If so, then at some point all, or almost all fitness computation requests are cached. In the research presented in [7] FFE per iteration may drop to zero for the whole remaining method run. Thus, a different computation load measure than FFE number is necessary because in the described situation FFE only shows that method is not consuming any computation load at all which is not true.

The issue of fair computation load measurement is also addressed in [14]. One of the assumptions of using FFE as a fair computation load measure is that the computation load required to compute a single fitness value is the same or, at least, similar. The research presented in [14] show that this assumption is not always true. For instance, the computation load may be dependent on the genotype. Another requirement to use FFE as a fair computation load measure is that the dependency between the overall computation load used by a method and FFE should be close to linear. In other words, the computation load necessary to compute the fitness value is significantly higher than the computation load used for all other method activities. If this condition is not true, then the use of FFE as computation load measure may not be reliable. This issue is discussed in details in [8] on the base of Bayesian Optimization Algorithm (BOA). During its run, at each iteration BOA constructs the model of gene dependencies. When the genotype is long the computation load necessary for model construction significantly exceeds the computation load spent on all other method activities making it ineffective.

III. Flow assignment in computer networks

The problem of flow assignment in computer networks is one of the main problems in the field of network design [11]. The other are the capacity assignment, flow and capacity assignment, topology, flow, and capacity assignment. In the flow assignment problem, the solution shall satisfy the set of demands. Each demand defines an amount of information that is to be sent between a particular pair of network nodes. The list of demands as well as the network topology are given and cannot be modified. For each demand, a route in the network must be set to satisfy it. Thus, the solution to the problem is a list of routes (one route for one demand) that satisfy all demands and do not break the network links capacity constraint. The solution quality may be measured in many different ways. Here, we employ the Lost Flow in Link (LFL) function. LFL describes how well the network topology is prepared for the link breakdown scenario. The optimization of LFL value increases the network survivability and the quality of service. The problem denominated as WP_LFL [15] and is NP-complete [11].

The notation used to represent the WP_LFL problem is presented below.

Sets

V - set of n vertices representing the network nodes
A - set of m arcs representing network directed links
P - set of q connections in the network

$\prod_p$ - the index set of candidate working paths (routes) for connection p
X_r - set (selection) of variables x_k^p, which are equal to one. X_r determines the unique set of currently selected working paths
Indices
p - connections (demands) in the network, used as subscript
k - candidate routes, used as superscript
a - arcs (directed links), used as subscript
r - selections, used as subscript
Other
$o(a)$ - the start node of arc a
$d(a)$ - the end node of arc a
Constants
δ_{pa}^k - equal to1, if arc a belongs to path k realizing connection p; 0 otherwise
Q_p - volume (estimated bandwidth requirement) of connection p
c_a - capacity of arc a
Variables
x_p^k - decision variable, which is 1 if working route $k \in \prod_p$is selected for connection p and 0 otherwise
f_a - flow of arc a
$g_v^{in} = \sum_{i:d(i)=v} f_i$ - aggregate flow of incoming arcs of v;
$e_v^{in} = \sum_{i:d(i)=v} c_i$ - aggregate capacity of incoming arcs of v;
$g_v^{out} = \sum_{i:o(i)=v} f_i$ - aggregate flow of outgoing arcs of v;
$e_v^{out} = \sum_{i:o(i)=v} c_i$ - aggregate capacity of outgoing arcs of v;

The $o : A \to V$ and $d : A \to V$ functions denote the origin and destination node of each arc. For each $a \in A$ the set of incoming arcs of $d(a)$ except a $in(a) = \{k \in A | d(k) = d(a), k \neq a\}$, and the set of outgoing arcs of $o(a)$ except a $out(a) = \{k \in A | o(k) = o(a), k \neq a\}$ are defined.

Definition 1. *The global non-bifurcated m.c. flow denoted by $\underline{f} = [f_1, f_2, ..., f_m]$ is defined as a vector of flows in all arcs. The flow $\underline{f}$ is feasible if for every arc $a \in A$ the following inequality holds*

$$\forall a \in A : f_a \leq c_a \tag{1}$$

Inequality 1 ensures that in every arc, flow is not greater than capacity. This inequality is called the capacity constraint.

For the sake of simplicity, the following function is introduced

$$\epsilon(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > k \end{cases} \tag{2}$$

The analysis of local repair properties is based on a scenario where the failure of arc $k \in A$ is considered. If local repair is used then flow on the arc k must be rerouted by the origin node of k. Therefore, residual capacity of arcs outgoing from node $o(k)$ except arc k is a potential bottleneck of the restoration process. Since

$$f_k \leq \sum_{i \in out(k)} (c_i - f_i) \tag{3}$$

then the flow of the failed k can be restored using the residual capacity of other links leaving the origin node of k. Otherwise, if

$$f_k > \sum_{i \in out(k)} (c_i - f_i) \tag{4}$$

then some flow of the failed link k cannot be restored because the residual capacity of other arcs leaving the origin node of k is too small. As a consequence, the 100% restoration is not possible and some flow of k is lost. Applying formulas 3 and 4, and the $g_{o(k)}^{out}$, $e_{o(k)}^{out}$ definitions, the LA^{out} function is defined as follows.

$$LA_k^{out}(\underline{f}) = \epsilon(g_{o(k)}^{out} - (e_{o(k)}^{out} - c_k)) \tag{5}$$

Formula 5 defines the flow lost for arc k as dependent on the whole flow leaving the origin node of k. Therefore, the function of flow lost for all arcs leaving node v is defined as follows.

$$LN_v^{out}(\underline{f}) = \sum_{a:o(a)=v} \epsilon(g_v^{out} - (e_v^{out} - c_a)) = \sum_{a:o(a)=v} LA_v^{out}(\underline{f}) \tag{6}$$

Function $LN_v^{in}(\underline{f})$ is analogous to $LN_v^{out}(\underline{f})$ and defines how much flow is lost in the arcs incoming to node v. The goal of defining a function that measures the network preparation to link breakdown is realized in 7.

$$LFL(\underline{f}) = \sum_{v \in V} (LN_v^{in}(\underline{f}) + LN_v^{out}(\underline{f}))/2 \tag{7}$$

More details about the LFL may be found in [13]. The optimization problem, WP_LFL [13], [15] is defined as follows.

$$\min_{\underline{f}} \quad LFL(\underline{f}) \tag{8}$$

subject to

$$\sum_{k \in \prod_p} x_k^p = 1 \quad \forall p \in P \tag{9}$$

$$x_k^p \in \{0, 1\} \quad \forall p \in P, \forall k \in \prod_p \tag{10}$$

$$f_a = \sum_{p \in P} \sum_{k \in \prod_p} \delta_{pa}^k x_p^k Q_p \quad \forall a \in A \tag{11}$$

$$f_a \leq c_a \quad \forall a \in A \tag{12}$$

Condition (9) guarantees that the each connection can use only one working route. Constraint (10) ensures that decision variables are binary ones. Formula (11) defines a link flow. Finally, (12) denotes the link capacity constraint. The

WP_LFL problem given by (8)-(12) is a 0/1 NP problem with linear constraints. For this problem, the solution space that includes all possible paths for each connection is large even for relatively small networks. Therefore, the problem is considered hard. Let us consider 2500 demands and ten routes available for each demand. The number of solutions (not all may be feasible) 10^{2500}.

IV. PEACh effect in considered practical problem

In this section, we present PEACh idea proposed in [17]. Since PEACh benefits are strictly dependent on the problem encoding, in the second subsection we give a detailed description of solution encoding for the considered WP_LFL problem. Finally, in the last subsection, we present the two considered competing methods and discuss whether they are suitable to use PEACh benefits.

A. General PEACh description

Let us consider a situation in which fitness computation process is built from two stages - solution creation and solution rating. For some problems, the main computational cost is paid on the solution creation stage, while the process of rating the created solution is relatively cheap [17]. In such situation, if we wish to rate an individual but we know a similar solution that was already rated it is reasonable to copy the rated solution, modify it and rate the resulting solution.

Let us consider the Traveling Salesman Problem (TSP). TSP can be represented as a graph $G = (V, E)$, where V is a set of n vertices (cities) and $E = \{ei, j\}_{n\times n}$ is a set of edges that represent connections between cities. The distance function $d : E \to \Re^+$ is used to assign each edge e a distance value. The goal is to find a Hamiltonian cycle of minimal distance that visits each vertex only once. Let us assume that in the considered TSP instance we need to visit 1000 cities, the genotype encodes the solution by storing a list of genes that are city identifiers. Some individual that was already rated is being mutated by changing two genes (cities order) in its genotype. To compute fitness for the mutated version of the individual, we may compute the cost generated by using 1000 routes that are encoded by the genotype of the mutated individual. Another option is to copy the fitness of individual before mutation, subtract from it the distance of two routes that were removed and add the distance yield by two new routes. Note, that the second option requires significantly lower computation load.

We may state that PEACh benefits are used if the following condition holds

$$cost(X) \gg cost(X_{mut}, fitInfo(X)) \tag{13}$$

subject to

$$diffr(X, X_{mut}) \ll size(X) \tag{14}$$

where
X, X_{mut} - individuals genotypes
$cost(X)$ - the computation load that must be paid to compute the fitness of X without any prior knowledge
$fitInfo(X)$ - the additional information about data produced during X fitness value calculation process (including X fitness value if necessary)
$cost(X, addInfo)$ - the computation load necessary to compute the fitness of X with additional information taken from the other fitness value calculation operation
$diffr(X, Y)$ - the number of genotype positions for which genotype X is different than the genotype Y
$size(X)$ - the number of genotype positions in genotype X

If an additional information (from another fitness value computation operation) is available, then the computation load necessary for the fitness value computation will be significantly lower than if the same computation was done without the additional knowledge (inequality (13) is true). Condition (14) guarantees that the additional information is supported by the fitness calculation process for similar genotype.

The above definition proposed in [17] does not define a unit of computation load amount (returned by $cost(X, addInfo)$ and $cost(X)$). In this paper, similar as in [17], we use the computation time. Another available choice is the number of processor instructions. Note, that FFE is not an allowed choice, since the amount of computation load necessary to compute a single fitness evaluation may significantly differ.

B. PEACh in WP_LFL

In WP_LFL the network topology is given. The network links connecting the nodes are directed, so the particular link transfers data only in one direction. Therefore, the network may be represented as a directed graph. The list of demands defining the nodes pairs between the communication channels must be established is also entry information. Each demand except the start and destination node defines the volume of the demanded communication channel. Each connection is using only one route. Thus, the solution to the problem is a set of routes (each route proposed for one demand). Many communication channels may go through a single network link. The summarized volume going through a single network link may not exceed the link capacity. The example of the 4-demand problem solution encoded in the GA-like manner may be as follows: [(4) (6) (11) (13)]. In this solution the first demand will use route number 1, second demand will use the sixth route, etc. Note, that the solution does not have to be feasible, i.e. the link capacity constraint may be broken.

In Fig. 1 we present the example of the network state in the form of the two-dimensional matrix. The numbers in the matrix represent the available capacity of network links. For instance, the link from node A to node B has 24 capacity units left. For the network state presented in the upper part of the figure, we wish to set up the connection channel using the route from node A through nodes B and C to node D. The demand size (volume) of the connection channel is 4 capacity units. In the result of this operation, the available capacity in links A to B, B to C and C to D is decreased by 4. In the WP_LFL instances consider in the research present in this

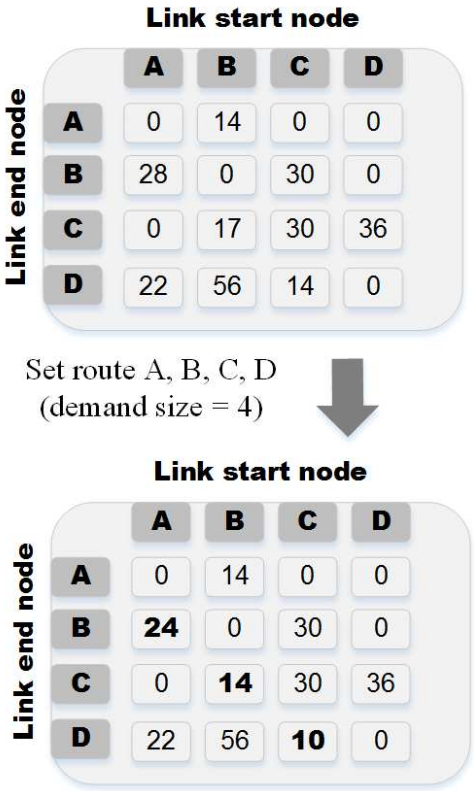


Fig. 1. The example of route establish

paper we use 1260 up to 2500 demands. To compute fitness of any individual we have to create a solution (by setting all connection channels for all demands) and then rate it (compute LFL function value for all links and summarize it). For the considered test cases the time necessary to construct the problem solution (T_{model}) is significantly larger than the time necessary to rate the constructed solution ($T_{quality}$). Therefore, the decreasing T_{model} time, shall significantly decrease the overall time necessary for fitness computation. Let us consider the following situation. The individual X encoding a solution to 2500-demands WP_LFL problem instance was rated and cost of this operation was $cost(X) = T_{model} + T_{quality}$. Then, individual X was mutated (one of its genes was modified) creating individual X_{mut}. We can compute fitness of individual X_{mut} without using any additional knowledge. If so, then $cost(X_{mut}) \approx costX$. However, we can also compute the fitness of X_{mut} differently and use the model created for individual X. It is enough to unset one connection channel that was encoded by individual X (set this channel with negative volume) and set the new channel that was defined in individual X_{mut}. In such situation we need to perform two channel setting operations instead of 2500 route setting operations $cost(X_{mut}, fitInfo(X)) = 2/2500 \cdot T_{model} + T_{quality}$. Since for the considered test cases $T_{model} \gg T_{quality}$, it is true that $cost(X) \gg cost(X_{mut}, fitInfo(X))$. Thus, we may expect that using the PEACh effect may significantly decrease the computation load consumed by a method during its run.

C. Competing methods and PEACh suitability

In this section, we present the two methods chosen for tests. The first is Hierarchical Evolutionary Algorithm for Flow Assignment in Non-bifurcated Commodity Flow (HEFAN 2.2) [13]. HEFAN 2.2 is a standard GA that is dedicated to solving flow optimization problem. The individuals are encoded in a standard GA-manner presented in the previous section. Since considering all available routes between every nodes pair would lead to the combinatorial explosion and would make the problem intractable, at the beginning of the run HEFAN 2.2 uses four shortest routes between each node pair and all routes that are not longer than 1-, 2-, 3-, and 4-hops. Such route set may be not enough to construct some high-quality solutions. Therefore, HEFAN 2.2 uses two problem-dedicated operators that were designed to propose new routes, namely the low-level crossover and low-level mutation operators. The idea behind both low-level operators is to interpret a single route (that is an ordered list of nodes) as an individual and apply to it standard GA operators. The routes resulting from low-level operators may not be feasible in the particular network topology. Therefore, the repair procedure is applied to their results. The details about dedicated operators employed by HEFAN 2.2 may be found in [13].

The second considered method is Multi Population Pattern Searching Algorithm for Flow Assignment in Non-bifurcated Commodity Flow (MuPPetS-FuN) [15], based on MuPPetS [8]. In MuPPetS-FuN two types of individual are employed. The classical GA-like individuals called *Competitive Templates* (CT) are used together with messy-coded individuals [2] called *viruses*. The messy-coded individuals do not encode the complete problem solution, but only a part of it. To rate such individual, the missing genes must be first supplemented. Therefore, each CT (complete problem solution encoded in GA-like manner) has a crowd of viruses assigned to it. The genes that are missing in each virus are supplemented from CT the virus is assigned to. Each virus population is separate, and the number of their populations is equal to the number of CTs. The virus population evolution process may be found as the optimization of the CT the viruses are assigned to - after their evolution is finished, if the fitness of best-found virus is better than its parental CT, then the genes from the virus infect the CT, improving it. The important feature of messy-coding concerning PEACh effect is that viruses encode only a small part of a complete problem solution. Thus, each virus may be interpreted as a modification of its parent CT. Since, $l_{vir} \ll size(CT)$, where l_{vir} is the genotype length of a virus, then the process of virus fitness computation seems to be suitable for using the benefits of PEACh effect.

V. Results

In this section, we present the results of experiments performed for two methods dedicated to solving the considered WP_LFL problem. HEFAN 2.2 is based on the idea of standard GA, while MuPPetS-FuN Active is a multi-population

method that uses messy-coding and dynamically manages the number of maintained subpopulations (i.e., during the method run new subpopulations are created, and some subpopulations are deleted). Since HEFAN 2.2 is based on standard GA idea, the use of PEACh effect will be limited only to mutation operator. On the other hand, MuPPetS-FuN uses messy-coding which is suitable for gaining the benefits brought by PEACh. The objective of the research is twofold. First, we wish to experimentally check how significant may be the optimization of computation load usage when PEACh is employed and how this optimization is dependent on the method type. Second, we wish to check how significantly the use of PEACh may influence the results quality for the considered WP_LFL problem.

The rest of this section is organized as follows. In the first subsection, we report the experiment setup. In the second subsection, we show and comment the differences in the computation load usage optimization brought by PEACh. The third subsection presents the influence of PEACh effect on the results quality for the two considered methods types. Finally, the last subsection contains the results discussion.

A. Experiment setup

In the experiments, the time-based stop condition was used. As pointed out in sections II and IV, since PEACh is employed, the use of FFE as a stop condition is doubtful for the research presented in this paper. The experiments were executed on PowerEdge R430 Dell Server Intel Xeon E5-2670 2.3 GHz 64GB RAM with Windows 2012 Server 64-bit installed. To ensure that the computation load used in each experiment is equal, the number of computation processes was always one less than a number of available CPU nodes. The time limit was set to 3 hours. All methods were programmed in C++, share all the possible pieces of code and are single-threaded. The experiments are executed in the clean environment – i.e., no other resource consuming processes are running, and the number of executed experiments is always one less than the number of available processor cores (1 core is spared for the operating system activities). Such assumptions are the same as in [7], [15], [8] and shall allow for fair comparison.

To compare the performance chosen methods we use ranking, defined as follows. The best method for a particular experiment receives the number of points equal to the number of competing methods; the second method receives one point less, etc. If more than one method takes the same place, then all such methods receive the same number of points as for one method. For instance, there are three competing methods A, B and C. Methods A and B were the best, and receive 3 points. Method C was the worst one – it receives 1 point.

The considered experiments may be grouped concerning the network type or flow parameters. Six different network topologies were considered. The networks parameters (minimal, maximal and average node degree) might be found as typical [19]. The mesh of five networks is irregular, the mesh of one network is grid-like. The parameters of all considered networks are presented in Table I.

TABLE I
CONSIDERED NETWORKS PARAMETERS

Network	**104**	**114**	**128**	**144**	**162**	**Grid**
Node number	36	36	36	36	36	36
Arc number	104	114	128	144	162	120
Minimal node deg.	2	2	3	3	3	2
Maximal node deg.	5	5	6	6	6	4
Average node deg.	2.89	3.17	3.56	4	4.5	3.33
Topology	Irregular mesh					Regular mesh

TABLE II
EXPERIMENT GROUPS PARAMETERS

Experiment group	**Group A**	**Group B**	**Group C**
Arc capacity	4800	4800	$km \cdot 1200$,where km=1,...,8
Connections to set	1260	2500	2500
Connection choice	1 for each pair	random	random
Demand size	equal for all connections	random	random

The classification using the demanded flow parameters divides the experiments into three groups: A, B and C. Each experiment group is characterized by arc capacities used, a number of connections to be set (demands) and the way the connections were chosen. The OC-12 and OC-48 standards, typical for transportation networks [4], were taken into consideration at the arc capacity design. The parameters for each experiment group are presented in Table II. The units of demands sizes and arc capacities were abandoned which is frequent for the papers concerning the problems of flow assignment in computer networks [11], [13], [15].

Ten experiments were used for each network and experiment group. Thus, the total number of test cases was $6 \cdot 3 \cdot 10 = 180$. The considered problem is NP-complete [4]. As presented in table II the number of demands is 1260 or 2500. In the employed solution encoding each gene refers to a single demand, which makes gene number equal to demand number. If for each demand we consider 10 different routes, then the number of solutions that may be encoded is equal to 10^{1260} or 10^{2500}.

The methods settings were adopted from [15] and are presented in Tables III and IV.

B. PEACh benefits depending on method type

One of the main objectives of this paper is to check how significant is the influence of PEACh effect on computation load optimization for the considered problem depending on

TABLE III
MUPPETS-FUN ACTIVE SETTINGS

Parameter name	**MuPPetS-FuN Active**
Virus generations	100
Virus subpopulation size	100
Cut	0.21
Splice	0.30
Mutation	0.20
Low Level Crossover	0.60
Low Level Mutation	0.20

TABLE IV
HEFAN 2.2 SETTINGS

Parameter name	**HEFAN 2.2**
Population size	1000
Crossover	0.9
Mutation	0
Low level crossover	0.1
Low level mutation	0.4
Uniform crossover	0.5
Crossover	0.9

TABLE V
FFE INCREASE RATIO CAUSED BY PEACH EFFECT IN MUPPETS-FUN ACTIVE RUNS

	Avr	**St. dev.**	**Min**	**Max**	**Mean**
All	207.92%	48.20%	123.14%	311.01%	207.18%
104	241.78%	34.17%	188.03%	311.01%	239.90%
114	216.76%	34.42%	170.45%	279.07%	207.64%
128	212.07%	23.90%	171.90%	265.80%	211.09%
144	158.89%	22.02%	129.44%	197.67%	157.62%
162	149.49%	20.17%	123.14%	187.52%	143.62%
Grid	255.66%	21.37%	225.07%	291.69%	261.73%
Group A	174.72%	39.80%	123.14%	251.10%	178.37%
Group B	192.74%	40.34%	140.48%	276.49%	205.40%
Group C	220.23%	45.64%	144.99%	311.01%	213.23%

the method type. Therefore, we compare the amount of FFE done by MuPPetS-FuN Active and HEFAN 2.2 when using and not using the benefits of PEACh effect. To assure that the comparison is precise, each experiment was executed with the same random seed. The runs in which the optimal solution was found were excluded from the comparison. The FFE increase ratio (FFE for the run using PEACh effect is divided by FFE for the run without PEACh) is presented in Tables V and VI.

The results presented in Tables V and VI show that using PEACh effect is far more beneficial for MuPPetS-FuN Active. For MuPPetS-FuN, the minimum increase obtained in all 180 runs is over 123%, while the maximum ratio for HEFAN 2.2 is less than 112%. Such results are expected since during most of its run MuPPetS-FuN process messy-coded individuals, which genotypes are much shorter than the genotype of full GA-like coded problem solution. Thus, PEACh effect benefits may be used at every fitness computation of messy-coded individual. On the other hand, for HEFAN 2.2 whish is based on a standard GA, PEACh is only beneficial during mutation.

It is also interesting that for MuPPetS-FuN Active PEACh

TABLE VI
FFE INCREASE RATIO CAUSED BY PEACH EFFECT IN HEFAN 2.2 RUNS

	Avr	**St. dev.**	**Min**	**Max**	**Mean**
All	107.98%	1.98%	104.14%	111.95%	108.14%
104	109.08%	2.24%	104.21%	111.95%	109.98%
114	107.91%	1.94%	104.64%	111.05%	108.04%
128	107.42%	2.37%	104.14%	111.76%	106.24%
144	107.71%	1.58%	105.17%	110.41%	108.48%
162	107.13%	0.94%	105.50%	108.79%	107.18%
Grid	107.98%	1.84%	104.43%	110.69%	108.56%
Group A	105.93%	1.06%	104.21%	108.42%	105.78%
Group B	109.12%	1.93%	104.14%	111.76%	109.14%
Group C	108.65%	1.45%	105.27%	111.95%	108.65%

TABLE VII
THE INFLUENCE OF PEACH EFFECT ON METHOD EFFECTIVENESS (RANKING)

		PEACh		**No PEACh**	
	LRH	**MuPPetS**	**HEFAN**	**MuPPetS**	**HEFAN**
All	2.15	**2.98**	2.33	2.62	2.12
104	2.43	**3.40**	2.97	2.70	2.77
114	1.77	**3.00**	2.03	2.60	1.87
128	1.87	**3.03**	2.10	2.43	1.80
144	1.43	**2.90**	2.50	**2.90**	2.17
162	1.80	2.43	2.37	**2.53**	2.13
Grid	**3.60**	3.13	2.03	2.53	1.97
Group A	1.12	**2.40**	2.34	2.30	2.38
Group B	1.58	2.50	**2.58**	2.02	2.36
Group C	2.88	**3.96**	2.26	3.58	1.70

effect is significantly more beneficial for some test case groups. For instance, the highest FFE increase ratio is found for networks *Grid* and *104*. On the other hand, FFE increase ratio is the lowest for experiments using networks *144* and *162*. The detailed analysis of this phenomenon is out of this paper scope and is one of the interesting future work directions. The reasonable explanation seems to be that in the experiments for networks *144* and *162* the average length of messy-individual genotypes is significantly longer than in the experiments for *Grid* and *104* networks. The longer is the messy-coded individual's genotype, the less significant is the fitness function computation speed-up caused by PEACh effect. This observation seems to be supported by the FFE ratio observed experiments in groups A, B, and C. The full GA-like encoded solution in experiments from group A contains 1260 genes, while in experiments in groups B and C this is 2500 genes. We may expect that in experiments of from group A, the average genotype of messy-individual contains a larger percentage of all necessary genes than in two other groups.

C. PEACh influence on results quality

In this section, we compare the effectiveness of both considered methods, depending on PEACh effect. In these experiments the randomizer seed was random. Thus, it is possible that a method version with PEACh may return a lower quality result than a version without PEACh. Note, that it was impossible for the experiments considered in the previous subsection because when the seed is set manually both method versions (with and without) were performing the same run. The only difference was that version with PEACh was working faster, so it was able to perform a higher number of iterations. The comparison of results quality in Table VII. In the experiments presented in this subsection the MuPPetS-FuN and HEFAN 2.2 are also compared with Lagrangian Relaxation Heuristic (LRH) [13], [15]. LRH is a hybrid algorithm that joins the Flow Deviation for Primary Routes algorithm [1] and Lagrangian Relaxation. It uses sub-gradient optimization to determine Lagrangian coefficients and was shown effective [13] in solving flow assignment problems.

As presented in Table VII the methods employing PEACh effect are the most effective for all experiment groups except two. The first is the Grid network group for which the most

TABLE VIII
THE COMPARISON OF MUPPETS-FUN EFFECTIVENESS WITH AND WITHOUT PEACH EFFECT ON THE BASE OF p-VALUE REPORTED BY SIGN TEST

	with PEACh better or equal	Equal	with PEACh worse or equal
All	**100.00%**	0.00%	0.00%
104	**99.95%**	0.37%	0.19%
114	**98.94%**	7.68%	3.84%
128	**100.00%**	0.07%	0.04%
144	59.27%	**100.00%**	59.27%
162	50.00%	**100.00%**	68.55%
Grid	**99.88%**	0.94%	0.47%
Group A	**92.52%**	28.10%	14.05%
Group B	**99.99%**	0.06%	0.03%
Group C	**100.00%**	0.01%	0.00%

TABLE IX
THE COMPARISON OF HEFAN 2.2 EFFECTIVENESS WITH AND WITHOUT PEACH EFFECT ON THE BASE OF p-VALUE REPORTED BY SIGN TEST

	with PEACh better or equal	Equal	with PEACh worse or equal
All	**99.98%**	0.07%	0.04%
104	**96.08%**	18.92%	9.46%
114	**92.83%**	33.23%	16.62%
128	**98.94%**	7.68%	3.84%
144	**96.82%**	16.71%	8.35%
162	**98.46%**	9.63%	4.81%
Grid	73.83%	**83.18%**	41.59%
Group A	50.00%	**100.00%**	64.94%
Group B	**99.00%**	4.70%	2.35%
Group C	**99.98%**	0.09%	0.04%

effective is LRH, the second is 162 network for which the most effective MuPPetS-FuN Active without PEACh effect. In the experiments employing the Grid network, LRH simply seems to be a more suitable method. Such observation is consistent with the previous research in this area [15]. For experiments using 162 network, the explanation may be as follows. MuPPetS-FuN Active seems to be effective in solving test cases from this group. The PEACh influence on computation load used by MuPPetS-FuN Active was the lowest for these experiments (less than 150% of average ratio presented in Table V), so it is likely that such result is the effect of noise. To check if the benefits of PEACh effect are statistically significant we have used Sign Test. The results are presented in Table VIII.

As presented in Table VIII the use of PEACh effect is not statistically significant for the method effectiveness for experiments using network *144* and *162*. Such results are not surprising since for both of these subgroups the average FFE increase ration was the lowest. The third subgroup for which the influence of PEACh benefits on results quality does not seem statistically significant are experiments from Group A. Note, that for this subgroup average FFE increase ratio was the third lowest one.

The same statistical tests were performed for HEFAN 2.2. As expected, for this method the results are less convincing - the p-value of tests checking that HEFAN 2.2 effectiveness with and without PEACh is equal is significantly higher and is above 9% for 6 of 10 experiment groups. Nevertheless, when all experiments are taken into account, the results are decisive. Thus, it is allowed to state that for HEFAN 2.2 the influence of PEACh significantly affects the results quality.

D. Results discussion

The results presented in this paper show that the influence of PEACh effect may significantly optimize the computation load used by a method. Thus, it may influence the method effectiveness when the available resources are limited. The research presented in this paper show that some methods are more suitable to use PEACh benefits than other. Here, MuPPetS-FuN Active that employs messy-coding was able to perform from 123% up 311% of FFE that would be computed without using PEACh. Such change seems significant. Note, that messy-coding is not the only mechanism suitable for using PEACh. For instance, the evolutionary method may be hybridized with the local search algorithm. If the local search is based on exchanging one gene value to the other, the optimization of computation load brought by using PEACh effect may be significant as well. In this paper we also show that for some methods like standard GA the use PEACh is limited. However, its influence may also lead to statistically significant effectiveness increase.

Another interesting observation refers to computation load measurement. When an evolutionary method is applied to solve a practical problem, it seems reasonable to use PEACh effect if possible. However, when PEACh is used the FFE is not a fair computation load measure because the cost of computing a single fitness evaluation may be significantly different depending on the situation in which it is computed - if fitness is computed after a small genotype change (eg. after mutation) the cost will be low, in other cases it will be high. Thus, in such cases, FFE is not a reliable computation load measure.

VI. CONCLUSION

The objective of this paper was to check how significant may be the optimization of computation load expenses when PEACh effect is employed by methods applied to solve hard, practical problem. The presented results show that the influence of PEACh may be significant even if the method is not suitable to employ it (HEFAN 2.2). On the other hand, for methods using mechanisms like messy-coding, the efficiency increase may exceed 300%. The results supported by statistical tests point out that the result quality differences caused by PEACh are significant. Thus, for the considered methods FFE is not a fair computation load measure.

The key direction of future research is further investigation of possible PEACh utilization, for instance in hybrid methods using local optimization to improve their effectiveness. The use of PEACh may also enable significant effectiveness breakthrough for methods employing the Baldwin effect [16] as it may significantly reduce its computational costs. Finally, new techniques that use problem features for computation load reduction shall be identified and proposed.

ACKNOWLEDGMENT

This work was supported by the Polish National Science Centre (NCN) under Grant 2015/19/D/ST6/03115.

REFERENCES

[1] L. Fratta M. Gerla, L. Kleinrock, "The Flow Deviation Method: An Approach to Store-and-Forward Communication Network Design," *Networks*, vol. 3, no. 2, 1973, pp. 97-133.

[2] D.E. Goldberg, B. Korb, K. Deb, "Messy genetic algorithms: Motivation, analysis, and first results," *Complex Systems*, vol. 3, 1989, pp. 493-530.

[3] B. W. Goldman, W. F. Punch, "Parameter-less Population Pyramid," *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation (GECCO '14)*, ACM, 2014, pp. 785-792.

[4] W. D. Grover, "Mesh-based Survivable Transport Networks: Options and Strategies for Optical, MPLS, SONET and ATM Networking," Prentice Hall PTR, New Jersey, 2004.

[5] S.-H. Hsu, T.-L. Yu, "Optimization by Pairwise Linkage Detection, Incremental Linkage Set, and Restricted / Back Mixing: DSMGA-II," *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation (GECCO '15)*, ACM, 2015, pp. 519-526.

[6] D.R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization*, vol. 21, 2001, pp.345-383.

[7] M. M. Komarnicki, M. W. Przewozniczek, "The influence of fitness caching on modern evolutionary methods and fair computation load measurement," *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '18)*, ACM, 2018, (in press).

[8] H. Kwasnicka, M. Przewozniczek, "Multi Population Pattern Searching Algorithm: a new evolutionary method based on the idea of messy Genetic Algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 5, pp. 715-734, 2011.

[9] P.B. Myszkowski, M. Przewozniczek, M. Skowronski, "Constructive heuristics for technology-driven Resource Constrained Scheduling Problem," *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, 2015.

[10] P.B. Myszkowski, M. Skowronski, L. Olech, K. Oslizlo, "Hybrid Ant Colony Optimization in solving Multi-Skill Resource-Constrained Project Scheduling Problem," *Soft Computing*, vol. 19, issue 12, 2015, pp 3599-3619.

[11] M. Pioro, D. Medhi, "Routing, Flow, and Capacity Design in Communication and Computer Networks," Morgan Kaufmann Publishers, 2004.

[12] R. J. Povinelli, X. Feng, "Improving Genetic Algorithms Performance By Hashing Fitness Values," *Artificial Neural Networks in Engineering*, 1999, 399-404.

[13] M. Przewozniczek, K. Walkowiak, "Quasi-hierarchical Evolutionary Algorithm for Flow Optimization in Survivable MPLS Networks ," *Lecture Notes in Computer Science,* vol. 4707, Springer Verlag, 2007, pp. 330-342.

[14] M. Przewozniczek, R. Goscien, K. Walkowiak, M. Klinkowski, "Towards Solving Practical Problems of Large Solution Space Using a Novel Pattern Searching Hybrid Evolutionary Algorithm - An Elastic Optical Network Optimization Case Study" in Expert Systems with Applications," vol. 42, 2015, pp. 7781-7796.

[15] M. Przewozniczek, "Active Multi Population Pattern Searching Algorithm for Flow Optimization in Computer Networks - the novel coevolution schema combined with linkage learning," *Information Sciences*, vol. 355-356, 2016, pp. 15-36.

[16] M. W. Przewozniczek, K. Walkowiak, M. Aibin, "The evolutionary cost of Baldwin effect in the routing and spectrum allocation problem in elastic optical networks," *Applied Soft Computing*, vol. 52, 2017, pp. 843-862.

[17] M. W. Przewozniczek, "Problem Encoding Allowing Cheap Fitness Computation of Mutated Individuals," *Proceedings of 2017 Congress on Evolutionary Computation (CEC 2017)*, 2017, pp. 308-316.

[18] D. Thierens, P. A. N. Bosman, "Hierarchical Problem Solving with the Linkage Tree Genetic Algorithm," *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation (GECCO'13)*, 2013, pp. 877-884.

[19] F. Zhang , X. Zheng, H. Zhang, Y. Guo, "A kind of topology aggregation algorithm in hierarchical wavelength-routed optical networks," *Photonic Network Communications*, vol. 9, 2005, pp. 167-180.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 67–72
DOI: 10.15439/2018F88
ISSN 2300-5963 ACSIS, Vol. 15

Representation Matters: An Unexpected Property of Polynomial Rings and its Consequences for Formalizing Abstract Field Theory

This paper is dedicated to Grzegorz Bancerek.

Christoph Schwarzweller
Intstitute of Informatics, Faculty of Mathematics, Physics and Informatics,
University of Gdańsk,
Wita Stwosza 57, 80-952 Gdańsk, Poland
Email: schwarzw@inf.ug.edu.pl

***Abstract*—In this paper we develop a Mizar formalization of Kronecker's construction, which states that for every field F and irreducible polynomial $p \in F[X]$ there exists a field extension E of F such that p has a root over E. It turns out that to prove the correctness of the construction the field F needs to provide a disjointness condition, namely $F \cap F[X] = \emptyset$. Surprisingly this property does not hold for arbitrary representations of a field F: We construct for almost every field F another representation F', i.e. an isomorphic copy F' of F, not satisfying this condition. As a consequence to F' our formalization of Kronecker's construction cannot be applied.**
All proofs have been carried out in the Mizar system. Based on Mizar's representation of the fields $\mathbb{Z}_p, \mathbb{Q}$ and $\mathbb{R}$ we also have proven that $\mathbb{Z}_p \cap \mathbb{Z}_p[X] = \emptyset$, $\mathbb{Q} \cap \mathbb{Q}[X] = \emptyset$, and $\mathbb{R} \cap \mathbb{R}[X] = \emptyset$ respectively.

I. Introduction

INTERACTIVE theorem proving aims at developing systems to be used to formalize, that is both formulate and prove, mathematical theorems and theories in an accurate and comfortable way. The ultimate dream is a system containing all mathematical knowledge in which mathematicians develop and prove new theorems. To come at least a little closer to this goal much effort has been spent building large repositories of computer-verified theorems such as the Coq library [4], the Isabelle2017 library [15], and the Mizar Mathematical Library [17]. A number of important mathematical theorems have been proven to illustrate the capability of interactive theorem proving, the most prominent examples being the proof of Kepler's conjecture in HOL Light [13], the Feit-Thompson theorem in Coq, and the Jordan curve theorem in Mizar (see also [25]).

Another interesting challenge in this context is Artin's solution of Hilbert's 17th problem, which asks whether a (multivariate) polynomial taking only non-negative values over the real numbers can be represented as a sum of squares of rational functions. Its formalization requires the development of real algebra: the theory of ordered fields and in particular the notion of field extensions and field adjunctions [23]. A key tool in field theory is Kronecker's construction which states that for every field F and every non-constant polynomial $p \in F[X]$ there exists a field extension E of F in which p has a root. The Mizar formalization of Kronecker's construction is the topic of this paper.

One dominating subject in abstract field theory is the construction of new larger fields containing the field (or ring) one has started with, for example constructing $\mathbb{C}$ from $\mathbb{R}$ or $F(X)$ from $F[X]$. Here only the general structure of the field, not the individual representation of the field's elements, is of interest; isomorphic fields are just considered to be the same field. For example, when constructing the complex numbers by $\mathbb{R}[X]/(X^2+1)$ the result is not the usual field $\mathbb{C}$ of complex numbers, yet we have $\mathbb{R}[X]/(X^2+1) \cong \mathbb{C}$. Also $\mathbb{R}[X]/(X^2+1)$ does not contain $\mathbb{R}$ as a subfield, but an isomorphic copy of it. From a mathematician's point of view this of course does not matter, because you can get what you want by exchanging the isomorphic fields. This kind of argument is omnipresent in abstract field theory. In fact, Kronecker's construction contains a similar argument and we will see that exactly this is the hardest part in the formalization: to carry it out, we need the disjointness condition $F \cap F[X] = \emptyset$ already mentioned in the abstract.

The plan of the paper is as follows. In the next section we give a brief overview of the Mizar system and illustrate how algebraic domains are constructed. In section III we discuss our Mizar formalization of Kronecker's construction placing emphasis on the disjointness condition. In the subsequent section we present the construction of fields which do not fulfill our disjointness condition. It turns out that our construction works for every field except $\mathbb{Z}_2$. In section V and VI we provide an intuitive, though not really helpful, condition implying $F \cap F[X] = \emptyset$ and finally prove the disjointness condition for Mizar's representation of the fields $\mathbb{Z}_p, \mathbb{Q}$, and $\mathbb{R}$. In the conclusion we discuss how a formalization of Kronecker's construction without a disjointness condition might look like.

II. The Mizar System

Mizar has often been described in the literature, for example in [3], [18], [12], [9], [8] and [11]. In this paper we will present only theorems, not proofs; therefore we give only a very rough description of Mizar. Mizar's logical basis is classical first-order logic, extended with so-called schemes. Schemes

introduce free second-order variables enabling the definition of induction schemes among others. In addition, Mizar objects are typed, the types forming a hierarchy with the fundamental type `set`. The user can introduce new (sub)types describing mathematical objects such as groups, fields, vector spaces, or polynomials over rings or fields. The development of the Mizar Mathematical Library relies on Tarski-Grothendieck set theory - a variant of Zermelo-Fraenkel set theory using Tarski's axiom about arbitrarily large, strongly inaccessible cardinals which can be used to prove the axiom of choice. Mizar proofs are written in natural deduction style. The rules of the calculus are connected with corresponding (English) natural language phrases so that the Mizar language is close to the one used in mathematical textbooks [10].

To define algebraic domains Mizar provides so-called structure modes fixing the domain's sets of elements and operations. So, for example[1]

```
definition
struct (addLoopStr,multLoopStr_0) doubleLoopStr
 (# carrier -> set,
    addF, multF -> BinOp of the carrier,
    OneF, ZeroF -> Element of the carrier #);
end;
```

defines the necessary backbone of rings and fields. Note that `doubleLoopStr` inherits from both `addLoopStr` and `multLoopStr_0`, that is it joins the operations of additive and multiplicative groups. Properties such as commutativity or the existence of inverse elements are described by attribute definitions such as

```
definition
let R be addLoopStr;
attr R is right_zeroed means
  for a being Element of R holds a + 0.R = a;
end;
```

Here for elements `a` and `b` of (the `carrier`) of `R` `a+b` is a shortcut for `(the addF of R).(a,b)`. A field then is a `doubleLoopStr` with the appropriate collection of attributes (compare [21], [19]).

```
definition
mode Field is
  Abelian add-associative right_zeroed
  right_complementable associative commutative
  well-unital almost_left_invertible
  distributive non empty doubleLoopStr;
end;
```

As a consequence a Mizar object of type `Field` obtains all properties described by the defining attributes. We note here, that Mizar types have to be non-empty.

Concrete algebraic domains are built by instantiation of structures. The field of rational numbers $\mathbb{Q}$, for example, is given by the set `RAT` of rational numbers and operations `addrat` and `multrat` defining addition and multiplication for elements of `RAT`. These are then glued together by the following

[1]Throughout the paper Mizar code is written in verbatim style.

```
definition
func F_Rat -> Field equals
  doubleLoopStr(#RAT,addrat,multrat,1,0#);
end;
```

Note, that the definition of the set `RAT` gives a particular representation of the rational numbers $\mathbb{Q}$; to be used when arguing about the rational numbers using the field `F_Rat`. There are other fields, that is fields with a different set of elements, isomorphic to $\mathbb{Q}$. In fact any field of characteristic 0 contains a subfield isomorphic to $\mathbb{Q}$.

III. Kronecker's Construction

In this section we discuss our Mizar formalization of Kronecker's construction, sometimes also called the main theorem of field theory. It can be stated as follows [24], [23]:

Theorem
Let F be a field and $p \in F[X]$ irreducible. Then there exists a field extension E of F such that p has a root over E.

Note that from this theorem easily follows the existence of such an extension for every non-constant $p \in F[X]$.

A. Field Extensions

We begin with the Mizar definition of field extensions: A field E is a field extension of a field F, if F is a subfield of E [24], or equivalently if F is a subset of E, which itself is a field. It is understood that the operations $+$ and $*$ in F are the restrictions of $+$ and $*$ in E. In Mizar this is stated as follows (see [6]).

```
definition
let F be Field;
mode Subfield of F -> Field means
  the carrier of it c= the carrier of F &
  the addF of it = (the addF of F)
                        || the carrier of it &
  the multF of it = (the multF of F)
                        || the carrier of it &
  1.it = 1.F & 0.it = 0.F;
end;
```

The mode `Subring of R`, where `R` is a ring is defined analogously. Based on this definition we can introduce field extensions as follows.

```
definition
let R be Ring, E be Field;
attr E is R-field-extending means
  R is Subring of E;
end;

definition
let F be Field;
mode FieldExtension of F is
                     F-field-extending Field;
end;
```

Note that instead of postulating that F is a subfield of E we demand that a ring R is a subring of E. This way our definition gets more flexible. For example, this allows to show that $\mathbb{Q}$ extends $\mathbb{Z}$. For a field F, however, our definition is equivalent

to the one from the literature given above, in particular one proves that

```
theorem
for F,E being Field
holds E is FieldExtension of F iff
      F is Subfield of E;
```

In any case the definition implies that a field E in order to be a field extension of a field F in particular must contain the elements of F, e.g. we must have `the carrier of` $F \subseteq$ `the carrier of` E as sets.

B. The Construction

Kronecker's proof is constructive [24] and consists of two parts: The first one observes is that if p is irreducible then $<p>$ is a maximal ideal in $F[X]$, and hence $E := F[X]/<p>$ is a field. The second step consists of showing that $[X]_{<p>}$ is a root of p in $E[X]$: If $p = a_0 + a_1 * X + \ldots a_n * X^n$ we get

$$\begin{aligned} p([X]) &= a_0 + a_1 * [X] + \ldots a_n * [X]^n \\ &= a_0 + a_1 * [X] + \ldots a_n * [X^n] \\ &= a_0 + [a_1 * X] + \ldots [a_n * X^n] \\ &= [a_0 + a_1 * X + \ldots a_n * X^n] \\ &= [p] \\ &= 0. \end{aligned}$$

Between these two steps one usually finds a remark that $F[X]/<p>$ is a field extension of F. Note that formally F is no subfield of $F[X]/<p>$ just because $F \not\subseteq F[X]/<p>$ as sets; and therefore $p \in F[X]$ is not even a polynomial over $F[X]/<p>$. However, $\varphi : F \longrightarrow F[X]/<p>$ given by $a \mapsto [a()]_{<p>}$ is a monomorphism, the so-called canonical monomorphism, and gives rise to the embedding of F into $F[X]/<p>$. The remark mentioned above then reads

> *We can identify φF with F in $F[X]/<p>$ and thus consider F as a subfield of $F[X]/<p>$.*

The Mizar formalization of the two steps is quite straightforward. The quotient field $F[X]/<p>$ has been defined in [16], [20] and $p([X]) = [p]$ can be easily shown by induction on the degree of p.

The main task is to formalize the aforementioned remark: Formally, identifying φF with F in a field E if $\varphi : F \longrightarrow E$ is a monomorphism means defining a new carrier $K := (E \setminus \varphi F) \cup F$ and modifying addition and multiplication of F appropriately. For example, $a + b$ for two elements a and b of K where $a \in F$ and $b \in E \backslash \varphi F$ actually means adding $\varphi a + b$ in E. The result $a + b$ then either is $\varphi a + b$ if this is not in φF or $\varphi^{-1}(\varphi a + b)$ if this is in φF. In this way we get a new field K isomorphic to E with $F \subseteq E$:

Theorem
Let F, E be fields and $\varphi : F \longrightarrow E$ a monomorphism. Then $K := (E \backslash \varphi F) \cup F$ is a field isomorphic to E. Moreover F is a subfield of K.

This field K then is the desired field extension for Kronecker's construction. Unfortunately we were not able to prove the theorem in this general setting: the problem is that there might be elements in E, more precisely in $E \backslash \varphi F$, already appearing in F, that is $F \cap (E \backslash \varphi F)$ might be non-empty. This leads to an identification of elements during the construction, which destroys the isomorphism between $(E \backslash \varphi F) \cup F$ and E. This has to be excluded, so we require a disjointness condition. We hence come up with two slightly weaker theorems in Mizar. Here E being a F-monomorphic field just means that there exists a monomorphism $\varphi : F \longrightarrow E$ and `emb f` is the field K defined above.

```
theorem
for F being Field,
    E being F-monomorphic Field
st F /\ E = {}
for f being Monomorphism of F,E
holds E,(emb f) are_isomorphic;

theorem
for F being Field,
    E being F-monomorphic Field
st F /\ E = {}
ex E' being Field st E',E are_isomorphic &
                     F is Subfield of E';
```

The Mizar proofs are straightforward, but quite tedious due to the number of different cases. Now our Mizar version of Kronecker's construction has to take into account the disjointness condition leading to the following theorem.

```
theorem
for F being Field,
    p being non constant
                  Element of Polynom-Ring F
st F /\ (Polynom-Ring F)/({p}-Ideal) = {}
ex E being FieldExtension of F
st p is_with_roots_in E;
```

The theorem's condition

```
F /\ (Polynom-Ring F)/({p}-Ideal) = {},
```

i.e. $F \cap F[X]/<p> = \emptyset$, is not really satisfying: it depends not only on the field F, but also on the given polynomial $p \in F[x]$. This can be improved by carrying out Kronecker's construction using another representation of $F[X]/<p>$: the isomorphic copy consisting of all polynomials $f \in F[X]$ with deg $f <$ deg p (see [24]). We denote this representation by $F[p]$. For $F[p]$ we have in particular $F[p] \subseteq F[X]$ as sets, so that the condition $F \cap F[X] = \emptyset$ suffices to apply the embedding theorems from above. With `polynomial_disjoint` denoting $F \cap F[X] = \emptyset$ we now get the following Mizar version of Kronecker's construction:

```
theorem
for F being polynomial_disjoint Field,
    p being non constant
                  Element of Polynom-Ring F
ex E being FieldExtension of F
st p is_with_roots_in E;
```

IV. Constructing Negative Examples

In the last section we discussed a Mizar formalization of Kronecker's construction and ended up with a version that does not hold for all fields in the first place: To apply Kronecker's construction to a given field F we have to ensure that $F \cap F[X] = \emptyset$.

At first glance this condition should be no restriction. Intuitively a polynomial $p \in F[X]$ is a more complex object than an element a of the underlying field F. In Mizar a polynomial $p \in F[X]$ is defined as a function $p : \mathbb{N} \longrightarrow F$, which returns the coefficients of p: for $n \in \mathbb{N}$ $p.n$ denotes the coefficient of X^n. Therefore it should be easy to show that $a \neq p$ and hence $F \cap F[X] = \emptyset$ for an arbitrary field F.

This, unfortunately, is not true in general. Of course it is easy to show $a \neq p$ if p includes a as a coefficient, that is $p.n = a$ for some $n \in \mathbb{N}$. This, however, does not exclude the existence of fields F with $F \cap F[X] \neq \emptyset$, and in the following we will construct for every field F, except for $\mathbb{Z}_2$, an isomorphic copy F' of F with $F' \cap F'[X] \neq \emptyset$.

A. A First Example

Perhaps the easiest example is a three-element field isomorphic to $\mathbb{Z}_3$. One takes 0 and 1 and sets

$$F' := \{0,\ 1,\ X\}$$

where X is the identity polynomial. The idea is that the polynomial X as a function $\mathbb{N} \longrightarrow F'$ is

$$X.i = \begin{cases} 1; & i = 1 \\ 0; & i \neq 1 \end{cases}$$

Therefore, having $0 \in F'$ and $1 \in F'$ we can built this function (over F') and hence $X \in F' \cap F'[X]$. The operations $+$ and $*$ of F' are just defined to mimic the ones of $\mathbb{Z}_3$ with X playing the role of 2. As a result we have an isomorphic copy of $\mathbb{Z}_3$ our Mizar version of Kronecker's construction cannot be applied to.

B. A Class of Negative Examples

The idea of the first example can be generalized to almost arbitrary fields F by observing that we in fact changed the representation of $\mathbb{Z}_3$ by just exchanging 2 with the polynomial X. This works for almost every field F; $\mathbb{Z}_2$ is the only exception. One can exchange an arbitrary element $a \in F \backslash \{0, 1\}$ with another arbitrary object o by setting

$$F_{a,o} := (F \backslash \{a\}) \cup \{o\}.$$

Defining $+$ and $*$ appropriately $F_{a,o}$ then is an isomorphic copy of F for an arbitary object o. Substituting X for o now shows that $X \in F_{a,X} \cap F_{a,X}[X]$ and gives the Mizar

```
theorem
for F being non almost_trivial Field
ex F' being non polynomial_disjoint Field
st F',F are_isomorphic;
```

Here, the property `non almost trivial` excludes $\mathbb{Z}_2$. In other words, for every field F (except for $\mathbb{Z}_2$) we constructed a representation of F our Mizar version of Kronecker's construction cannot be applied to.

Note also that X is non-constant and hence $X \notin \varphi F$, so that identifying φF with F will not adjust the intersection. In fact - as o is arbitrary - one can substitute o with the polynomial X^n for $n \in \mathbb{N}^+$. X^n as a function is

$$(X^n).i = \begin{cases} 1; & i = n \\ 0; & i \neq n \end{cases}$$

so an analogous argument shows $X^n \in F_{a,X^n} \cap F_{a,X^n}[X]$. Hence, we get the following

```
theorem
for F being non almost_trivial Field
for n being non zero Nat
ex F' being non polynomial_disjoint Field,
   p being Polynomial of F'
st F',F are_isomorphic &
   deg p = n &
   p in (the carrier of F') /\
              (the carrier of Polynom-Ring F');
```

so that the degree of the polynomial p in the intersection $F' \cap F'[X]$ is not bounded.

As the main result from our counterexamples we get that $F' \cap F[X] = \emptyset$ is a property not invariant under isomorphisms (of fields). Consequently the application of Kronecker's theorem depends on the representation of the given field F.

V. An Intuitive "Solution"

In the last section it turned out that in order to apply Kronecker's construction with a given field F we have to ensure that $F \cap F[X] = \emptyset$, depending on the actual representation of F. This is in particular true for the basic fields $\mathbb{Q}, \mathbb{R}$, and $\mathbb{Z}_p$, where p is prime: it becomes important how these fields are represented in Mizar.

A first approach to solve this problem again relies on the intuitive feeling that a polynomial is a more complex object than an element of the underlying field. So, if all elements of a field F are of the same complexity, then $F \cap F[X]$ should be empty just because all polynomials $p \in F[X]$ are more complex than - and therefore are not equal to - all elements $a \in F$. Note, that our counterexamples from section IV do not fulfill this condition.

A possibility to measure the complexity of mathematical objects o is the so-called rank of o (see [5]). Here every object o is understood as a set and the rank of o is the least ordinal number greater than the rank of every member of the set o. In Mizar the notion of rank has been formalized as a function `the_rank_of` from objects into ordinal numbers [1]. Using this function we define

```
definition
let F be Field;
attr F is flat means
  for a,b being Element of F
  holds the_rank_of a = the_rank_of b;
end;
```

to express that all elements of a field F are of the same complexity. As already mentioned a Mizar polynomial over F is defined as a function $p : \mathbb{N} \longrightarrow F$, i.e. formally is a set of pairs $p = \{ [n, p.n] \mid n \in \mathbb{N}\}$.[2] From this immediately follows that if F is flat, then the rank of all polynomials $p \in F[X]$ is greater than the rank of all elements $a \in F$ and thus

```
theorem
for F being flat Field
              holds F is polynomial_disjoint;
```

Note that in particular the definition of functions in terms of set of pairs enabled the proof of this theorem.

Unfortunately the criterion of being flat is not really helpful, as it does not apply to standard representations of fields. The reason is that in Mizar 0 is defined as the empty set - and the empty set is the only mathematical object of rank 0. Consequently every field containing 0 is non-flat, so that in particular $\mathbb{Q}, \mathbb{R}$, and $\mathbb{Z}_p$ are non-flat.

VI. Some Positive Examples

To provide some examples our Mizar version of Kronecker's construction can be applied to, we prove $\mathbb{Z}_p \cap \mathbb{Z}_p[X] = \emptyset$, $\mathbb{Q} \cap \mathbb{Q}[X] = \emptyset$, and $\mathbb{R} \cap \mathbb{R}[X] = \emptyset$ by hand. To do so we have to rely on Mizar's representation of these fields: essentially we have to check the definitions. Note again, that for a field F the condition $F \cap F[X] = \emptyset$ suffices to apply our Mizar version of Kronecker's construction, because our formalization uses $F[p]$ instead of $F[X]/<p>$.

We already mentioned that a Mizar polynomial over F is a function $p : \mathbb{N} \longrightarrow F$, that is $p = \{[n, p.n] \mid n \in \mathbb{N}\}$ as a set. We now need to show that no Mizar polynomial p can equal any Mizar number $r \in \mathbb{R} \supseteq \mathbb{Q} \supseteq \mathbb{Z} \supseteq \mathbb{N}$. To be more precise, this has to be shown for the Mizar sets `REAL`, `RAT`, `INT`, and `NAT`, which have been used to define the fields `INT.Ring p`, `F_Rat`, and `F_Real`. To keep the following more readable we will, however, continue writing $\mathbb{N}$ for `NAT`, $\mathbb{Z}$ for `INT`, and so on.

In Mizar all numbers beginning with $\mathbb{N}$ are constructed from sets following the well-known set-theoretic approaches. So for $\mathbb{N}$ we find $0 = \emptyset,\ 1 = \{0\},\ 2 = \{0, 1\}, \ldots$ and in general $n = \{m \mid m < n\}$ for $n, m \in \mathbb{N}$.

Because the carrier of $\mathbb{Z}_n$ equals $\{0, 1, \ldots n-1\} \subset \mathbb{N}$ we already can show $\mathbb{Z}_n \cap \mathbb{Z}_n[X] = \emptyset$. For if we have $p = n$ for a polynomial p and a natural number n it follows that $\{[i, p.i] \mid i \in \mathbb{N}\} = \{m \mid m < n\}$, hence there is a $j \in \mathbb{N}$ smaller than n such that $j = [n, p.n] = \{\{n\}, \{n, p.n\}\}$. Then, because j is a natural number, j must equal $\{0, 1\} = \{\emptyset, 1\}$,[3] but neither $\{n\}$ nor $\{n, p.n\}$ equals $\emptyset$, a contradiction.

The proofs of polynomial disjointness for $\mathbb{Z}, \mathbb{Q}$ and $\mathbb{R}$ use similar set-based argumentations. To give an impression how Mizar's set-based definition of numbers is used, we briefly discuss the case of $\mathbb{Q}$. In Mizar first the non-negative rational numbers $\mathbb{Q}^+$ are introduced as pairs of natural numbers, i.e. elements of the set `NAT` (see [2]):

```
reserve i,j,k for Element of NAT;

definition
func RAT+ -> set equals
  ({[i,j]: i,j are_coprime & j <> {}}
                        \ the set of all [k,1])
  \/ NAT;
end;
```

Note that the embedding $\mathbb{N} \subseteq \mathbb{Q}^+$ is performed by hand substituting all pairs `[k,1]` for `k` $\in \mathbb{N}$. Then in a second step the rational numbers $\mathbb{Q}$ are defined as

```
definition
func RAT -> set equals
  RAT+ \/ [:{0},RAT+:] \ {[0,0]};
end;
```

that is a negative rational number r is represented as a pair $[0, r']$, where r' is a non-negative rational number.

Now assuming that there is a polynomial p and a positive rational number r with $p = r$ we get $[i, j] = \{[n, p.n] \mid n \in \mathbb{N}\}$ for some $i, j \in \mathbb{N}$, $[i, j] \in \mathbb{Q}^+$ and hence that $[i, p.i] \in [i, j] = \{\{i\}, \{i, j\}\}$. Then both cases - $[i, p.i] = \{i\}$ and $[i, p.i] = \{i, j\}$ - lead to a contradiction. Here we just mention that in one (sub) case we even use that i and j are coprime.

With $\mathbb{Q}^+ \cap \mathbb{Q}^+[X] = \emptyset$ it is then straightforward to also show $\mathbb{Q} \cap \mathbb{Q}[X] = \emptyset$: For if $p = r$ for a polynomial p and a rational number r, then r must be negative, that is $r = [0, r']$ with $r' \in \mathbb{Q}^+$. But then because $[1, p.1] \in p = r = [0, r'] = \{\{0\}, \{0, r'\}\}$ we either get $[1, p.1] = \{0\} = 1$ or $[1, p.1] = \{0, r'\} = \{\emptyset, r'\}$ - in both cases a contradiction.

Summarizing, to show that our Mizar formalization of Kronecker's construction applies to $\mathbb{Z}_p, \mathbb{Q}$, and $\mathbb{R}$ we have to provide quite involved proofs using the set-based definition of numbers in Mizar.

VII. Conclusion

We have presented a Mizar formalization of Kronecker's construction. The main drawback is the necessary disjointness condition $F \cap F[X] = \emptyset$. This condition forbids the application of the construction with an arbitrary representation of the given field F.

Though the proofs have been carried out in Mizar, we claim that similar constructions should be possible in other proof assistants such as HOL [14] or Isabelle [15] as well: From a technical point of view all that was necessary to construct our counterexamples was the possibility to aggregate arbitrary objects in a set. In this way we defined the carrier of the fields by uniting elements of F with elements of $F[X]$. This should be possible in most proof assistants - if not some rank or typing argument forbids aggregating elements of different complexities.

Mathematicians solve our disjointness problem in a somewhat intuitive way:

[2] We already mention here, that Mizar uses Kuratowski's definition of pairs, that is $[x, y] := \{\{x\}, \{x, y\}\}$. This will be used in section VI.

[3] In fact, $j = \{0\}$ is possible if $n = p.n$; but in this case we get $j = \{\{n\}\}$, and hence $\{n\} = 0 = \emptyset$.

Of course $F \cap F[X]$, and also $F \cap F[X]/<p>$, can be considered non-empty, for if not just rename elements appropriately.

is their comment, and in fact for every field F there exists another representation $F' \cong F$ such that $F' \cap F'[X] = \emptyset$. It would be desirable to eliminate the disjointness condition in such a way. The comment can be stated as a theorem with F' denoting the renamed version of F:

```
theorem
for F being Field ex F' being Field
st F',F are_isomorphic &
   F' /\ (Polynom-Ring F') = {};
```

or more general for arbitrary fields

```
theorem
for F,E being Field ex F' being Field
st F',F are_isomorphic & F' /\ E = {};
```

These theorems would allow for formalizing Kronecker's construction without any condition. To prove them, it would be necessary to exchange a possibly infinite number of elements with new ones. The emphasize here is on "new", because one has to ensure that the adjoined elements are in fact new, that is appear neither in F nor in $F[X]$ (nor in E). Note also that the construction of our counterexamples uses precisely the technique of exchanging elements. So the key of the proof is the assumption that there is always an infinite stock of new objects which can be stated as a

```
theorem
for Y being set
ex Z being infinite set st Y /\ Z = {};
```

With such a theorem one could construct the required isomorphic copy F' by taking the elements of $F \cup F[X]$ (or $F \cup E$) as Y and then exchanging the elements of F that are in $F \cap F[X]$ (or in $F \cap E$) with elements from Z. Note, however, that one has to keep track of exactly which element of F is replaced with which element of Z. This is necessary to define the operations in F' appropriately.

We believe that the above theorem follows from Zermelo's axioms of set theory, namely the axiom of power set. Carrying out these proofs would call for a non-trivial amount of additional work. Nevertheless it might be helpful when further developing abstract field theory - and in fact would give a formalization of Kronecker's construction as found in the literature. Again the proofs would make use of basics of set theory showing that field theory heavily relies on the (informally used) foundations of mathematics. Therefore the further development of abstract field theory will remain a challenge.

REFERENCES

[1] G. Bancerek, *Tarski's Classes and Ranks*. Formalized Mathematics 1(3), 563–567, 1990.

[2] G. Bancerek, *Arithmetic of Non Negative Rational Numbers*. Mizar Mathematical Library, 1998.

[3] G. Bancerek et.al., *Mizar: State-of-the-art and Beyond*. in: M. Kerber et.al. (eds.), Proceedings of the 2015 International Conference on Intelligent Computer Mathematics, Lecture Notes in Computer Science 9150, 261–279, 2015. http://dx.doi.org/10.1007/978-3-319-20615-8_17

[4] *The Coq Proof Assistant*. available at www.coq.inria.fr.

[5] H.B. Enderton, *Elements of Set Theory*. Elsevier, 1977.

[6] Y. Futa, H. Okazaki, and Y. Shidama, *Set of Points on Elliptic Curve in Projective Coordinates*. Formalized Mathematics 19(3), 131–138, 2011. http://dx.doi.org/10.2478/v10037-011-0021-6

[7] A. Grabowski, A. Korniłowicz, and A. Naumowicz, *Mizar in a Nutshell*. Journal of Formalized Reasoning 3(2), 153–245, 2010. https://doi.org/10.6092/issn.1972-5787/1980

[8] A. Grabowski, A. Korniłowicz, and C. Schwarzweller, *Equality in Computer Proof-Assistants*. in: Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds.), ACSIS, Vol. 5, 45–54, 2015. http://dx.doi.org/10.15439/2015F229

[9] A. Grabowski, A. Korniłowicz, and C. Schwarzweller, *On Algebraic Hierarchies in Mathematical Repository of Mizar*. in: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, 363–371, 2016. http://dx.doi.org/10.15439/2016F520

[10] A. Grabowski and C. Schwarzweller, *Translating Mathematical Vernacular into Knowledge Repositories*. in: M. Kohlhase (ed.), Proceedings of the 4th International Conference on Mathematical Knowledge Management, Lecture Notes in Artificial Intelligence, 3863, 49–64, Springer Verlag, 2006.

[11] A. Grabowski and C. Schwarzweller, *On Duplication in Mathematical Repositories*. in: S. Autexier et.al. (eds.), Intelligent Computer Mathematics, Lecture Notes in Artificial Intelligence, 6167, 300-314, Springer Verlag, 2010.

[12] A. Grabowski, A. Korniłowicz, and A. Naumowicz, *Four Decades of Mizar*. Journal of Automated Reasoning, vol 55(3), 191–198, 2015. http://dx.doi.org/10.1007/s10817-015-9345-1

[13] J. Harrison, The HOL Light Theorem Prover. available at www.cl.cam.ac.uk/~jrh13/hol-light.

[14] *The HOL Interactive Theorem Prover*. available at hol-theorem-prover.org.

[15] *Isabelle*. available at isabelle.in.tum.de.

[16] A. Korniłowicz, *Quotient Rings*. Formalized Mathematics 13(4), 573–576, 2005.

[17] *Mizar Home Page*. available at www.mizar.org.

[18] A. Naumowicz and A. Korniłowicz, *A Brief Overview of Mizar*. in: Theorem Proving in Higher Order Logics 2009, S. Berghofer, T. Nipkow, C. Urban, M. Wenzel (eds.), Lecture Notes in Computer Science, 5674, 67–72, *Springer Verlag*, 2009.

[19] P. Rudnicki, A. Trybulec, and C. Schwarzweller, *Commutative Algebra in the Mizar System*. Journal of Symbolic Computation, vol. 32(1/2), pp. 143–169, 2001. http://dx.doi.org/10.1006/jsco.2001.0456

[20] C. Schwarzweller, A. Korniłowicz, and A. Rowińska-Schwarzweller, *Some Algebraic Properties of Polynomial Rings*. Formalized Mathematics 24(3), 227–237, 2016. http://doi.org/10.1515/forma-2016-0019

[21] W.A. Trybulec, *Vectors in Real Linear Space*. Formalized Mathematics 1(2), 291–296, 1990.

[22] A. Trybulec, A. Korniłowicz, A. Naumowicz, and K. Kuperberg, *Formal Mathematics for Mathematicians*. Journal of Automated Reasoning 50(2), 119–121, 2013. http://dx.doi.org/10.1007/s10817-012-9268-z

[23] B.L. van der Waerden, *Algebra Vol. I*. 8th edition Springer Verlag 1990.

[24] S. Weintraub, *Galois Theory*. 2nd edition Springer Verlag, 2008.

[25] F. Wiedijk, *Formalizing 100 Theorems*. available at www.cs.ru.nl/~freek.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 73–80
DOI: 10.15439/2018F202
ISSN 2300-5963 ACSIS, Vol. 15

A Non-Deterministic Strategy for Searching Optimal Number of Trees Hyperparameter in Random Forest

Kennedy Senagi
Department of Information Technology
Dedan Kimathi University of Technology
Kenya
Email: kennedy.senagi@dkut.ac.ke

Nicolas Jouandeau
LIASD
University Paris8
France
Email: n@ai.univ-paris8.fr

***Abstract*—In this paper, we present a non-deterministic strategy for searching for optimal number of trees hyperparameter in Random Forest (RF). Hyperparameter tuning in Machine Learning (ML) algorithms is essential. It optimizes predictability of an ML algorithm and/or improves computer resources utilization. However, hyperparameter tuning is a complex optimization task and time consuming. We set up experiments with the goal of maximizing predictability, minimizing number of trees and minimizing time of execution. Compared to the deterministic search algorithm, the non-deterministic search algorithm recorded an average percentage accuracy of approximately 98%, number of trees percentage average improvement of 44.64%, average time of execution mean improvement ratio of 175.62 and an average improvement of 94% iterations. Moreover, evaluations using Jackknife Estimation show stable and reliable results from several experiment runs of the non-deterministic strategy. The non-deterministic approach in searching hyperparameter shows a significant accuracy and better computer resources (i.e cpu and memory time) utilization. This approach can be adopted widely in hyperparameter tuning, and in conserving utilization of computer resources like green computing.**

I. Introduction

ML performance tuning is aimed at improving the predictability of ML algorithms. Improving performance of a ML systems can be done by configuring a set of hyperparameters. Most ML algorithms have several hyperparameters to be configured. Hyperparameters specify the interoperability of the underlying model. ML algorithms hyperparameter tuning is aimed at getting optimal values that can improve the algorithm's predictability considering minimum consumption of computer system resources [6]. When adopting ML algorithm to a specific dataset, hyperparameter tuning can be cumbersome and time consuming [13].

Manual, grid search and bayesian optimization are methods of hyperparameter optimization. Grid search is deterministic. It does an exhaustive search. It uses a predefined parameter space $S = \{0, 1, 2, ..., n\}$. The goal is to search an optimal hyperparameter s in S that records an optimal accuracy. Grid search consumes substantial amount time and is computationally expensive. However, it gives accurate results [4]. Manual search involves randomly selecting a value s in S. The value s is configured in the algorithm, the experiment executed and the accuracy observed. The process is repeated comparing the accuracy. The hyperparameter that records the optimal accuracy is selected. Manual search is cumbersome and difficult to reproduce results [1]. Bayesian optimization stochastically and efficiently trades off exploration and exploitation of the parameter space. It also explores historical information to find the parameters that maximize functions to inform user the configurations that best optimize predictability of the ML algorithm [5].

This paper introduces a non-deterministic search algorithm. The algorithm randomly selects 10% of elements in a parameter space. It then uses heuristics and termination conditions to maximize accuracy (acc) and minimize time of execution (t). This algorithm was applied and tested in selecting optimal number of trees (θ) in random forest (RF). In this paper, Section II covers related works, Section III discusses methodology and Section IV concludes this paper.

II. Related Works

In the paper by Hazan et al. (2017), large scale machine learning systems at times involves large number of parameters that are fixed manually. This is time consuming and at times inaccurate and difficult for a human expert. A hyper-parameter optimization strategy is proposed inspired by analysis of boolean function focusing on high-dimension datasets. The algorithm is an iterative application of compressed sensing techniques for orthogonal polynomials. The algorithm is tested in deep neural networks. In terms of running time, the algorithm records at least an order of magnitude faster than Hyperband and Bayesian Optimization and outperform Random Search 8x [Hazan et al., 2017]. Hazan et al. (2017) guides this work as they develops an algorithm and tests it in another algorithm; their algorithm establishes heuristics for reducing the search space.

Experiments showed that accuracy increased when number of trees in RF was doubled. However, there was a threshold beyond which there was no significance gain in accuracy. Therefore, increasing number of trees does not always mean

a better performance can be attained [15]. We note that, there was no significant variable that used to measure use of computing resources consumed when varying number of trees.

MapReduce was used to optimize regularization parameters for boosted trees and random forests (RF). For RF[2], two parameters were tuned: the number of trees in the model and the number of features selected to split each node. Experiments showed that performance was sensitive to the number of trees but less sensitive to the number of features in each split. Results showed that MapReduce could make parameter optimization feasible on a massive scale. However, it created possibilities for overfitting that could reduce accuracy and lead to inferior learning parameters [6].

In the technical report by [3], they discuss manually setting up, using and understanding RF. They note that RF grows trees rapidly and setting up a large number of trees (e.g. 1000) is okay. They further note that, if there are many variables, they can grow more trees (of up-to 5000) Beiman, (2003). From this work we can set up experiments with variable number of trees and see their effects on computing resources.

ML algorithms often involve careful tuning of learning parameters and model hyper-parameters. Parameter tuning is often a "black art" that requires expert experience, rules of thumb or sometimes brute-force search. To solve this problem, the following techniques were used: a full Bayesian treatment expected improvement, and algorithms (e.g ANN) for dealing with variable time regimes and running experiments in parallel. Results of this experiment surpassed a human expert at selecting hyper-parameters on the competitive CIFAR-10 dataset; beating the state of the art by over 3%. SVM was used as a case study algorithm [13].

A novel idea for approximate tree learning is seen in sparsity-aware algorithm for sparse data and weighted quantile sketch. The algorithm (XGBoost) proposes candidate splitting points according to percentiles of feature distribution, then maps the continuous features into buckets split, aggregates the statistics and finds the best solution among proposals based on the aggregated statistics. The algorithm also provides an insights on cache access patterns, data compression and sharing to build a scalable tree boosting system. The algorithm has been widely used and recognized in machine learning and data mining challenges e.g. Kaggle and KDDCup 2015. The algorithm can be applied to machine learning systems and in solving real-world scale problems using a minimal amount of resources [4].

Optimizing parameters of an evolutionary algorithm values is a challenging activity. CMA-ES tuning algorithms gave better results in terms of utility, in evolution algorithms. It is noted that using algorithms for tuning parameters of evolutionary algorithms does pay off in terms of performance. However, tuning algorithms gave better tuning parameter values than relying on intuitions and the usual parameter setting conventions [14].

It is challenging to create a large dataset and improve train ability of deep neural network models (DNNs). A selection of supplemental training datasets was used in fine-tuning a high-performing neural network model. Natural Language Processing system ability is improved after being evaluated by the Item Response Theory ability scores without negatively affecting generalization due to overfitting [9].

Large scale machine learning systems at times involve large number of parameters that are fixed manually. This is time consuming and at times inaccurate and difficult for a human expert. A hyper-parameter optimization strategy is proposed inspired by analysis of boolean function focusing on high-dimension datasets. The algorithm is an iterative application of compressed sensing techniques for orthogonal polynomials. The algorithm is tested in deep neural networks. In terms of running time, the algorithm records at least an order of magnitude faster than Hyperband and Bayesian Optimization and outperform Random Search 8x. The algorithm requires only uniform sampling of the hyperparameters and is easily parallelizable [7].

In the department of Soil Survey in Kenya Agriculture and Livestock Research Organization (KALRO) [10] and other soil research organizations, land evaluation is done manually, is stressful, takes a long time and is prone to human errors [11][12]. Parallel RF experiment prototypes are set up in [11] and further experiments in [12]. Parallel RF, Linear Regression, Linear Discriminant Analysis, KNN, Gaussian Naive Bayesian and Support Vector Machine are applied in predicting land suitability for crop (sorghum) production, given soil properties information. Parallel RF had a better accuracy of 0.96 and time of execution of 1.7 sec [12].

Besides assertions regarding performance reliability of default parameters in RF, many RF experiments fit using these values. An examination of parameter sensitivity of RF in computational genomic was studied. Experiments were evaluated using Area Under Curve (AUC), Root Mean Square Error (RMSE) and cross-fold validation. It was seen that RF performance was strongly affected by number of trees, sample size and number of random variables used at each split. It was noted that tuned RF gave better results than when default parameters/values are used. Effects of parameterization were analyzed using selection methods and showed that tuning can successfully improved prediction accuracy of non-parametric ML algorithms [8].

III. Methodology

In this research, we considered 14 standardized datasets collected from UCI Machine Learning website, namely: Balance Scale (1), Breast Cancer Wisconsin - Original (2), Car Evaluation (3), Habermans Survival (4), Pen-Based Recognition of Handwritten Digits (5), Website Phishing (6), Yeast (7), Banknote Authentication (8), Contraceptive Method Choice (9), Diabetic Retinopathy Debrecen (10), EEG Eye State (11), Pima Indians Diabetes (12), Wine Quality - White (13) and Wine Quality (14). In each dataset, we used simple random sampling without replacement strategy to sample 10% of elements in the search space. All experiments were run 10 times and results averaged. Number of trees (θ) was varied accordingly as we measured accuracy (acc) and time of

execution (t). The computer had the following specifications: Intel(R) Xeon(R) CPU W3505 @ 2.53GHz x 2.

A. Considering 2 to 4096 Number of Trees

We considered a finite set of sorted number of trees in the parameter space. RF predictability was evaluated by acc defined in equation 1 with n samples, where $\hat{y}_i$ is the predicted label and y_i is the original label. The results of acc and t are tabulated in Tables I and II respectively.

$$acc(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \tag{1}$$

Table I shows a general trend of accuracy increasing steadily with increase in number of trees, then flattens. RF classification employs bagging principles, where a committee of trees each, cast a vote for the predicted class. However, RF classifier introduces modifications in bagging where it builds a large collection of de-correlated trees, and then averages them. When the number of trees become huge, we see RF accuracy varying insignificantly meaning the average accuracy of de-correlated trees varying insignificantly. Average accuracy varies because of the random nature of RF, for example, randomly selecting features when building trees. We further observed an interesting trend in the number of trees against accuracy; increasing the number of trees does not significantly contribute to a positive accuracy. The maximum accuracy values are in bold, in Table I. Moreover, we see 13 out of the 14 dataset's maximum accuracy values found between 2 and 512 trees. Dataset 6 with 2048 number of trees recorded an accuracy of 88.7% and 6.42 seconds. It's second best accuracy is 88.4% with 0.89 seconds observed at 256 number of trees. In this case, we think 256 number of trees is better because the change in accuracy rather insignificant (-0.3) while it runs faster (approximately 7x faster). Generally, we observed better results between 2 and 512, and we assume these results can be extended to other datasets. We call the region between 2 and 512, the $fertile\ region$.

Table II shows a general trend of time of execution increasing steadily with increase in number of trees. This tells us that more number of trees demand more computing resources. We also observed a relative significant change in time of execution, the threshold values are in bold. Generally, after 64 number of trees, we see a significant change in time difference. Increase in number of trees increases time of execution. More number of trees requires more computer resources to build and average the de-correlated trees in RF.

Different datasets give different values of accuracy and time of execution with the same number of trees. The selected datasets have different complexity i.e dimensionality, number of records and classes. This leads to a variation in accuracy and time of execution. For us to have an optimal number of trees hyperparameter in RF classifier, it is important we consider maximizing accuracy and minimizing number of trees.

However, we see the 6$^{\text{th}}$ dataset maximum accuracy of 88.7% and time of execution of 6.42 seconds being out of the fertile region i.e 2048 number of trees. As per our experiments, this is a probability of 0.07 i.e 1 out of 14 datasets can exhibit this. The second best accuracy of 87.9% is observed in the fertile region i.e 128 number of trees with 0.5 seconds time of execution. In such instances, we can compromise accuracy to get a better time of execution, for this case, we compromise 0.8% accuracy to gain 5.92 seconds.

B. Considering 2 to 512 Number of Trees

In the fertile region, we observed lower time of execution and maximum accuracy, therefore, we will have avoided searching out regions (> 512) that show higher time of execution and significantly same or lower accuracy. We defined a finite set of sorted number of trees from the parameter space θ. We configured, trained and tested RF with the respective θ and recorded acc and t. The results are show in Fig. 1 and 2. Fig. 1 is a box plot of accuracy for number of trees against datasets across 14 datasets in the fertile region. Most datasets had a low inter-quartile range, low difference between the low and maximum points and more outliers below the lower whiskers. Some box plots also recorded some outliers above the upper whisker. A low difference in quartile ranges means there was a low variation in accuracy from the median and 50% of the accuracy records are within this region. However, the outliers inform us that, some maximum accuracy values were very far away from the median and some lowest accuracy values were very far away from the median. The goal of any data scientist is to have the maximum accuracy when configuring RF with a specific number of trees. Nonetheless, we see variations in accuracy on different datasets, i.e. different datasets record different accuracy levels. This make the search problem more difficult because we need to have a strategy that will be dynamic to search the best accuracy in different datasets. This research was interesting in finding number of trees (i.e. the outliers in the upper whisker) that maximize accuracy.

Fig. 2 is a box plot of time of execution of number of trees against datasets across 14 datasets in the fertile region. We see the lower whisker having almost the same time of execution. This means there are some number of trees that could give almost the same minimum time of execution when configured in RF. We also see the lower whiskers being shorter than the upper whiskers. A shorter lower whisker means most lower time of executions were closer to the median. This research was interesting in these number of trees that minimize time of execution.

From these analysis, we formulated deterministic, non-deterministic and automatic configuration (having 8 number of trees by default) algorithmic approaches in searching optimal number of trees hyperparameter in the fertile region.

C. Deterministic Hyperparameter Search

Deterministic search algorithm is defined in equation 2. We developed a deterministic hyperparameter search algorithm from equation 2 as outlined in Algorithm 1. We considered number of trees θ, time t and accuracy acc descriptions and results from Section III-B. The deterministic hyperparameter search algorithm's goal is to maximize acc and minimize θ.

Table I: Accuracy (percentage) of RF with θ trees for 14 datasets (DS)

DS	Number of Trees											
	2	4	8	16	32	64	128	256	512	1024	2048	4096
1	80.3	81.9	83.0	82.4	84.6	**85.6**	84.6	84.0	84.0	84.0	84.6	84.6
2	91.7	93.7	97.1	**98.0**	97.6	97.6	97.6	97.1	97.1	97.1	97.1	97.1
3	86.3	85.5	83.6	83.8	**84.8**	84.4	84.6	84.4	84.8	84.8	84.6	84.6
4	76.1	79.3	75.0	76.1	**79.3**	79.3	78.3	78.3	78.3	79.3	78.3	79.3
5	92.5	96.8	98.3	98.6	98.4	98.9	99.0	**99.1**	99.0	99.1	99.1	99.1
6	81.5	86.9	86.2	87.4	85.7	87.4	87.9	88.4	87.7	87.9	**88.7**	88.2
7	48.6	47.8	52.9	57.3	56.5	59.5	**59.8**	58.8	58.8	58.5	58.5	58.8
8	96.6	**97.8**	97.6	97.6	97.3	97.6	97.8	97.8	98.1	97.8	97.8	97.8
9	46.4	48.4	49.1	**51.6**	49.5	49.8	51.1	49.5	50.7	51.4	50.9	51.1
10	61.3	64.7	65.3	65.0	**69.9**	66.5	67.6	67.9	68.2	67.1	67.9	67.3
11	77.9	84.2	87.9	89.3	91.3	**92.7**	92.0	92.2	92.2	92.1	92.3	92.2
12	66.7	71.0	74.9	74.5	76.6	76.6	76.6	75.8	**77.5**	76.6	77.1	77.1
13	54.9	59.4	64.7	64.6	65.7	65.9	67.1	**67.3**	67.1	66.6	67.3	67.4
14	54.4	69.7	63.3	67.3	69.2	69.2	69.6	**70.2**	69.8	69.2	69.8	69.8

Table II: Time of execution (sec) of RF with θ trees for 14 datasets (DS)

DS	Number of Trees											
	2	4	8	16	32	64	128	256	512	1024	2048	4096
1	0.21	0.21	0.22	0.23	0.25	**0.30**	0.51	0.90	1.60	3.29	6.49	12.45
2	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	0.90	1.59	3.09	5.98	12.35
3	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	1.00	1.80	3.39	6.79	13.57
4	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	0.80	1.60	3.30	5.99	12.06
5	0.21	0.21	0.22	0.23	**0.26**	0.41	0.60	1.10	2.20	4.01	8.23	15.87
6	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	0.89	1.89	3.46	6.42	13.14
7	0.21	0.21	0.22	0.23	0.26	**0.30**	0.50	1.00	1.88	3.71	7.13	14.06
8	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	0.90	1.69	3.17	6.64	12.76
9	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	1.00	1.89	3.47	6.95	13.70
10	0.21	0.21	0.22	0.23	0.25	**0.30**	0.50	0.90	1.79	3.47	6.93	14.41
11	0.21	0.21	0.22	0.33	**0.46**	0.71	1.20	2.40	4.70	9.18	18.27	36.62
12	0.21	0.21	0.22	0.24	0.25	**0.30**	0.50	0.79	1.68	3.46	6.43	12.64
13	0.21	0.21	0.22	0.23	**0.25**	0.51	0.70	1.40	2.69	5.28	10.45	21.20
14	0.21	0.21	0.22	0.23	0.25	0.40	**0.60**	1.10	2.09	3.76	7.33	14.96

We note that, $\exists acc_{max} \in acc$ that has θ_{best}. The deterministic search algorithm is exhaustive, i.e., it does a linear search and returns acc_{max}, with θ_{best} and the time needed t. Experiment results are tabulated in Tables III, IV and V.

$$\theta^*_{best}, acc^*_{best} = \underset{\theta \in \mathcal{T}}{argmax}\ \hat{Q}\ (\theta, acc) \tag{2}$$

Algorithm 1 The Deterministic Hyperparameter Search

```
procedure DETERMINISTICSEARCH(train, test)
    t_i ← CURRENTTIME()
    T ← [θ_1, θ_2, θ_3, ..., θ_n]
    acc_max ← 0
    for each θ in T do
        rf ← RANDOMFOREST(θ, train)
        acc_new ← ACCURACY(rf, test)
        if acc_new > acc_max then
            (acc_max, θ_best) ← (acc_new, θ)
    time_spent ← CURRENTTIME() − t_i
    return (acc_max, θ_best, time_spent)
```

D. The Non-Deterministic Hyperparameter Search Algorithm

In this research, we were interested in maximizing accuracy and minimizing number of trees. Tables 1 and 2 shows almost the same accuracy but with different time of execution. Table 2 shows more NoTs require more ToE (i.e. memory and cpu time). With this analogy, this research formulated a non-deterministic search approach to converge close/to maximize accuracy and minimize number of trees and save time of execution. The algorith is outlined Algorithm 2, where $\theta_i = random(\in \mathcal{T}), \psi_1 = 1 + \frac{lim}{100},\ and\ \psi_2 = 1 - \frac{lim}{100}$.

We considered θ, acc and t descriptions and results from Section III-B. The goal of this algorithm was to maximize acc and minimize t through randomization. In this algorithm we assumption that, $\exists acc_{best} \in acc$ that has θ_{best}. Note that the function GENERATE( ) returns 26 elements which is approximately 10% of elements in the parameter space. We iterate through the random selected number of trees as we configure RF. We considered percentage upper bound and lower bound of the acc_{best}. If acc_{rand} falls in the upper boundary, then $acc_{best} \leftarrow acc_{rand}, \theta_{best} \leftarrow \theta_{rand}$ and we $break$, with the assumption that we do not anticipate further percentage Δ acc_{best}. If acc_{rand} falls in the lower boundary and θ_{rand} is less than θ_{best}, then $acc_{best} \leftarrow acc_{rand}, \theta_{best} \leftarrow \theta_{rand}$ and we also $break$, with the assumption that we have an insignificant Δ acc_{best} and we have a better t_{best}. Moreover, if acc_{rand} falls above the upper boundary, then $acc_{best} \leftarrow acc_{rand}, \theta_{best} \leftarrow \theta_{rand}$, and we continue looping with the assumption that

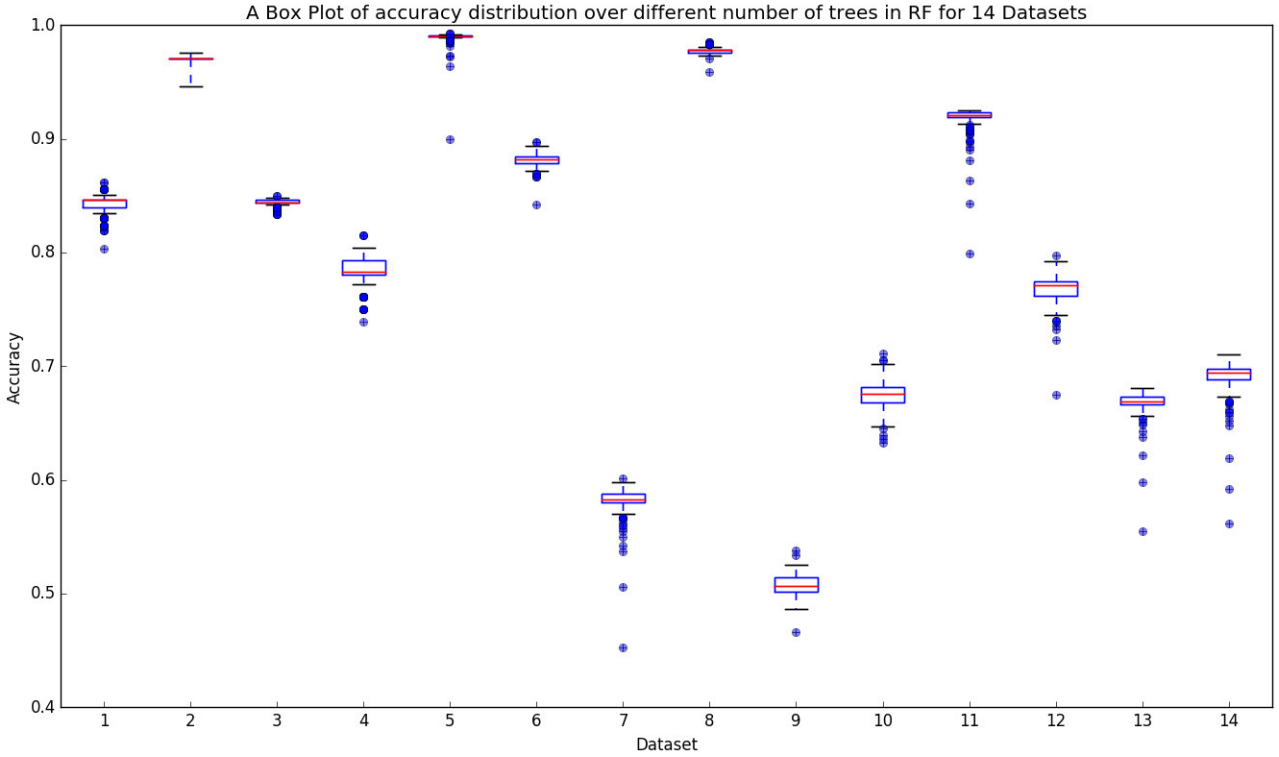


Fig. 1. Number of trees (many) against datasets of Accuracy in RF for 14 Datasets

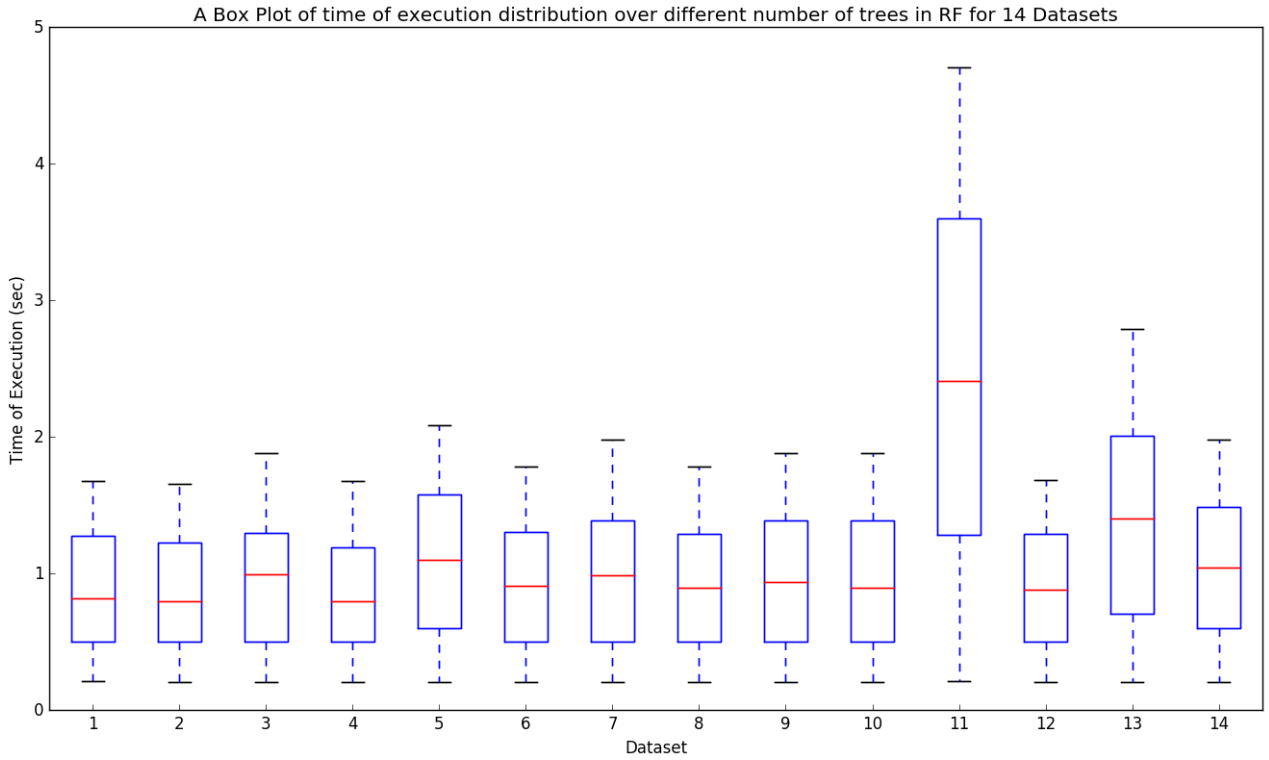


Fig. 2. Number of trees (many) against datasets of Time of Execution in RF for 14 Datasets

we anticipate further percentage $\Delta\ acc_{best}$. Lastly, we $break$ when iteration counts are 10% of the parameter space, with the assumption that we have uniformly sampled the whole parameter space. We set the percentage boundary as 1% to increase the algorithm's accuracy. Experiment results are tabulated in Tables III, IV and V.

Algorithm 2 The Non Deterministic Hyperparameter Search

```
vals = []
procedure GENERATE()
    while LEN(vals) ≤ 26 do
        val = 2 + rand()%512
        if val is not in vals then
            add val in vals
    return val
procedure NONDETERMINISTICSEARCH(train, test)
    t_i ← CURRENTTIME()
    acc_rand, θ_rand, acc_best, θ_best, count ← 0
    T ← GENERATE( )
    for each θ_rand in T do
        rf ← RANDOMFOREST(θ_rand, train)
        acc_rand ← ACCURACY(rf, test)
        if count == 0 then
            (acc_best, θ_best) ← (acc_rand, θ_rand)
        if ψ_1.acc_best > acc_rand > ψ_2.acc_best then
            if acc_rand < acc_best then
                if θ_rand < θ_best then
                    (acc_best, θ_best) ← (acc_rand, θ_rand)
                    break
            else
                (acc_best, θ_best) ← (acc_rand, θ_rand)
                break
        else if acc_rand > ψ_1.acc_best then
            (acc_best, θ_best, count) ← (acc_rand, θ_rand, 0)
        count ← count + 1
        if count >= 10 then break
    time_spent ← CURRENTTIME() − t_i
    return (acc_best, θ_best, time_spent)
```

E. Determinstic and Non-Deterministic Hyperparameter Search Algorithms, and Auto-Configured RF

Table III contains results and analysis of minimum number of trees selected by deterministic and non-deterministic hyperparameter search algorithms. We see a considerably good percentage improvement of number of trees in the non-deterministic search algorithm. At some instances, for example, in datasets 8 and 13, the non-deterministic search algorithm was able to perfectly converged to the minimum number of trees with 26 and 2 iterations respectively. In some datasets e.g dataset 1, the percentage number of trees improvement was poor. Moreover, as observed in Table III, 50% of the datasets used less than 50% (i.e. less than 5% of random values in the search space) of random values while iterating, to converge close/to maximum accuracy and minimum number of trees. With this observation, in some cases, we can have an assumption that sometimes increasing the search space would not have much scientific significance. Generally, the percentage number of trees improvement was 44.6% and the average number of iterations used were 14.5.

Table IV has results and analysis of accuracy recorded from running deterministic, non-deterministic and auto-configured RF algorithms. The auto-configured RF had a mean percentage difference -5.46 while the non-deterministic search algorithm had a considerably better percentage change of -2.1. In non-deterministic search algorithm, datasets 2, 8 and 13 recorded a zero percentage change in accuracy. 50% of the datasets recorded a percentage change of more than 1%.

Table V has results and analysis of time of execution of deterministic and non-deterministic search algorithms, and auto-configured RF. The ratio of deterministic:non-deterministic algorithms and deterministic: auto-configured RF are calculated. Their averages are also calculated. Both auto-configured RF and non-deterministic algorithm record a very high average ratio of 5623 and 176 respectively.

As discussed in Section III-C, the deterministic search algorithm is exhaustive and selects the minimum number of trees that has the maximum accuracy. With these results, we benchmark the non-deterministic search algorithm and auto-configured RF. The non-deterministic search algorithm, as discussed in Section III-D, uses the principle of randomization, heuristics and terminating policies as outlined in Algorithm 2. With this strategy, the non-deterministic search algorithm recorded $\approx 98\%$ average accuracy, and could run at an average of 175.62 faster, on an average of 14.5 iterations. Using the strategy formulated in Algorithm 2, the non-deterministic search algorithm recorded 100% accuracy at three instances and recorded zero number of trees percentage improvement on two instances. Moreover, in the non-deterministic search algorithm, we recorded number of trees that are below the number of trees threshold (64 trees), that showed a significant change in time of execution, as discussed in Section III-A. This means the formulated strategy worked quite well. Considering dataset 2, we note that 0% percentage accuracy change, was got with more number of trees (48 trees instead of 46 trees) but at 34.8 times faster. These shows 100% accuracies got, at more number trees but takes a shorter searching time. This makes the strategy formulated in this research more relevant. Despite the 1% boundary policy and breaking policies strategies, 50% of the datasets recorded less than 1% change in percentage accuracy. The other 50% scored fairly good results too. Generally, a shorter time of execution means the process will take a shorter time in memory and shorter cpu time, when tuning RF. We see the non-deterministic search algorithm run $\approx$ 175 faster on average, achieving an average of $\approx 98\%$ accuracy, on an average of 5.6% iterations (i.e 14.5 of 256 iterations in the parameter space). This is an improvement in iterations by 94.4%. Therefore, the non-deterministic search algorithm can improve utilization of computing resources while maintaining a significant accuracy.

Auto-configuring (having 8 number of trees by default) RF

Table III: Recorded minimum number of trees (θ_{best}) and iterations for deterministic and non-deterministic search algorithms across 14 datasets (DS), and their mean (μ)

DS	Deterministic	Non-Deterministic		
	θ_{best}	θ_{best}	θ % improvement	Iteration
1	26	32	-23.08	5
2	46	48	-4.35	26
3	116	46	60.34	26
4	70	18	74.29	26
5	48	16	66.67	26
6	216	26	87.96	26
7	118	34	71.19	3
8	44	44	0.00	26
9	48	42	12.50	2
10	18	10	44.44	4
11	196	50	74.49	26
12	164	10	93.90	2
13	46	46	0.00	2
14	150	50	66.67	3
μ	93.28	33.71	44.64	14.5

Table IV: Maximum accuracy (acc_{best}) recorded across 14 datasets (DS), and their mean (μ)

DS	Deterministic	Auto-Configured		Non-Deterministic	
	acc_{max}	acc_{best}	% Δ	acc_{best}	% Δ
1	0.862	0.819	-4.99	0.856	-0.70
2	0.976	0.971	-0.51	0.976	0.00
3	0.850	0.846	-0.47	0.846	-0.47
4	0.815	0.761	-6.63	0.804	-1.35
5	0.993	0.973	-2.01	0.990	-0.30
6	0.897	0.855	-4.68	0.887	-1.11
7	0.601	0.552	-8.15	0.593	-1.33
8	0.985	0.976	-0.91	0.985	0.00
9	0.538	0.480	-10.78	0.505	-6.13
10	0.711	0.627	-11.81	0.682	-4.08
11	0.925	0.890	-3.78	0.919	-0.65
12	0.797	0.740	-7.15	0.736	-7.65
13	0.681	0.636	-6.61	0.681	0.00
14	0.710	0.654	-7.89	0.679	-4.37
μ	0.81	0.77	-5.46	0.80	-2.10

Table V: Time of execution (sec) recorded across 14 datasets (DS), and their mean (μ)

DS	Deterministic	Auto-Configured		Non-Deterministic	
	i (sec)	t (sec)	Ratio	t (sec)	Ratio
1	224.11	0.03	7470	1.22	183.7
2	217.97	0.02	10899	6.27	34.8
3	239.22	0.03	7974	6.45	37.1
4	216.26	0.02	10813	6.43	33.7
5	282.42	0.07	4035	6.38	44.3
6	235.94	0.03	7865	6.25	37.7
7	249.68	0.04	6242	0.78	319.7
8	230.44	0.03	7681	6.34	36.3
9	246.37	0.03	8212	0.51	484.0
10	246.37	0.04	6159	0.94	263.2
11	622.88	0.29	2148	10.20	61.1
12	227.91	0.03	7597	0.46	497.6
13	360.73	0.11	3279	0.59	613.5
14	260.52	0.05	5210	0.77	338.8
μ	227.11	0.05	5623	3.15	175.62

showed good results. It recorded $\approx 94.5\%$ average accuracy change and very good time of execution ratio of 5623; probably had fewer iterations.

Table VI: Jackknife Estimates for deterministic and non-deterministic search algorithms across 14 datasets (DS), and their mean (μ)

DS	Bias-Corrected Jackknifed Estimate		Confidence Interval			
			Deterministic		Non-Deterministic	
	Determ-inistic	Non-Determ-inistic	Lower	Upper	Lower	Upper
1	0.86	0.85	0.86	0.87	0.85	0.85
2	0.98	0.98	0.98	0.98	0.98	0.98
3	0.85	0.85	0.85	0.85	0.85	0.85
4	0.82	0.79	0.82	0.82	0.79	0.8
5	0.99	0.99	0.99	0.99	0.99	0.99
6	0.89	0.88	0.89	0.89	0.88	0.89
7	0.61	0.59	0.6	0.61	0.59	0.59
8	0.99	0.99	0.99	0.99	0.98	0.99
9	0.53	0.52	0.53	0.53	0.51	0.52
10	0.71	0.69	0.71	0.71	0.69	0.69
11	0.93	0.91	0.93	0.93	0.91	0.92
12	0.8	0.77	0.79	0.8	0.77	0.78
13	0.68	0.67	0.68	0.68	0.66	0.67
14	0.71	0.69	0.71	0.71	0.68	0.69
μ	0.81	0.80	0.81	0.81	0.80	0.80

F. Evaluation using Jackknife Estimation

Jackknife is used to evaluate the quality of the prediction of computational models. It uses resampling to calculate standard deviation error and estimate bias of a sample statistic, as shown in equations 3 and 4 [16]. We computed Jackknife across the 14 datasets and tabulated results as shown in Table VI. We recorded a zero for bias and standard errors across all datasets.

$$Var(\theta) = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{\theta}_i - \bar{\theta}_{jack})^2, \quad \bar{\theta}_{jack} = \frac{1}{n}\sum_{i=1}^{n}(\bar{\theta}_i) \tag{3}$$

$$\bar{\theta}_{BiasCorrected} = N\bar{\theta} - (N-1)\bar{\theta}_{jack} \tag{4}$$

In Table VI we see different datasets record different values of Bias-Corrected Jackknifed Estimates. We also observe stable results are per the predictions in Table IV. Standard error is used for null hypothesis testing and for computing confidence intervals (upper and lower bounds). This explains why we observe confidence intervals deviating insignificantly. We also see the bias-corrected Jackknifed estimate deviating minimally because the standard error were zero across all the records. These results show that the non-deterministic search algorithm predictions are stable and reliable.

IV. Conclusion

In this research, we formulated a non-deterministic strategy in searching for the best hyperparameter in random forest algorithm considering number of trees, accuracy and time of searching hyper-parameter. The non-deterministic search strategy recorded significantly good results in maximizing accuracy, minimizing number of trees and minimizing searching time. Evaluations using Jackknifed Estimation show that its predictions are stable. Moreover, the non-deterministic search strategy had a significant accuracy levels and better utilization cpu processing and time in memory. This research can be widely adopted in algorithms hyperparameter search and in green computing to preserve computing resources.

ACKNOWLEDGMENT

We would like to express our appreciation to the French Embassy in Kenya and the French Government's Ministry of Foreign affairs for the financial support in this research.

REFERENCES

[1] J. Bergstra and Y. Bengio. "Random search for hyper-parameter optimization." Journal of Machine Learning Research, pp. 281-305, 2012.

[2] L. Breiman. "Random forests." Machine learning, Kluwer Academic Publisher, 45(1), DOI:10.1023/A:1010933404324, pp. 5-32, 2001.

[3] Breiman, L., and Cutler, A. (2003), "Random forests manual v4.0", Technical report, UC Berkeley. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf Date Accessed: July 2018.

[4] T. Chen and C. Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, DOI:10.1145/2939672.2939785, pp. 785-794. ACM, 2016.

[5] I. Dewancker, M. McCourt, S. Clark, P. Hayes, A. Johnson and G. Ke. "A stratified analysis of bayesian optimization methods." Cornell University Library, arXiv:1603.09441 [cs.LG], 2016.

[6] Y. Ganjisaffar, T. Debeauvais, S. Javanmardi, R. Caruana and C.V. Lopes. "Distributed tuning of machine learning algorithms using MapReduce clusters." In Proceedings of the Third Workshop on Large Scale Data Mining: Theory and Applications, DOI:10.1145/2002945.2002947 USA, 2011.

[7] E. Hazan, A. Klivans and Y. Yuan. "Hyperparameter optimization: A spectral approach.", arXiv:1706.00764 [cs.LG], 2017.

[8] B.F. Huang and P.C. Boutros. "The parameter sensitivity of random forests." BMC bioinformatics, 17(1), DOI: https://doi.org/10.1186/s12859-016-1228-x, 2016.

[9] J.P. Lalor, H. Wu and H. Yu. "CIFT: Crowd-informed fine-tuning to improve machine learning ability". arXiv:1702.08563 [cs.CL], 2017

[10] Kenya Agricultural and Livestock Research Organization. www.kalro.org/. Date Accessed: July 2018.

[11] K. Senagi, N. Jouandeau and P. Kamoni. "Machine learning algorithms for soil analysis and crop production optimization: A review". In Proceedings of the International Conference on Mass Data Analysis of Images and Signals (MDA), USA, pp. 1-15, 2017.

[12] K. Senagi, N. Jouandeau and P. Kamoni. "Using parallel random forest classifier in predicting land suitability for crop production". Journal of Agricultural Informatics. Vol. 8 (3), 2017.

[13] J. Snoek, H. Larochelle and R.P. Adams. "Practical bayesian optimization of machine learning algorithms." In Advances in neural information processing systems, pp. 2951-2959, 2012.

[14] S.K. Smit and A.E. Eiben. "Comparing parameter tuning methods for evolutionary algorithms." In Evolutionary Computation CEC'09, IEEE, DOI:10.1109/CEC.2009.4982974, pp. 399-406, May 2009.

[15] T.P. Oshiro, S.J. Perez and A. Baranauskas. "How many trees in a random forest?" In Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer, Berlin, Heidelberg, DOI: https://doi.org/10.1007/978-3-642-31537-4_13, pp. 154-168, 2012.

[16] S. Wager, T. Hastie and B. Efron. "Confidence intervals for random forests: The Jackknife and the infinitesimal Jackknife." Journal of Machine Learning Research 15(1), 1625-1651, 2014.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 81–84
DOI: 10.15439/2018F295
ISSN 2300-5963 ACSIS, Vol. 15

Testing the Algorithm of Area Optimization by Binary Classification with Use of Three State 2D Cellular Automata in Layers

1st Miroslaw Szaban
Institute of Computer Science,
Siedlce University of Natural Sciences and Humanities,
Poland
mszaban@uph.edu.pl

2nd Anna Wawrzynczak
Institute of Computer Science,
Siedlce University of Natural Sciences and Humanities
and National Centre for Nuclear Research,
Poland
awawrzynczak@uph.edu.pl

***Abstract*—The paper is dedicated to a new algorithm of optimization in the sense of the area. Proposed method joins a few issues. First one is utilizing data from the set of sensors monitoring the area put into optimization. The second one is using the classification method based on two-dimensional three-state cellular automata, working on the data reported by the sensors. This method classifies all points of the area based on the data received from the sensors and designates optimal subarea. The third issue is applying the categorization layers to the data received from sensors. Such, approach gives a possibility to specify the areas in the different levels and, in consequence, after analysis, optimal subarea or subarea including the optimal point can be designated. This method can be used in different optimization tasks, starting from simple one as optimization of n-dimensional function, through specifying the contaminated area utilizing data from mobile sensors and finally estimating the contamination source-term. In this paper are presented results of testing for the proposed algorithm on a few selected functions from the set of dedicated for this purpose.**

***Index Terms*—area optimization, cellular automata, classification, sensors**

I. Introduction

In a classification problem, we wish to determine to which class new observations belong, based on the training set of data containing observations whose class is known. The binary classification deals with only two classes, whereas in a multiclass classification observations belong to one of the several classes. The well-known classifiers are neural networks, support vector machines, k-NN algorithm, decision trees, and others. The idea of using cellular automata (CA) in the classification problem was described by Maji et al. [2], Povalej et al. [3] and by Fawcett [1]. Fawcett designed the heuristic rule based on the von Neumann neighborhood (so-called voting rule); moreover, tested its performance on different sets of data. Results of Fawcett's study indicated his method, based on CA, as better than the other compared methods, like as (a) J48, a decision tree induction algorithm, (b) k-NN, a nearest-neighbor learning algorithm, (c) SMO, implementation of support vector machines. Recently, in the papers [4] were proposed and analyzed the Fawcett's method modifications into a probabilistic form of such method. These modifications were examined on the different sets of data, and obtained results show in general its higher effectiveness for classification (lower number of incorrect classifications), also in general better accuracy (the shortest scattering range).

The reconstruction of the source of an airborne contaminant may be obtained by using forward approaches, in which source characteristics are inferred from concentration or deposition measurements at different locations and time intervals by establishing source-concentration relationships. In e.g.[10] authors presented the reconstruction of the airborne contaminant source utilizing the Bayesian approach in conjunction with Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC). A comprehensive literature review of past works on solutions of the inverse problem for atmospheric contaminant releases can be found in (e.g.[9]). This class of problems is a potential area of application for the newly presented algorithm of area optimization and binary classification with use of three state 2D cellular automata.

This paper is organized as follows. Section 2 describes two-dimensional CAs, binary classification problem and binary classification methods based on 3-state CA. In Section 3 is presents the construction of the algorithm of area optimization. The stages of proposed approach examining and experimental results are presented in Section 4. The last Section concludes the paper, and is a study of application possibilities of the newly proposed algorithm and plans of future work.

II. Two-Dimensional Cellular Automata and Binary Classification Problem

CAs and they potential to efficiently perform complex computations are described by S. Wolfram in [8]. In this paper is considered two-dimensional CA. CA is a rectangular grid of $N \times M$ cells, each of which can take on k possible states. After determining initial states of all cells (i.e. the initial configuration of a CA), each cell changes its state according to a rule - transition function TF which depends on states of cells in a neighborhood around it. In this paper is considered finite CA (finite length of CA) with the periodic boundary conditions (bordered cells are neighbour cells each other).

Two types of the neighborhood are commonly used: the von Neumann neighborhood (the four cells orthogonally surrounding the central cell) which can be described as $a_{i,j}^{(t+1)} = TF[a_{i,j-1}^{(t)}, a_{i-1,j}^{(t)}, a_{i,j}^{(t)}, a_{i+1,j}^{(t)}, a_{i,j+1}^{(t)}]$, where $a_{i,j}^{(t)}$ denotes the state of a cell at position i, j in the two-dimensional cellular grid, at time step t. Also, the Moore neighborhood (the eight cells around the central cell) which can be described as $a_{i,j}^{(t+1)} = TF[a_{i-1,j-1}^{(t)}, a_{i,j-1}^{(t)}, a_{i+1,j-1}^{(t)}, a_{i-1,j}^{(t)}, a_{i,j}^{(t)}, a_{i+1,j}^{(t)}, a_{i-1,j+1}^{(t)}, a_{i,j+1}^{(t)}, a_{i+1,j+1}^{(t)}]$.

The square state of the data space in classification problem should be i.e. $[0, 1] \times [0, 1]$. Suppose that $N \times M$ data-points $p_{(i,j)}$=(x_i, y_j), where i=1, 2, ..., N and j=1, 2, ..., M are given as a training set from two classes: class 1 and class 2. When each of $p_{(i,j)}$ data-points is known as one of two classes then we have the classification. On the other hand, when even one of the data-points is not one of two known classes we have the classification problem. Moreover, to answer the question, to which of class (1 or 2) unclassified data points belong to, the classification method should be applied. In CA the data space of such problem should be mapped from $[0, 1] \times [0, 1]$ into the grid of $N \times M$ cells (in this paper $N \times N$ for the simplicity). Each cell can take one of 3 states, classified the state 1 (class 1) and state 2 (class 2) and also unclassified state (class 0). Classifier - the rule of CA will analyze the unclassified cells and changes its states into one of two known.

The classification methods based on two-dimensional three-state cellular automata was applied for classifying whole points of the area on the base data received from the sensors. The goal was expected designation of the optimal subarea. For this purpose, three kinds of the classifiers were studied. The first classification method was proposed in [1] (the rule of CA known as $n4\_V1\_stable$). The second and third one were modifications of Fawcett's rule into two patterns: partially and fully probabilistic (see, [4]). A proposed modification should strengthen an original and more accurately classify binary data, especially for large CA grid.

III. Algorithm of the Area Optimization Method

The proposed method uses the data reported by the set of sensors monitoring the area put into optimization. This data are the input for the classification method based on two-dimensional three-state cellular automata, which classifies all points of the area to designate optimal subarea. The layers are categorized based on the level of values received from sensors. The steps of the algorithm of the optimization method are presented below.

The algorithm of area optimization by layers and binary classification with use of three state 2D cellular automata:

1) Downloading input parameters:
   - CA size,
   - Threshold - the minimum value for which the recorded indication is acceptable (sensor is in positive state - class 1), the lover values recorded by the sensor are considered as 0 (sensor is in negative state - class 2),
   - Step of Threshold - the value by which the Threshold is increased during processing, it designates the levels of the layers,
   - Number of sensors,
   - Method of data classification (including type of neighborhood);
2) Preparation of cellular automaton,
   - Mapping optimizing area into (discrete) CA grid,
   - The random distribution of sensors in CA grid - CA cells with included sensors are the classified cells (class 1 or class 2), other CA cells are unclassified (class 0),
3) Searching for solution:
   - Preparation of layers for cellular automaton work (with use of Threshold and Step of Threshold): values of sensors in layer $i \in [Threshold + (i - 1) * Step; Threshold + i * Step)$, where $i = 1, 2, ...$,
   - For each layer, specify classes of CA cells with sensors: where the sensor value is $> Threshold + (i - 1) * Step$, where $i = 1, 2, ...$ Then the sensor is in a positive state (class 1). Otherwise, the sensor is in negative state - class 2,
   - For each layer perform classification - during processing of a cellular automaton are specifying classes 1 or 2 for CA cells being unclassified (class 0):
4) Elaboration of received results:
   - Designation of optimal area - development of common parts from classified as class 1 areas on each of analyzed layers,

IV. Experimental Results

The above presented and described algorithm for area optimization by layers, which apply in its work binary classification on three states two-dimensional CA should be examined in the sense of its efficiency for optimization. In this kind of test are usually used dedicated sets of different n-variables functions, as Test Functions for Unconstrained Global Optimization [7]. This set contains the testing functions with one or multi-global optima. For our problem were used functions having one optimum. Because functions have one global minimum, and proposed algorithm searching for maximum, as it was mentioned earlier, the functions $f(x)$ were inverted into $f^*(x) = f_{max}(x) - f(x)$, $f_{max}(x)$ is the maximal value of function in analyzed search domain. In the tests were used three selected and inverted functions: Booth and Matyas Function for which the results are described in this paper, and Sphere Function for which results are presented in [5], [6].

In conducted experiments, the applied CA was two-dimensional with size 500×500. The higher size of CA the more accurate results we retrieve. For proper analysis, Threshold in the algorithm was specified as: 1000 for Booth function and 10 for Matyas. The tests were conducted for a varying number of sensors ({5, 10, ..., 45, 50}) and steps of threshold ({2, 4, 6, 8, 10}).

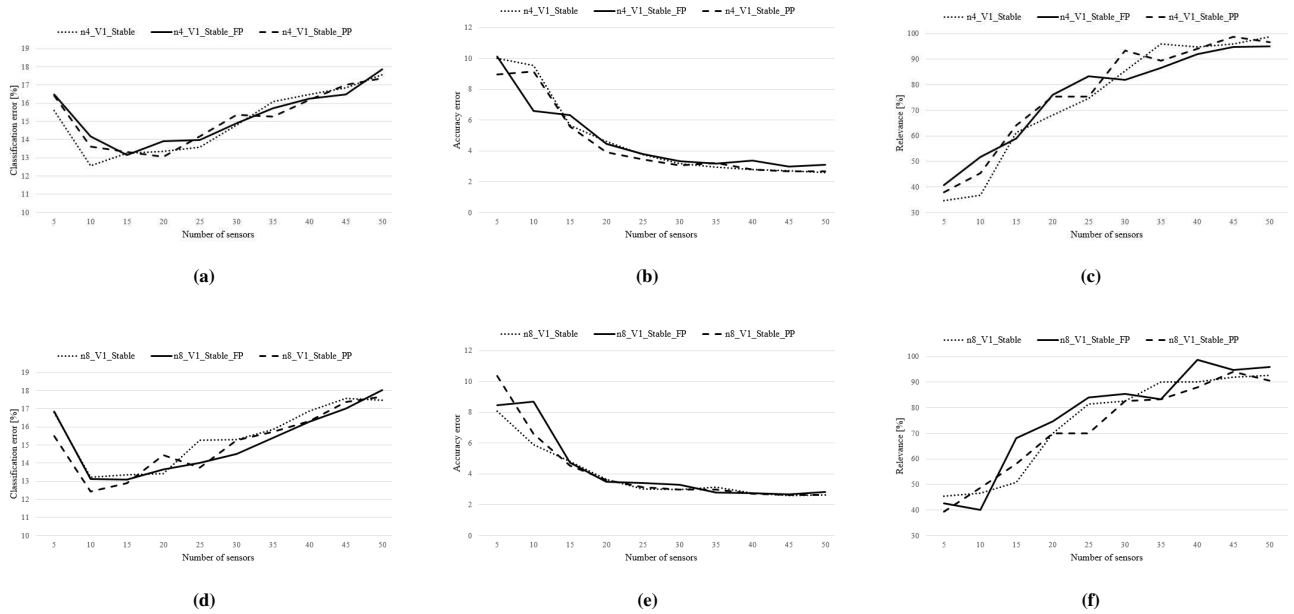


Fig. 1. Comparison of average values of **(a)** Classification error, **(b)** Accuracy error (real distance) and **(c)** Relevance (in [%]) resulted for varying number of sensors with use of each classification method ($n4\_V1\_stable$, $n4\_V1\_stable\_PP$ and $n4\_V1\_stable\_FP$) for von Neumann neighborhood on Booth function. Also, comparison of average values of **(d)** Classification error, **(e)** Accuracy error (real distance) and **(f)** Relevance (in [%]) obtained for varying number of sensors with use of each classification method ($n8\_V1\_stable$, $n8\_V1\_stable\_PP$ and $n8\_V1\_stable\_FP$) for Moore neighborhood on Booth function.

Each of tests series contains 500 single runs of the algorithm with the different random sensors spatial setup. From the set of conducted experiments were calculated the average values of classification error, accuracy error, and relevance for varying number of sensors.

A. Testing With the Booth Function

Figure 1 presents these results for Booth function with use of each classification method with von Neumann neighborhood. We can see that with growing number of sensors, the quality of results is generally getting better, except the classification error. The classification error presented in Figure 1(a) is not higher than 18% for the bordered numbers of sensors and has a parabolic trend. The lowest classification error has value about 13% for a number of sensors fluctuated from 10 to 25. This result is independent of the classification method. The error of accuracy (see, Figure 1(b)) is not higher than ~ 10 for only five sensors case and is generally going down near to 3 for 50 sensors setup. From the level of 25 sensors, accuracy error is lower than 4. The Relevance for 20 sensors is not lower than 70% and is going up. Since the setup of 40 sensors relevance being higher than 90%, where for the partially probabilistic method of classification ($n4\_V1\_stable\_PP$) and $n4\_V1\_stable$ method is near to 100% (see, Figure 1(c)).

Next step of experiments for Booth function is presented in Figure 1, where each of classification method was used with Moore neighborhood. In this case, we can see similar results and trends in particular for errors of classification and accuracy. Interesting is the fact that the fully probabilistic method of classification ($n8\_V1\_stable\_FP$) characterizes better results than for von Neumann neighborhood. Furthermore, in the case of relevance, the scores for $n8\_V1\_stable\_FP$ are slightly better than for other analyzed methods of classification. Relevance for this method and for 40 sensors is near to 100% (see, Figure 1(f)).

B. Testing With the Matyas Function

In this subsection are presented results of testing for Matyas function. Figure 2 shows these results for each of classification method with von Neumann neighborhood. As we can expect, with growing number of sensors, the quality of results is generally getting better, except the classification error. The classification error presented in Figure 2(a) is not higher than 18% only up to 15 sensor setup. For higher number of sensors classification error going up to $\sim 25\%$ for 50 sensors setup. This result is independent of the classification method. The error of accuracy (see, Figure 2(b)) is not higher than ~ 12 for only five sensors case and is generally going down near to 1 for 50 sensors setup (it is generally better than for Booth function). From the level of 20 sensors, accuracy error is lower than 4. The Relevance for 25 sensors is not lower than 70% and is going up to $\sim 90\%$, independently on the classification method (see, Figure 2(c)).

The experiments for Matyas function, where each of classification method was used with Moore neighborhood is presented in Figure 2. In this case, we can see similar results and trends to obtained for Matyas function and von Neumann neighborhood for each of assessment criteria. Moreover, classification error for the fully probabilistic method of classification ($n8\_V1\_stable\_FP$) seems to be better than other methods of classification (see, Figure 2(d)).

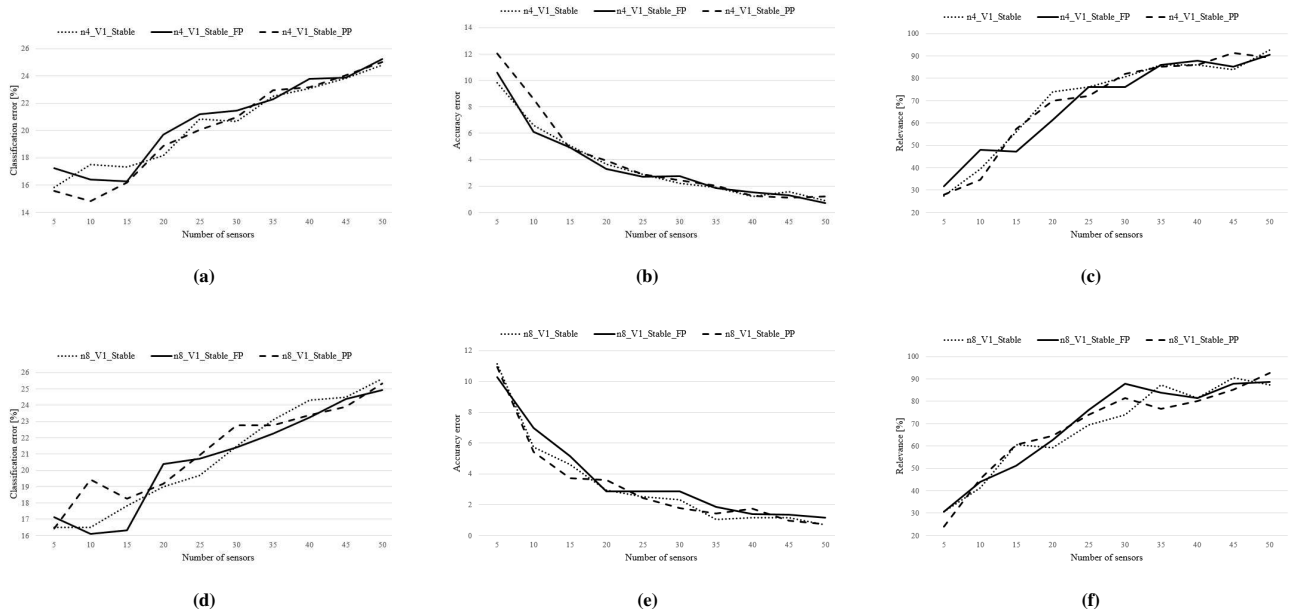


Fig. 2. Comparison of average values of **(a)** Classification error, **(b)** Accuracy error (real distance) and **(c)** Relevance (in [%]) resulted for varying number of sensors with use of each classification method ($n4\_V1\_stable$, $n4\_V1\_stable\_PP$ and $n4\_V1\_stable\_FP$) for von Neumann neighborhood on Matyas function. Also, comparison of average values of **(d)** Classification error, **(e)** Accuracy error (real distance) and **(f)** Relevance (in [%]) obtained for varying number of sensors with use of each classification method ($n8\_V1\_stable$, $n8\_V1\_stable\_PP$ and $n8\_V1\_stable\_FP$) for Moore neighborhood on Matyas function.

CONCLUSIONS

In this paper is proposed an algorithm of area optimization by layers and binary classification with use of three state 2D cellular automata. The proposed algorithm as an input utilizes the data reported by the set of sensors monitoring the area put into optimization. This data are subject to the classification based on two-dimensional three-state cellular automata, which classifies all points of the area to designate optimal subarea. The layers are categorized based on the level of values received from sensors. Such, approach gives a possibility to specify the areas in the different levels and after analysis could be selected optimal subarea or subarea included the optimal point.

The algorithm was verified and tested with use of two functions: Booth and Matyas included in the set of the functions applied for testing optimization/optimizing algorithms. The methods for interpretation of obtained results were introduced in conjunction with algorithms assessment criteria, like classification error, accuracy error, and relevance. The values of the algorithm characteristics corresponding to the algorithm run with each of three classification methods for different setups, i.e., the varying number of sensors (input data) were presented.

Conducted studies show that quite good results characterize proposed algorithm. Reasonably high relevance value, i.e. $\sim 90\%$, and higher was reached. Furthermore, the accuracy error are characterized by the low value. Performed tests show that this method could be used in different optimization problems starting from simple ones, as optimization of n-dimensional functions, and in more complicated tasks as designating the contaminated area based on the restricted number of mobile sensors data or estimating the source of airborne toxin.

Presented experiments prove that proposed algorithm is able to reduce the scanned area to the little size (even optimal). Furthermore, studies of the relevance of results obtained by the proposed algorithm indicate that it can be used as an optimization tool, which going to indicate the area including the optimum.

ACKNOWLEDGEMENTS

This work is supported by The Polish National Science Centre grant awarded by decision number DEC-2012/07/D/ST6/02488.

REFERENCES

[1] T. Fawcett, “Data mining with cellular automata.” *ACM SIGKDD Explorations Newsletter*, 10(1), pp. 32–39, 2008.
[2] P. Maji, B. Sikdar, and P. Chaudhuri, “Cellular automata evolution for pattern classification.” *LNCS 3305* , pp. 660–669. Springer Verlag 2004.
[3] P. Povalej, M. Lenic, and P. Kokol, “Improving ensembles with classificational cellular automata.” *LNCS 3305*, pp. 242–249. Springer 2004.
[4] M. Szaban: “Probabilistic 2D Cellular Automata Rules for Binary Classification.” M. Ganzha, L. Maciaszek, M. Paprzycki (eds.): Annals of Computer Science and Information Systems, Volume 8 (FedCSIS 2016), 2016, pp. 161-164, DOI: 10.15439/978-83-60810-90-3
[5] M. Szaban, A. Wawrzynczak: “The Algorithm of Area Optimization by Layers and Binary Classification with Use of Three State 2D Cellular Automata.” Proceedings of the International Conference on Control, Artificial Intelligence, Robotics and Optimization, 2018 (in print)
[6] L. Swider: “Optimization of two variables functions with use of data classification algorithms based on cellular automaton”. Master thesis, (2016), (in Polish).
[7] Test Functions for Unconstrained Global Optimization: $http://www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar\_files/TestGO\_files/Page364.htm$
[8] S. Wolfram: *A New Kind of Science*. Wolfram Media, 2002.
[9] Hutchinson, M., Oh, H. & Chen, W.H.: “A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors.” Information Fusion 36, (2017), pp. 130-148.
[10] Wawrzynczak A., Kopka P., Borysiewicz M.: “Sequential Monte Carlo in Bayesian assessment of contaminant source localization based on the distributed sensors measurements.”, LNCS 8385, (2014), pp. 407-417.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 85–93
DOI: 10.15439/2018F231
ISSN 2300-5963 ACSIS, Vol. 15

Modular Multi-Objective Deep Reinforcement Learning with Decision Values

Tomasz Tajmajer
Institute of Informatics, University of Warsaw
ul. Banacha 2, 02-097, Warsaw, Poland
Email: t.tajmajer@mimuw.edu.pl

***Abstract*—In this work we present a method for using Deep Q-Networks (DQNs) in multi-objective environments. Deep Q-Networks provide remarkable performance in single objective problems learning from high-level visual state representations. However, in many scenarios (e.g in robotics, games), the agent needs to pursue multiple objectives simultaneously. We propose an architecture in which separate DQNs are used to control the agent's behaviour with respect to particular objectives. In this architecture we introduce *decision values* to improve the scalarization of multiple DQNs into a single action. Our architecture enables the decomposition of the agent's behaviour into controllable and replaceable sub-behaviours learned by distinct modules. Moreover, it allows to change the priorities of particular objectives post-learning, while preserving the overall performance of the agent. To evaluate our solution we used a game-like simulator in which an agent - provided with high-level visual input - pursues multiple objectives in a 2D world.**

I. Introduction

MANY recent works on Reinforcement Learning focus on single-objective methods such as Deep Q-learning [1], [2]. As those methods provide great performance in tasks such as playing video games, many real-life problems require satisfying multiple objectives simultaneously. In single objective reinforcement learning the agent receives a single reward each time it performs an action. In multi-objective reinforcement learning (MORL) the agent receives multiple rewards - one for each objective. In particular, agents dealing with complex environments, such as autonomous robots or agents playing real-time video games, need to pursue multiple, often conflicting objectives.

To have a real-life example, lets consider an autonomous cleaning robot, which is able to clean floors, navigate through obstacles and autonomously return to charging station. The observable aggregated behaviour of such robot may be decomposed into three sub-behaviours: collision avoidance (ca), floor cleaning (fc) and recharging (rg). We may describe the objectives of the robot for each identified sub-behaviour in a multi-objective manner, or we can aggregate the sub-behaviours and define a single objective. In the former case, the robot-agent will receive a set of three rewards ([r_{ca}, r_{fc}, r_{rg}]) after each action. If the robot collides with a wall, it receives a negative reward related to collision avoidance (r_{ca}), yet the rewards related to floor cleaning and recharging do not depend on this event. However, in single-objective case, the robot will receive only one reward value ([r]) dependent on any of the three sub-behaviours. In case of collision, the the single-objective robot will receive a negative reward, but it will be indistinguishable from any negative reward provided with respect to other sub-behaviours such as depletion of batteries.

In single objective scenarios, we may find an optimal policy for which the sum of rewards collected by the agent is the highest possible. Methods such as Q-learning should converge to optimal policies [3]. However, for multi-objective problems, many such optimal policies may exist, depending on the trade-offs between satisfying particular objectives [4].

Autonomous agents, such as our example cleaning robot, are not really independent - they usually have a purpose defined by another agent: human. This aspect is often neglected in the literature, but is significant when considering practical applications of intelligent agents in robotics, automation or even when designing AIs for video games (always winning AI is not the one that many humans would like to play against). Our cleaning robot may follow a policy for which collision avoidance has greater importance than floor cleaning - in such case the robot should focus on avoiding collisions even at the cost of worse performance at floor cleaning. It is however for the user of such robot to decide, what should be the proportion between carefulness and cleanliness. The user may even want to fully disable some functions (behaviours) of the robot. Yet, state of the art reinforcement learning methods, such as Deep Q-Learning, do not allow to modify the behaviour of the agent after it was trained.

We see that when considering practical applications it is desired to have a multi-objective reinforcement learning method with the following features available post-learning: 1) ability to select the sub-set of pursued objectives and 2) ability to change the impact of particular objectives on the overall policy of the agent. As we will show later, the method presented in this paper possesses those features.

Multi-objective problems may be approached using *single-policy* or *multi-policy* methods. The simplest single-policy method uses a *scalarization function* [5], which converts multiple objectives into a single objective. Scalarization methods utilize a weight matrix to obtain a single score from multiple action-value functions. Some techniques assign linear priorities to objectives [6], [7]. This allows to obtain a single optimal policy with respect to objectives ordered by those priorities.

In contrast to single-policy methods, multi-policy MORL methods are used to find a set of policies. Their aim is to

find a set of policies that contains an approximately optimal policy for every possible user's preference [4]. In multi-policy methods, the preference of objectives does not need to be set a priori as a Pareto optimal policy for any preference may be obtained at runtime [8].

A natural approach in MORL is to use separate learning modules for each objective [9]. Modularity allows to decompose the problem into components that are to some extent independent [10]; modularity may be required for providing features desired in practical applications that were listed earlier. Some works deal with transforming complex single-objective problems to many simpler objectives [11]. Such methods may be used to benefit from modular approach while solving single-objective problems.

Although Deep Q-Networks gained much attention in recent years, not many works consider the use of DQNs in multi-objective problems. Recently authors of [12] proposed a multi-policy learning framework that utilizes Deep Q-Networks.

Learning behaviours in embodied agents, such as robots, is a problem well suited for reinforcement learning methods. In *embodied artificial intelligence*, the idea of *parallel, loosely coupled processes* [13] is proposed as a principle for designing embodied agents. It states, that the control logic for embodied agents should consists of many independent components dedicated for particular aspects of the agent's behaviour. The aggregated behaviour of an agent emerges from cooperation or competence among those components.

In this work we will present a method for combining multiple Deep Q-Networks for solving multi-objective problems. We will introduce decision values used for more advanced scalarization of multiple Q-functions. Furthermore we will combine decision values with user define priorities, to have an architecture that can dynamically adapt its behaviour with respect to user's preferences.

In section II we will briefly describe single- and multi-objective reinforcement learning. Next, in section III we will describe how many separate DQNs may be used together and we will define decision values. In section IV we will present a simple 2D game - a virtual environment including an autonomous agent that has a local (situated) sensory inputs and may pursue different objectives. Finally in the last section we will evaluate our solution and present the results of our experiments.

II. Background

A. Single Objective Reinforcement Learning

In the single-objective reinforcement learning an agent interacts with the environment by perceiving the state $s_t \in S$ and performing an action $a_t \in A$ for each step t. The actions are chosen by the agent according to some policy π. After performing an action, the agent receives a reward r_t. Then the agent observes the next state s_{t+1} and the process repeats. The goal of the agent is to maximize the expected discounted reward $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where $\gamma \in [0, 1]$ is the discount factor.

In Q-learning actions are selected based on $Q(s, a)$, which represents the expected discounted reward for performing action a in state s. For given state s, $a_t = \arg\max_a Q(s, a)$ is the optimal action. The policy of an agent, denoted by π, is the probability of selecting action a in state s. If the agent always selects the optimal action, then we say that it follows an optimal policy $\pi_\star$. Knowing the $Q(s, a)$ allows to create an optimal policy simply by selecting the action with the highest Q-value. Deep Q-learning utilizes Deep Neural Networks for approximating $Q(s, a)$ values, thus enabling this method to be used in many real-world applications. Deep Q-Networks [2] may be used used with high-level visual inputs such as those provided by video games.

B. Multi-Objective Reinforcement Learning

We may consider a more complex reinforcement learning scenario in which multiple objectives are pursued by the agent. Let O be the set of objectives of an agent. We may assign a priority p to each objective $o \in O$ such that o_k will have lower priority than o_j when $p(o_k) < p(o_j)$. For further analysis we will assume that $\forall_{o \in O} p(o) \geq 0$, so that priorities may be interpreted as weights.

The agent, instead of a single reward, receives a vector of rewards at each time-step t with respect to each objective o_i, i.e: $\vec{r}_t = [r_{1,t}, r_{2,t}, \dots r_{n,t}]$, where $r_{i,t}$ corresponds to objective o_i. For each objective o_i and step t we may define the discounted return as:

$$R_{i,t} = \sum_{k=0}^{\infty} \gamma^k r_{i,t+k} \tag{1}$$

Moreover, for each objective o_i there is a Q-function $Q_i(s, a)$ that represents the expected discounted return $R_{i,t}$, i.e: $Q_i(s, a) = \mathbb{E}\left[R_{i,t} \mid s_t = s, a_t = a\right]$.

We may define a vector of Q-functions, which includes $Q(s, a)$ for each objective o_i:

$$\vec{Q}(s, a) = [Q_1(s, a), Q_2(s, a), ..., Q_n(s, a)] \tag{2}$$

The function $Q_i(s, a)$ may be used by the agent to determine the optimal action with respect to objective o_i at time-step t, given state s_t:

$$a_{i,t} = \arg\max_a Q_i(s_t, a) \tag{3}$$

The vector $\vec{a}_t = [a_{1,t}, a_{2,t}, ..., a_{n,t}]$ consists of actions optimal with respect to particular objectives at a given time-step t. Because at each step, the agent may perform only a single action, a method of reducing $\vec{a}_t$ to a single action is required.

A common method for selecting a single action is the scalarization [5] of $\vec{Q}(s, a)$ using some scalarization function and a weight vector $\vec{w}$. Typically a linear scalarization is applied, so that:

$$SQ(s, a) = \sum_{i=1}^{N} w_i Q_i(s, a) \tag{4}$$

Then $SQ(s,a)$ may be used as in equation 3 to select an action. The weight wector in this case corresponds to priorities assigned to particular objectives.

In the further sections of this paper, we will show how to apply scalarization in Deep Q-Networks and we will introduce Decision Values to dynamically adjust the weights for improved performance of the agent. For simplicity, further in the text we will use the index i to note that a particular value or function is defined for any objective o_i, and by N we will define the number of objectives.

III. Using multiple DQNs

We have considered an agent that have multiple objectives, receives rewards with respect to those objectives and has a separate Q-function for each objective. In this section we will describe how to merge q-values obtained from Deep Q-Networks for different objectives and how the impact of particular DQNs on the behaviours of the agent may be controlled by using Decison Values. Finally we will describe the learning process utilizing DQNs with Decision Values. We will refer to our method as to Multi-Objective Deep Q-Network with Decision Values (MODQN-DV).

A. Combining Q-values

In case of multi-objective agent, we may use a separate DQN as an approximator for each $Q_i(s,a)$ in the $\vec{Q}(s,a)$ vector. Such agent would be controlled by multiple Deep Q-Networks working in parallel. Each DQN provides a list of q-values and we want to use q-values from all DQNs to select a single action a that will be performed by the agent

Let us define a vector $\vec{q_i}$ that consists of q-values provided by $Q_i(s,a)$ for each possible action $a \in A$ and a single objective o_i, i.e.:

$$\vec{q_i} = [Q_i(s,a_0), Q_i(s,a_1), ..., Q_i(s,a_j)] \tag{5}$$

In the single-objective case the optimal action a would be equal to a_j for such j that $\vec{q}_{i,j} = \max \vec{q_i}$. For multi-objective case we can use scalarization to sum up all $\vec{q}$ vectors and then select the action corresponding to the maximal value of such scaled q-value vector. In this approach, q-values may be interpreted as votes of certain DQN, which are summed-up and the highest-voted action is selected. We need to stress here that simply adding the vectors does not produce a meaningful result yet. The q-values produced by different Q-functions are not scaled. In general q-values may be any real numbers. If we want them to represent votes for particular actions, each $\vec{q_i}$ vector needs to be rescaled to $[0,1] \subseteq \mathbb{R}$. Many approaches for scaling the vector may be applied. In our experiments we use the following scaling function for which $\min(\vec{q_i})$ is mapped to 0 and $\max(\vec{q_i})$ to 1:

$$scale(\vec{x}) = \frac{\vec{x} - \min(\vec{x})}{\max(\vec{x} - \min(\vec{x}))} \tag{6}$$

The scalarized q-vector is then defined as:

$$\vec{q_s} = \sum_{i=1}^{N} w_i scale(\vec{q_i}) \tag{7}$$

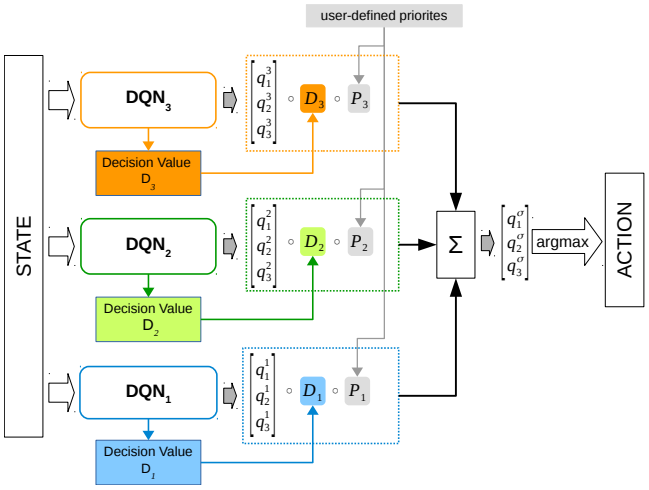


Fig. 1: Three Deep Q-Networks are working in parallel based on the same sensory input. Each DQN corresponds to different task pursued by the agent. Each DQN has an additional decision value output which acts as a dynamic weight used while summing up q-vectors from particular DQNs. User defined priorities are also used for weighting the decision from particual DQNs.

Now, using the rescaled $\vec{q_i}$ vectors we can sum them up and select one action with the highest total q-value. For example, let have actions a_1, a_2, a_3, weight vector $\vec{w} = [1,1,1]$, objectives o_1, o_2 and corresponding q-vectors $\vec{q_1} = [0, 0.6, 1]$ and $\vec{q_2} = [1, 0.5, 0]$. Adding them will result in vector $[1, 1.1, 1]$, for which the second element is the maximal, thus the corresponding action a_2 should be selected.

B. Decision Value

The scalarization allows to combine outputs from multiple DQNs. However, such a combination does not guarantee a meaningful action selection. Let us return to previous examples and consider a vacuum cleaner approaching a wall; actions a_1, a_2, a_3 correspond to turning left, going straight, and turning right respectively. If the vacuum cleaner perform the action proposed in q_1 it will turn right, alternatively if it uses q_2 then it will turn left. Using the sum will however lead to going straight forward and hitting the wall. So while both DQNs suggested a meaningful action, their sum is not meaningful at all. We see that using constant weights while summing q-values does not provide a solution for this problem.

To solve this issue, we would need to dynamically choose which q-value vectors are more important in a particular state. In other words, we would like to have a meta-policy for choosing the actual policy of the agent. However, as the agent pursues many objectives, it is hard to define this meta-policy with respect to all objectives. To overcome this problem we propose to indicate the *value* of the *decision* provided by each

DQN with respect to corresponding objective pursued by the agent.

The proposed *decision values* may be indicated independently by each DQN based on the current state and used as additional weights while summing up q-value vectors. Going back to the previous example: let assume that q_1 is the output from DQN associated with collision avoidance and q_2 is the output from DQN associated with cleaning. As the robot approaches a wall, the decision regarding collision avoidance is clearly more important than the decision regarding cleaning. This is because if the robot does not make any decision, it will collide with the wall and receive a negative reward with respect to collision avoidance objective. However, not making the decision will not affect cleaning objective (assuming that the cleanliness of the floor in front of him is not different than in other places). Thus, at this particular state the value of q_1 is higher than the value of q_2 and q_1 should be summed with a higher weight.

We may define the decision value signal $d \in [0, 1] \subseteq \mathbb{R}$, and by d_i denote the decision value associated with DQN_i. Now the scalarized q-vector would use decision values instead of constant weights:

$$\vec{q_d} = \sum_{i=1}^{N} d_i scale(\vec{q_i}) \tag{8}$$

We may additionally include the external preferences indicated by values of priorities p_i assigned to objectives as introduced in II-B. This way the q-values will be scaled both by dynamic decision values and static priorities. Moreover, for technical reasons, we need to add $\vec{\mu}$, which is a vector containing very small random values. This will ensure that in a rare cases when all decision values are equal to 0, a random action will be chosen. Finally the scaled, decision value- and priority- weighted q-value vector denoted by $\vec{q_\sigma}$ is equal to:

$$\vec{q_\sigma} = \vec{\mu} + \sum_{i=1}^{N} d_i p_i scale(\vec{q_i}) \tag{9}$$

C. Acquiring values of decisions

Now, as we have a method of applying decision values in the scalarization of multiple objectives, let us explain in more details how decision values are defined and how they can be learned by reinforcement learning.

First we should consider how objectives of an agent are defined. Again let us refer to the vacuum cleaning robot example. If the agent had only two objectives: a) to seek dirt and b) to avoid colliding with obstacles, then we could define two reward/terminal states: state A - state in which dirt is collected, state B - state in which the robot is colliding with something. There is a notable difference between those two states. In the first case, the agent should be rewarded positively, but in the latter case, it should be rewarded negatively. Moreover, if the agent is not in any of those states, it should be not rewarded at all. We can describe the first objective as being *attractive* (as it attracts the agent by positive rewards) and the second as being *repulsive* (as it repulses the agent by negative rewards). Many problems in robotics, games or other fields of AI may be presented using a set of attractive or repulsive objectives. In particular some problems may be decomposed into such set of objectives to promote more granular learning and control. Such decomposition is usually simpler and more intuitive compared to more advanced reward shaping techniques.

Let us consider an agent moving in a state-space with attractive and repulsive states. As the agent approaches one of those states, it becomes more critical to perform an action that will either move the agent towards such state or away from it. The value of the decision made with respect to an objective near a rewarding state rises as the distance to this state becomes shorter. This is a simple and intuitive heuristic: if an agent pursues multiple equally weighted objectives, then it probably should focus most on the objective that is already very close to being accomplished.

We can thus create a *decision reward* - the reward provided to the agent for performing a decision - which would be simply the absolute value of the reward provided with respect to an objective: $\rho_i = abs(r_i)$. Now we can define the decision value as a *state-value* function [3], returning the value of the state s under policy π, with respect to the decision rewards of a particular objective:

$$D_i(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \, \rho_{i,t+k+1} \,\middle|\, s_t = s \right] \tag{10}$$

Such defined decision value will provide high values around rewarding states (either positive or negative) and low values in states which are far from rewarding states. In any state, the decision value will provide the importance of particular objective. The proposed decision value function will hence provide values representing the chances of achieving a rewarding state (with respect to some objective o) given the current state s and following policy π. Where policy π is the policy provided by the Q-function for a particular objective.

It is important to note, that the decision value, as defined, can not be directly used for scalarization, because its value may be any positive number. Moreover, the range of the values provided for different objectives may be very broad. To overcome this problem, the decision value needs to be scaled to be in range $[0, 1]$ as noted in section III-B. However, the unscaled decision value is needed during learning as it will be shown in the next section. We will therefore denote the unscaled decision value by D_i and define the scaled decision value by d_i as follows:

$$d_i = \sigma \left(\frac{(D_i - \alpha_i)}{\beta_i} \right) \tag{11}$$

Where σ is the sigmoid function; α_i and β_i are derived during learning: α_i is an approximation of the mean value of D_i, while β_i is an approximation of the standard deviation of D_i.

D. Learning

Having defined decision values, we may move to the method of learning such values along with learning policies for particular objectives. We use Deep Q-Networks to approximate the values of Q-functions. Following the state-of-the-art in this field a DQN provides the approximated function $Q(s,a;\theta)$, where θ are the learnable parameters of the neural network. As in our model we use multiple DQNs, there is a function $Q_i(s,a;\theta_i)$ for a DQN_i related to objective o_i. Each DQN_i is optimised iteratively, using the following loss function for each iteration j:

$$L^Q_{i,j}(\theta_{i,j}) = \mathbb{E}_{(s,a,r_i,s')\sim U(M_i)}[(r_i + \\ + \gamma \max_{a'} Q_i(s',a';\theta^-_{i,j}) - Q_i(s,a;\theta_{i,j}))^2] \quad (12)$$

As introduced in [1], there are in fact two neural networks involved in the learning process of a single DQN. The *on-line network* $Q_i(s,a;\theta)$ is updated at each iteration, while the *target network* $Q_i(s',a';\theta^-)$ is updated only each K iterations. Moreover *experience replay* is used to further improve the learning process. The agent stores experienced states, actions and rewards in a *replay memory* M_i for each DQN_i respectively. Then at each iteration, each DQN_i is trained using a sample of past experiences selected uniformly at random from the corresponding replay memory M_i. Those samples are used as mini-batches for gradient descent optimization.

The Decision Value may be updated using TD-learning [3] similarly as for any state value function, by using the following update rule:

$$D_i(s_t) \leftarrow D_i(s_t) + \alpha\left[\rho_i + \gamma D_i(s_{t+1}) - D_i(s_t)\right] \quad (13)$$

As we use a neural network for approximating $D_i(s)$, we may define the loss function as follows:

$$L^D_{i,j}(\theta_{i,j}) = \mathbb{E}_{(s,\rho_i,s')\sim U(M_i)}[(\rho_i + \\ + \gamma D_i(s';\theta^-_{i,j}) - D_i(s;\theta_{i,j}))^2] \quad (14)$$

The decision value is provided by an additional output of the DQN and the learning procedure is analogical to Q-function. Moreover the decision value requires scaling, for which the parameters α and β need to be learned. If we include α and β in the neural network parameters θ, then the additional loss function for the decision value scaling would be defined as:

$$L^d_{i,j}(\theta_{i,j}) = \mathbb{E}_{(s)\sim U(M_i)}[(0.5 - \sigma(D_i(s;\theta_{i,j})))^2 + \\ + (1 - \max_s(D_i(s;\theta_{i,j})) + \min_s(D_i(s;\theta_{i,j})))^2] \quad (15)$$

The neural network is optimized using a combined loss function for Q-values, decision values and scaling of the decision values:

$$L_{i,j}(\theta_{i,j}) = L^Q_{i,j}(\theta_{i,j}) + L^D_{i,j}(\theta_{i,j}) + L^d_{i,j}(\theta_{i,j}) \quad (16)$$

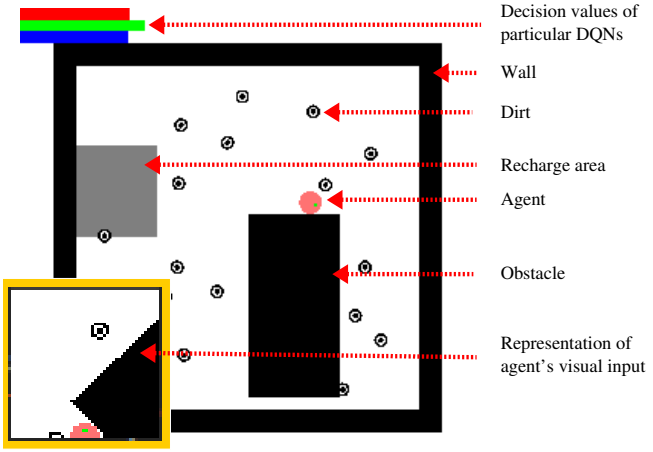


Fig. 2: Cleaner - a game-like virtual environment with agent pursuing multiple-objectives. The environment consists of the agent, walls, obstacles, recharge area and dirt. The agent perceives the environment by a visual input (a view from the top limited to a square located in the front of the agent). Agent may move forward and turn; its area of movement is limited by walls. Agent has three objectives: avoid walls, consume dirt and recharge.

IV. Evaluation

A. Cleaner - a 2D game-like virtual environment

To evaluate the solution presented in this paper we created *Cleaner* - a simple game-like virtual environment, simulating the behaviour of an autonomous vacuum cleaner. The environment consists of an agent, walls, recharge areas and dirt consumable by the agent. Cleaner is presented in Figure 2. The agent is a circular object that may move around the map by performing one of three actions: move forward, turn left and turn right. The map is a continuous space. The agent perceives the environment only by visual sense, i.e. a W x H pixel (width and height) rectangle situated in front of him. This visual input is converted to gray-scale (8bit). Agent's world (white) is surrounded by walls and filled with obstacles (black rectangles) which agent can not pass. Agent may pick up dirt and recharge itself. Dirt is indicated by three small coaxial circles (black), while recharging field is indicated by a gray rectangle. Dirt re-spawns at random positions on the map after being consumed by the agent. The quantity of dirt, recharge fields and obstacles is constant during the episode. Cleaner is a simplified simulation of a mobile robot moving on a flat surface (e.g. floor) with a video camera attached at the top of the robot pointed towards the floor.

The agent has a battery level $E \leq E_{max}$, which is decreased at each time step by E_{step}. The battery level may be increased when the agent enters the recharging area by $(1-E)\cdot 0.1$ each step. An episode ends when the agent's energy level drops to 0 or when 2000 steps pass. The agent starts each game with initial battery level $E = E_{start}$. The position of dirt, recharge

TABLE I: MODQN-DV learning hyperparameters

Parameter	Value
learning steps	1000000
replay memory size	10000
target network update rate	1000
learning rate	0.001
ϵ start value	1
ϵ end value	0.1
ϵ end step	100000
discount	0.99
batch size	32
optimizer	Adam

fields and obstacles as well as the initial position of the agent are chosen randomly at the start of the episode.

The agent has three objectives: (ca) collision avoidance, (fc) cleaning and (rg) recharging.

The rewards for particular objectives are as follows: objective (ca): -1 for collision, 0 otherwise; objective (fc): $+1$ for collecting dirt, 0 otherwise; objective (rg): -1 for for each step when $E < 0.1$, $(1 - E) \cdot 0.1$ while charging and 0 otherwise.

In all experiments described in this chapter, the game options were as follows: $E_{start} = E_{max} = 1.0$, $E_{step} = 0.001$. The size of the agent sight rectangle is $W = 50\,\text{px}$, $H = 50\,\text{px}$. The quantity of dirt is 20. The number of obstacles varies randomly from 1 to 5, and the number of charging areas varies randomly from 1 to 3.

B. MODQN-DV implementation

Our implementation of the MODQN-DV[1] was based on the baseline DQN implementation [14] developed by OpenAI using TensorFlow[15]. We expanded the standard DQN with additional decision value outputs and mechanism for scalarizing q-values from multiple DQNs. Each single DQN in a MODQN-DV consists of a convolutional network with three convolution layers and no pooling layers, followed by a fully connected layer and the output layer. Dueling [16] and double q-learning [17] were used. The additional decision value output is a single neuron linear layer connected to the state score layer used for dueling.

The parameters of the convolution network were kept default as provided in the baselines implementation. The size of the fully connected layer in our models is set to 128, and the size of the input image is our case is 50x50x1, thus the q-values are provided based only on an image input from a single state. The memory replay was modified to store rewards with respect to all objectives separately. The prioritized experience replay[18] was not used in our implementation. The hyperparameters used for training DQNs during evaluation are presented in Table I.

During training of the MODQN-DV, loss functions are used as specified in section III-D. DQNs for all objectives are trained simultaneously and scaled decision values are used for scalarization during learning.

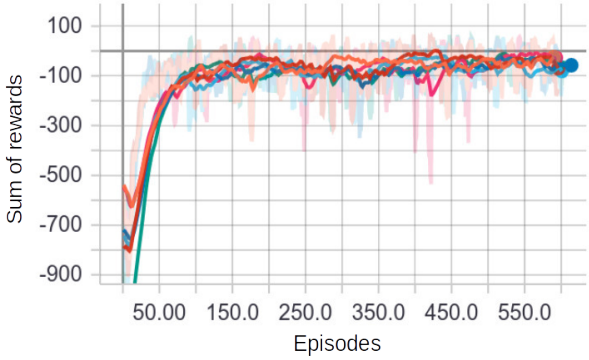


Fig. 3: The sum of rewards (smoothed) collected by MODQN-DV in cleaner over episodes of training. The plot shows data from 6 different runs.

C. Experiments

To evaluate our method we conducted a series of experiments utilizing MODQN-DV and the cleaner environment. In particular we compared the performance of multiple Deep Q-Networks for case a) where decision values were enabled for scalarization and case b) where the decision values were disabled. This comparison gave us a clear indication of the impact of decision values on the performance. We will refer to case (a) as MODQN-DV (b) as to MODQN.

The experiment for both cases (a) and (b) were conducted as follows. First the DQNs were trained using the implementation and parameters as provided in section IV-B and table I. In (a) the decision values were trained and used for scalarization. In (b) the decision values were disabled during training and their values were forcefully set to 1. During training, the user defined priorities for objectives were set to 1 in all cases (all objectives were weighted equally during scalarization). For each case the training procedure was repeated 6 times and all trained neural networks were saved. As show in figure 3, the learning of MODQN-DV is stable over time.

Next, the trained MODQN-DV and MODQN networks were used for evaluation with 10 different sets of user-defined priorities (p_{ca}, p_{fc}, p_{rg}) as provided in tables IIa and IIb. In a single evaluation, 100 episodes were played. The same sequence of randomly generated map layouts were used for each run. The sum of collected rewards were recorded for each run. For each set of priorities, 6 runs performed by 6 separately trained MODQN-DV and MODQN instances were averaged.

D. Results

The results presented in the tables IIa and IIb are averaged sums of collected rewards with respect to each objective, namely: Σr_{ca} for objective (ca) - collision avoidance, Σr_{fc} for objective (fc) - cleaning and Σr_{rg} for objective (rg) - recharging. $\Sigma\Sigma r_i$ is the sum of the sums of rewards - it indicates

[1]Source code available: https://github.com/ttajmajer/morl-dv

TABLE II: Evaluation results - sums of collected rewards for experiments with different priorities assigned to objectives with either enabled or disable decision values. Each case is compared against the baseline agent with all priorities set to 1.0. Values in bold denote objectives with the highest priority assigned. Green/red colour of the cell indicates whether a score is higher/lower compared to a corresponding score with opposite decision values configuration (enabled/disabled).

(a) decision values enabled

p_{ca}	p_{fc}	p_{rg}	Σr_{ca}	Σr_{fc}	Σr_{rg}	$\Sigma\Sigma r_i$
1	1	1	-88.4	47.6	-35.0	-75.9
$\Delta_{baseline}$			—	—	—	—
1	0	0	**-51.9**	24.0	-46.2	-74.1
$\Delta_{baseline}$			**41.4%**	-49.6%	-32.0%	2.37%
0	**1**	0	-303.0	**50.0**	-40.3	-293.3
$\Delta_{baseline}$			-242.7%	**5.1%**	-15.2%	-286.74%
0	0	**1**	-311.8	20.6	**-35.9**	-327.2
$\Delta_{baseline}$			-252.6%	-56.7%	**-2.7%**	-331.32%
0.5	0.3	0.2	**-45.7**	42.9	-39.2	-42.1
$\Delta_{baseline}$			**48.4%**	-9.9%	-12.2%	44.55%
0.5	0.2	0.3	**-68.4**	38.3	-39.5	-69.6
$\Delta_{baseline}$			**22.7%**	-19.5%	-12.9%	8.27%
0.2	**0.5**	0.3	-143.7	**51.3**	-33.2	-125.6
$\Delta_{baseline}$			-62.5%	**7.9%**	4.9%	-65.63%
0.3	**0.5**	0.2	-90.0	**50.2**	-34.7	-74.4
$\Delta_{baseline}$			-1.7%	**5.6%**	0.9%	1.93%
0.2	0.3	**0.5**	-140.6	45.2	**-34.7**	-130.1
$\Delta_{baseline}$			-59.0%	-4.9%	**0.6%**	-71.54%
0.3	0.2	**0.5**	-123.1	42.4	**-33.8**	-114.5
$\Delta_{baseline}$			-39.2%	-10.9%	**3.3%**	-51.01%

(b) decision values disabled

p_{ca}	p_{fc}	p_{rg}	Σr_{ca}	Σr_{fc}	Σr_{rg}	$\Sigma\Sigma r_i$
1	1	1	-61.0	51.3	-28.8	-38.5
$\Delta_{baseline}$			—	—	—	—
1	0	0	**-77.6**	32.0	-45.0	-90.6
$\Delta_{baseline}$			**-27.2%**	-37.6%	-56.6%	-135.26%
0	**1**	0	-518.2	**33.3**	-58.5	-543.4
$\Delta_{baseline}$			-749.1%	**-35.0%**	-103.4%	-1310.52%
0	0	**1**	-126.9	31.4	**-27.8**	-123.3
$\Delta_{baseline}$			108.0%	-38.7%	**3.3%**	-220.12%
0.5	0.3	0.2	**-35.7**	47.7	-35.7	-23.7
$\Delta_{baseline}$			**41.6%**	-7.0%	-24.1%	38.54%
0.5	0.2	0.3	**-40.3**	45.2	-32.6	-27.7
$\Delta_{baseline}$			**34.0%**	-11.9%	-13.4%	28.08%
0.2	**0.5**	0.3	-236.2	**49.8**	-37.8	-224.2
$\Delta_{baseline}$			-287.0%	**-2.8%**	-31.5%	-482.04%
0.3	**0.5**	0.2	-218.9	**50.3**	-38.7	-207.3
$\Delta_{baseline}$			-258.7%	**-1.8%**	-34.5%	-438.10%
0.2	0.3	**0.5**	-86.7	41.9	**-29.4**	-74.2
$\Delta_{baseline}$			-42.1%	-18.3%	**-2.1%**	-92.71%
0.3	0.2	**0.5**	-80.8	40.7	**-29.1**	-69.3
$\Delta_{baseline}$			-32.4%	-20.7%	**-1.3%**	-79.83%

the total performance of the agent. Priorities (p_{ca}, p_{fc}, p_{rg}) correspond to objectives (ca), (fc) and (cg).

The set of priorities: $(p_{ca} = 1, p_{fc} = 1, p_{rg} = 1)$ was used as the baseline for evaluation (also those priories were used during training). For each row in the tables IIa and IIb there is an additional row marked as $\Delta_{baseline}$ with values showing the percentage of gain or loss of collected rewards with respect to the baseline value for each case. The green and red colours of the cells indicate if the reward gain for a particular set of priorities was better compared to the corresponding case in the second table.

The aim of the evaluation was to test how the overall performance of the agent changes when priorities are different from the initial values used during training. As we can see in table IIa on 7 of 9 cases, the use of MODQN-DV helped to preserve (or even increased) the overall performance compared to the baseline (all priorities set to 1). Moreover, in almost all cases, the performance of the agent with respect to the objective with the highest priority (marked in bold in the tables) increased when decision values were used. On the contrary, when decision values were not used, changes in the priorities usually led to a decrease in the agent's performance, as presented in table IIb. The results show that the proposed solution has a significant impact on the performance when priorities are modified post-training. The average change in the agent's performance, calculated over all evaluation cases, is -27.5% when using decision values and -69.1% when decision values are not used. The average change for the objective with the highest priority is 11.3% and 4.4% respectively.

It should be noted however, that in the baseline case (all priorities set to 1), the overall performance of the agent was lower when decision values were enabled. A possible explanation of this issue is that decision values introduce additional noise to action selection. In some cases, the final q-values associated with particular actions may be very similar (e.g. when the agent does not perceive any objects). Then, action selection depends heavily on the decision values; if there is no dominating decision value, then there may be a lot of variance in action selection, thus actions may be selected based on different policies (from different DQNs) in each step. This may lead to a chaotic behaviour in states that are "far" from any rewarding states. One possible way of overcoming this issue is providing a sequence of states as the input to DQNs rather than a single state to stabilize the outputs.

It is also worth noticing how the decision values change as the agent moves. As expected, the decision value for a particular objective rises when the agent approaches a state where it could receive a reward. For instance, the value of collision avoidance rises significantly when the agent is very close to a wall or an obstacle. Moreover, the decision value

drops when the agent is in a state far from receiving a reward. For example, if the agent does not perceive any walls or obstacles, then the collision avoidance decision value is lower than average. The agent thus usually selects the action, which is related to the most promising objective at a particular state.

V. Conclusions

In this paper, we presented a method for using multiple Deep Q-Networks to approach multi-objective problems called Multi Objective Deep Q-Networks with Decision Values (MODQN-DV). We introduced decision values to DQNs in order to improve the scalarization of outputs from multiple DQNs. Our method requires only slight modification of existing DQN architectures, while it introduces a number of benefits: 1) it enables the decomposition of problems in to smaller sub-problems, for which independent DQNs may be trained simultaneously, 2) it provides a method for robust manipulation of priorities after the training, which also allows to completely disable DQNs responsible for particular behaviour/objective, 3) it allows to add new objectives to already trained agent without the need of retraining and to tune their impact on the behaviour of the agent.

In the experimental part, we shown that in most cases MODQN-DV improves the performance of the agent, that uses a different set of priorities compared to the training phase. The results are promising, however more tests should be performed using other benchmarks. Moreover, more work needs to be done to reduce the impact of the noise in decision values on the overall performance of the agent.

In this paper, we also introduce *cleaner* - a benchmark for multi-objective reinforcement learning problems that provides visual state representation. The authors are not aware of any other existing multi-objective benchmark that would be comparable to atari games benchmark or other provided by OpenAI.

In future work we want to improve the performance of MODQN-DV; one possible improvement is the use of common convolutional layers for all DQNs. It is particularly interesting to use MODQN-DV in very complex environments, such as video games. Recently published Starcraft 2 learning environment may be a good choice for further tests of MODQN-DV architecture as strategy games may be perceived as multi-objective problems.

Acknowledgment

Author would like to thank Piotr Wasilewski and Andrzej Janusz for their scientific supervision and valuable suggestions. This work was supported by the National Science Centre in programme SONATA 1, grant no. 2011/01/D/ST6/06981.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *NIPS Deep Learning Workshop*, 2013.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb 2015, letter. [Online]. Available: http://dx.doi.org/10.1038/nature14236

[3] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998. ISBN 0262193981

[4] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *J. Artif. Int. Res.*, vol. 48, no. 1, pp. 67–113, Oct. 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2591248.2591251

[5] K. V. Moffaert, M. M. Drugan, and A. Nowé, "Scalarized multi-objective reinforcement learning: Novel design techniques," in *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, April 2013. doi: 10.1109/ADPRL.2013.6615007. ISSN 2325-1824 pp. 191–199.

[6] L. Barrett and S. Narayanan, "Learning all optimal policies with multiple criteria," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008. doi: 10.1145/1390156.1390162. ISBN 978-1-60558-205-4 pp. 41–47. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390162

[7] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Machine Learning*, vol. 84, no. 1, pp. 51–80, Jul 2011. doi: 10.1007/s10994-010-5232-5. [Online]. Available: https://doi.org/10.1007/s10994-010-5232-5

[8] K. Van Moffaert and A. Nowé, "Multi-objective reinforcement learning using sets of pareto dominating policies," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3483–3512, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2750356

[9] N. Sprague and D. Ballard, "Multiple-goal reinforcement learning with modular sarsa(o)," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, ser. IJCAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 1445–1447. [Online]. Available: http://dl.acm.org/citation.cfm?id=1630659.1630892

[10] P. Raicevic, "Parallel reinforcement learning using multiple reward signals," *Neurocomputing*, vol. 69, no. 16–18, pp. 2171 – 2179, 2006. doi: http://doi.org/10.1016/j.neucom.2005.07.008 Brain Inspired Cognitive SystemsSelected papers from the 1st International Conference on Brain Inspired Cognitive Systems (BICS 2004)1st International Conference on Brain Inspired Cognitive Systems (BICS 2004). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231205003036

[11] T. Brys, A. Harutyunyan, P. Vrancx, M. E. Taylor, D. Kudenko, and A. Nowe, "Multi-objectivization of reinforcement learning problems by reward shaping," in *2014 International Joint Conference on Neural Networks (IJCNN)*, July 2014. doi: 10.1109/IJCNN.2014.6889732. ISSN 2161-4393 pp. 2315–2322.

[12] H. Mossalam, Y. M. Assael, D. M. Roijers, and S. Whiteson, "Multi-objective deep reinforcement learning," *CoRR*, vol. abs/1610.02707, 2016. [Online]. Available: http://arxiv.org/abs/1610.02707

[13] R. Pfeifer and C. Scheier, *Understanding Intelligence*. Cambridge, MA, USA: MIT Press, 2001. ISBN 026266125X

[14] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Openai baselines," https://github.com/openai/baselines, 2017.

[15] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[16] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 1995–2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045601

[17] H. V. Hasselt, "Double q-learning," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 2613–2621. [Online]. Available: http://papers.nips.cc/paper/3964-double-q-learning.pdf

[18] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *CoRR*, vol. abs/1511.05952, 2015.

3rd International Workshop on Artificial Intelligence in Machine Vision and Graphics

THE main objective of the 3rd Workshop on Artificial Intelligence in Machine Vision and Graphics (AIMaViG'18) is to provide an interdisciplinary forum for researchers and developers to present and discuss the latest advances of artificial intelligence in the context of machine vision and computer graphics. The workshop covers the whole range of AI-based theories, algorithms, technologies and systems for diversified and heterogeneous areas of vision and graphics.

TOPICS

The topics and areas include but are not limited to:

- image processing
- scene analysis, modeling, and understanding
- machine vision
- pattern matching and pattern recognition
- image synthesis, including three-dimensional imaging and solid modeling
- computer-aided graphic arts and animation
- mathematical approaches to image processing, analysis, and synthesis
- computational geometry
- image models and transforms
- visualization and graphical data presentation
- diagrammatic knowledge representation and reasoning
- monocular and stereo vision
- modeling of human visual perception
- innovative uses of various graphic and vision devices and systems

EVENT CHAIRS

- **Kwaśnicka, Halina,** Wrocław University of Science and Technology, Poland
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates

PROGRAM COMMITTEE

- **Dias, Jorge,** Khalifa University, United Arab Emirates
- **Foresti, Gian Luca,** University of Udine, Italy
- **Janusz, Andrzej,** University of Warsaw, Poland
- **Karhang, Maylor Leung,** Universiti Tunku Abdul Rahman, Malaysia
- **Kasprzak, Włodzimierz,** Warsaw University of Technology, Poland
- **Kulikowski, Juliusz,** Institute of Biocybernetics and Biomedical Engineering, Poland
- **Sikos, Leslie F.,** University of South Australia, Australia
- **Subbotin, Sergey,** Zaporizhzhya National Technical University, Ukraine
- **Tomczyk, Arkadiusz,** Łódź University of Technology, Poland
- **Werghi, Naoufel,** Khalifa University of Science and Technology, United Arab Emirates

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 97–100
DOI: 10.15439/2018F89
ISSN 2300-5963 ACSIS, Vol. 15

Baker's Cyst Classification Using Random Forests

Adam Ciszkiewicz, Grzegorz Milewski
Institute of Applied Mechanics
Cracow University of Technology,
al. Jana Pawła II 37, 31-864 Cracow, Poland
Emails: acisz@poczta.fm,
milewski@mech.pk.edu.pl

Jacek Lorkowski
Department of Orthopaedics and Traumatology
Central Clinical Hospital
of the MSWiA in Warsaw
Wołoska 137, 02-507 Warsaw, Poland
Email: jacek.lorkowski@gmail.com

***Abstract*—In this paper, a classification procedure for Baker's cysts was proposed. The procedure contained two subprocedures: the image preprocessing (dual thresholding, labeling, feature extraction) and the classification (Random Forests, cross validation). In total, five features were required to classify the cysts. These geometric features represented the location, the area and the convexity of the cyst. The procedure was proven effective on a set 436 varied MRI images. The set contained 68 images with cysts ready for aspiration and was oversampled with the SMOTE approach. The proposed method operates on 2D MRI images. This reduces the time of diagnosis and, with the ever increasing demand for MRI scanners, is justified economically. The method can be employed in systems for autonomous and semi-autonomous Baker's cyst aspiration or as a standalone package for MRI images annotation. Furthermore, it can be also extended to other fluid-based medical conditions in the knee.**

I. Introduction

WITH the ageing of the world population and the ever present health sector shortages [1], automation of surgical procedures is increasingly more common and important. For some of the simpler medical conditions, it is possible to develop robotic systems to treat them nearly autonomously. This in turn offloads the medical staff – their time and experience can be devoted to difficult and complicated surgical procedures.

A Baker's cyst is a very common medical condition (see Fig. 1). The cyst is a synovial capsule filled with fluid, which often occurs when the knee is in an inflammatory state. This condition is not dangerous, but it can affect the patient's quality of life through pain and reduced range of motion of the knee. The safest and easiest method of treatment for this condition is aspiration [2]. The aspiration is usually preceded by Magnetic Resonance Imaging (MRI) [3]. Often, full 3D scans are obtained, as they can be used to precisely diagnose the cyst and chose the best approach for aspiration. Despite its simplicity, fully autonomous aspiration of Baker's cyst remains largely unexplored in the literature. To automate this procedure, it is necessary to propose a classification procedure, capable of determining whether the knee contains a Baker's cyst that is ready for aspiration.

Two major approaches to medical image classification can be distinguished. In the first one, the learning algorithm is supplied a raw image and it learns meaningful features from the image automatically [4]. This approach is intuitive but requires large training sets and time-consuming training to provide good results. Such datasets can only be obtained through scientific collaborations, which are often focused on pressing medical issues. In case of simpler medical conditions, the second approach to classification can be employed. This method involves feature extraction coupled with image processing [5]. As the features are defined by the user, smaller training sets can be used. To extract the features it is often necessary to segment the image first. While there are many methods available for the medical image segmentation [6], the thresholding remains one of the more popular approaches [7]. The threshold can be set manually by the user or automatically with one of the available methods [8], [9]. Both approaches were utilized in this study.

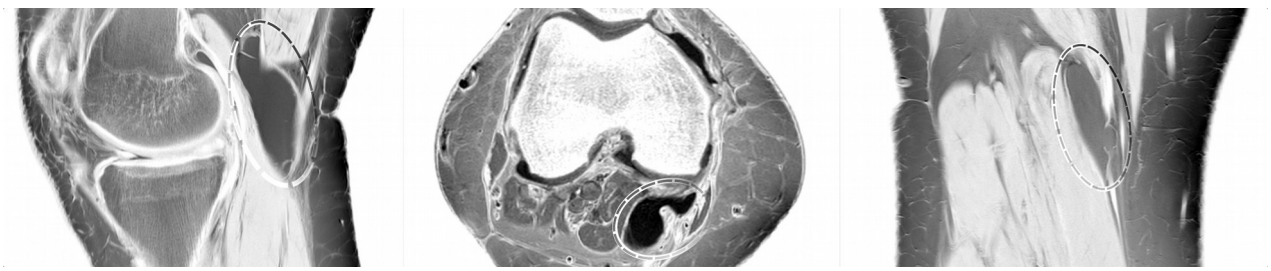


Fig 1. The Baker's cyst.

In a broader context, a classification procedure can be seen as a module in systems for autonomous and semi-autonomous surgery, in which it is used for diagnosis. The other problems in this area include: physical simulation of soft tissues and joints, medical tool path planning, surgery planning and surgery optimization [10]–[15].

In this paper, a procedure for Baker's cyst classification was proposed. The procedure was based on Random Forests and operated on 2D scans from Magnetic Resonance Imaging (MRI). The scans were preprocessed using custom, dual thresholding and labeling. Then, five geometric features were extracted from the preprocessed images. These features were then used to train the Random Forest (with cross-validation). The procedure was repeated 40 times to ensure that the obtained results were independent of the random number generator.

II. Method

A. The input dataset

The input dataset contained 12 spatial MRI sets of the knee. In total, the number of 2D MRI slices was 436. Out of them only 68 contained a Baker's cyst that was ready for aspiration. These 68 images were selected by an experienced orthopedic surgeon. The disproportion between the classes is very common in medical classification problems. Typical machine learning algorithms don't perform well with such datasets. Therefore, it is necessary to balance them. In this study, the input dataset was balanced with the Synthetic Minority Over-sampling Technique (SMOTE) [16] using Imbalanced-learn [17].

The proposed procedure was tested on three different MRI sequences: PDW SPAIR (Proton Density Weighted Spectral Attenuated Inversion Recovery), STIR (Short Tau Inversion Recovery) and PDW FatSat (Proton Density Weighted Fat Saturation) MRI sequences. In all of these sequences, the fluid-based structures appear hyper-intense, as seen in Fig. 2. The images were imported into Python using Pydicom [18].

The Baker's cysts are typically diagnosed with 3D MRI scans. The idea to use 2D slices instead of full 3D sets was partially inspired by the FAST USG, often employed in abdominal cavity diagnosis. In FAST USG the patient is diagnosed with only a few sweeps of the transducer. The diagnosis takes between 10 and 20 seconds, which is less expensive than a full sweep. These advantages also apply to MRI. With the ever growing need for MRI-based diagnosis, the availability of MRI scanners is a significant factor. Furthermore, some patients can't undergo full MRI scanning and others may feel uncomfortable during the procedure. When using singular 2D MRI scans for diagnosis, as in the proposed procedure, these issues are no a longer a concern.

B. Image processing

The initial threshold for the image segmentation was obtained using the method proposed in [9]. This threshold was then modified as follows:

$$thr_{fin} = mfn[\boldsymbol{I}(x,y) * (\boldsymbol{I}(x,y) > thr)], \quad (1)$$

where: thr_{fin} – the final threshold, thr – the threshold computed with the method presented in [9], $\boldsymbol{I}(x, y)$ – the input image, $mfn()$ – a function that computes the mean value of a matrix using only the nonzero elements.

In the next step the image was labeled with a 3 by 3 pixel mask. After the labeling, the largest object in the segmented image was selected. This object (see Fig. 3b) was assumed to be the cyst candidate. In the third step, the boundary of the knee was obtained (see Fig. 3c) using the following threshold:

$$thr_b = k\, max[\boldsymbol{I}(x,y)] + (1-k)\, min[\boldsymbol{I}(x,y)], \quad (2)$$

where: k – the thresholding parameter (here: k = 0.07). Finally, the obtained bounding box of the knee was applied to the image of the segmented object (see Fig. 3d).

C. Features

In this study, the following 5 features were used in the classification procedure:

a) $object_{area}$ – a typical feature used in medical classification procedures, which represented the area of the segmented object [mm^2],

b) *circularity* – a dimensionless measure of how circular the segmented object was; computed as a ratio between the area of the Largest Empty Circle (LEC) inside the object LEC_{area} and the object area $object_{area}$:

$$circularity = LEC_{area} / object_{area}. \quad (3)$$

The LEC (see Fig. 4b) was obtained using the method presented in [19],

c) *convexity* – a relative feature that measured the convexity of the segmented object; computed as a ratio between the object area $object_{area}$ and the area of the convex hull of the object $convexhull_{area}$ (see Fig. 4a):

$$convexity = object_{area} / convexhull_{area}. \quad (4)$$

The convex hull of the object was computed using Scipy [20],

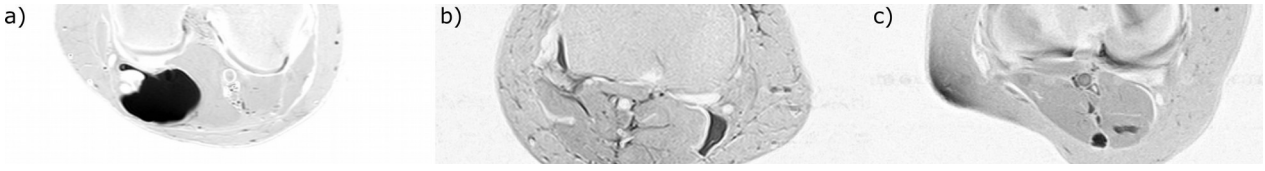


Fig 2. The supported MRI sequences: a) PDW SPAIR, b) STIR, c) PDW FatSat.

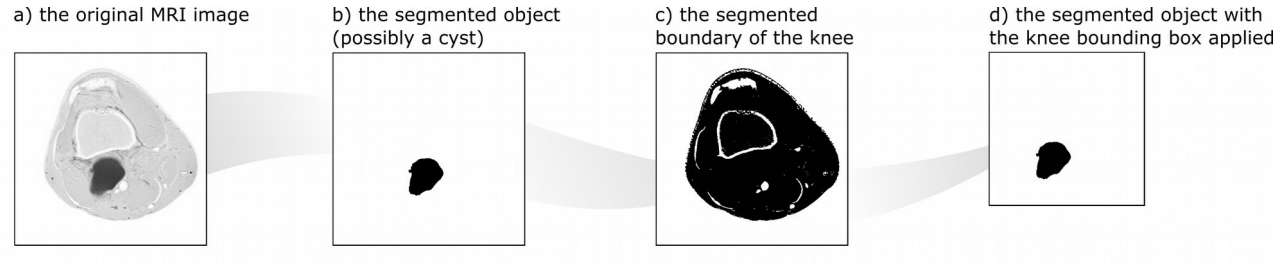


Fig 3. The image processing.

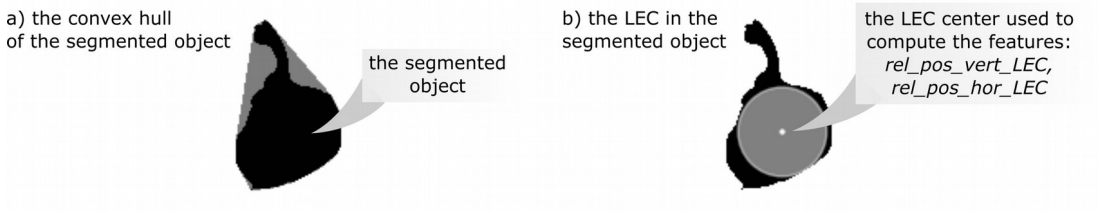


Fig 4. The convex hull and the LEC in the segmented image.

d) *rel_pos_vert_LEC* – the relative vertical position of the LEC center (see Fig. 4b),

e) *rel_pos_hor_LEC* – the relative horizontal position of the LEC center.

The first feature separated the cysts from the veins and the arteries – the cysts ready for aspiration often had larger areas. The second and the third feature helped distinguish cases with segmentation errors (multiple, small and connected objects or images with no hyper-intense objects), Furthermore, the second and the third feature were computed by dividing the area of the LEC and the convex hull respectively by the object_area. This made them independent of the first feature.

The fourth and the fifth feature differentiated the cysts from other fluid-related conditions in the knee, such as knee effusions – the cysts usually occur in the posterior side of the knee. With the proposed skin segmentation procedure, the LEC center was computed with regards to the actual boundary of the knee – not the image (see Fig. 3d). This made the procedure more general and reliable.

D. The classifier

Decision trees are among the most popular classification algorithms in machine learning. They are easy to implement and quick to train. Despite these advantages, the trees tend to overfit data. This is a serious drawback, especially when working with limited datasets. A natural extension to this approach, which addresses this flaw, is Random Forest [21]. In this algorithm several decision trees are trained on different subsets of the training dataset and the final classification is based on majority vote. It is worth mentioning that, unlike many classification algorithms, Random Forests do not require data preconditioning. This means that raw feature values can be used to train them and classify new samples.

In this study, Random Forest contained 10 decision trees, based on Gini impurity. The depth of the tress was not limited. The classifier was implemented using Scikit-learn [22]. The training and the testing were performed with a 4-fold cross validation, while the performance of the classifier was measured with the following indicators:

$$\begin{aligned} Sen &= TP/(TP+FN), \\ Spec &= TN/(TN+FP), \\ F_1 &= 2*TP/(2*TP+FN+FP), \end{aligned} \tag{5}$$

where: *Sen* – the sensitivity, *Spec* – the specificity, F_1 – the F_1 score (also referred to as balanced *F*-score), *TP* – the number of true positive cases (the knee contains a cyst ready for aspiration and is classified for aspiration), *FP* – the number of false positive cases (the knee does not contain a cyst ready for aspiration and is classified for aspiration), *TN* – the number of true negative cases (the knee does not contain a cyst ready for aspiration and is not classified for aspiration), *FN* – the number of false negative cases (the knee contains a cyst ready for aspiration and is not classified for aspiration).

III. Results and discussion

The training and test procedures were repeated 40 times to factor in the effects of shuffling and oversampling. The average and the best (based on the F1-score) results over the 40 runs were summarized in Table 1.

As seen in Table 1, the mean values of the performance indicators were higher than 95.0%. Furthermore, the best run achieved F_1 of 99.4 % and specificity of 100.0 %. This proves that the proposed method is capable of classifying the cysts for aspiration based on 2D MRI images. It is worth mentioning, that this classification problem would be easier with 3D MRI images. In three dimensions, the cyst can be easily distinguished from veins and arteries based on its spherical shape. Nevertheless, 3D MRI scanning can be time

TABLE I.
THE PERFORMANCE INDICATORS BASED ON THE 40 RUNS

	TP	*FP*	*TN*	*FN*	*Sen* [%]	*Spec* [%]	F_1 [%]
avg	88.6± 3.0	2.0± 3.0	94.0± 2.8	4.1± 3.8	95.6± 4.2	97.8 ± 3.4	96.7± 2.6
best	89	0	94	1	98.9	100.0	99.4

consuming. Furthermore, the demand for MRI scanners exceeds the supply in most medical centers in the world. Therefore, good performance using 2D slices can be seen as an advantage of the proposed procedure.

Currently, the procedure analyzes only the largest segmented object in the knee (these objects were classified by the surgeon and used to train/test the Random Forest). Nevertheless, in some rare cases the cyst may not be the largest segmented object. This issue can be solved by iteratively checking and classifying all of the segmented objects.

IV. CONCLUSION

In this paper, a classification procedure for Baker's cyst was proposed. The procedure was composed of two subroutines: the image preprocessing (coupled with feature extraction) and classification using Random Forests. The procedure was proven effective on a set 468 varied MRI images. The images were obtained using three, different MRI sequences. Good performance using 2D slices can be seen as an advantage of the proposed procedure. This reduces the time of diagnosis, improves patient's comfort and, with the ever increasing demand for MRI scanners, is justified economically. The method can be employed in systems for autonomous and semi-autonomous Baker's cyst aspiration or as a standalone package for MRI images annotation. Furthermore, it can be also extended to other fluid-based medical conditions in the knee.

REFERENCES

[1] X. Scheil-Adlung, T. Behrendt, and L. Wong, "Health sector employment: A tracer indicator for universal health coverage in national Social Protection Floors," Hum. Resour. Health, vol. 13, no. 1, pp. 1–8, 2015, http://dx.doi.org/10.1186/s12960-015-0056-9

[2] S. Jamshed and L. M. Snyder, "An Intact Dissecting Baker's Cyst Mimicking Recurrent Deep Vein Thrombosis," J. Investig. Med. High Impact Case Reports, vol. 4, no. 2, p. 2324709616650703, Apr. 2016, http://dx.doi.org/10.1177/2324709616650703

[3] T. J. Frush and F. R. Noyes, "Baker's Cyst: Diagnostic and Surgical Considerations.," Sports Health, vol. 7, no. 4, pp. 359–65, Jul. 2015, http://dx.doi.org/10.1177/1941738113520130

[4] W. Sun, T.-L. Tseng, J. Zhang, and W. Qian, "Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data," Comput. Med. Imaging Graph., vol. 57, pp. 4–9, 2017, http://dx.doi.org/10.1016/j.compmedimag.2016.07.004

[5] P. Zarychta, P. Badura, and E. Pietka, "Comparative analysis of selected classifiers in posterior cruciate ligaments computer aided diagnosis," Bull. Polish Acad. Sci. Tech. Sci., vol. 65, no. 1, pp. 63–70, 2017, http://dx.doi.org/10.1515/bpasts-2017-0008

[6] D. J. Withey and Z. J. Koles, "Three Generations of Medical Image Segmentation: Methods and Available Software," Int. J. Bioelectromagn., vol. 9, no. 2, pp. 67–68, 2007.

[7] T. Markiewicz et al., "Thresholding techniques for segmentation of atherosclerotic plaque and lumen areas in vascular arteries," Bull. Polish Acad. Sci. Tech. Sci., vol. 63, no. 1, pp. 269–280, 2015, http://dx.doi.org/10.1515/bpasts-2015-0031

[8] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Trans. Syst. Man. Cybern., vol. 9, no. 1, pp. 62–66, Jan. 1979, http://dx.doi.org/10.1109/TSMC.1979.4310076

[9] J. C. Jui-Cheng Yen, F. J. Fu-Juay Chang, and S. Shyang Chang, "A new criterion for automatic multilevel thresholding," IEEE Trans. Image Process., vol. 4, no. 3, pp. 370–378, Mar. 1995, http://dx.doi.org/10.1109/83.366472

[10] O. A. Pappalardo et al., "Mass-spring models for the simulation of mitral valve function: Looking for a trade-off between reliability and time-efficiency," Med. Eng. Phys., vol. 47, pp. 93–104, 2017, http://dx.doi.org/10.1016/j.medengphy.2017.07.001

[11] N. Chentanez et al., "Interactive Simulation of Surgical Needle Insertion and Steering," ACM Trans. Graph., vol. 28, no. 3, 2009, http://dx.doi.org/10.1145/1531326.1531394

[12] G. J. Vrooijink, M. Abayazid, S. Patil, R. Alterovitz, and S. Misra, "Needle path planning and steering in a three-dimensional non-static environment using two-dimensional ultrasound images," Int. J. Rob. Res., vol. 33, no. 10, pp. 1361–1374, Sep. 2014, http://dx.doi.org/10.1177/0278364914526627

[13] K. C. Assi, S. Grenier, S. Parent, H. Labelle, and F. Cheriet, "A physically based trunk soft tissue modeling for scoliosis surgery planning systems.," Comput. Med. Imaging Graph., vol. 40, pp. 217–28, Mar. 2015, http://dx.doi.org/10.1016/j.compmedimag.2014.11.002

[14] A. Ciszkiewicz, J. Lorkowski, and G. Milewski, "A novel planning solution for semi-autonomous aspiration of Baker's cysts," Int. J. Med. Robot., p. e1882, 2018, http://dx.doi.org/10.1002/rcs.1882

[15] A. Ciszkiewicz and G. Milewski, "Path planning for minimally-invasive knee surgery using a hybrid optimization procedure," Comput. Methods Biomech. Biomed. Engin., vol. 21, no. 1, 2018, http://dx.doi.org/10.1080/10255842.2017.1423289

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002, http://dx.doi.org/10.1613/jair.953

[17] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," J. Mach. Learn. Res., vol. 18, no. 17, pp. 1–5, 2017.

[18] D. Mason, "SU-E-T-33: Pydicom: An Open Source DICOM Library," Med. Phys., vol. 38, no. 6, p. 3493, 2011, http://dx.doi.org/10.1118/1.3611983

[19] G. T. Toussaint, "Computing largest empty circles with location constraints," Int. J. Comput. Inf. Sci., vol. 12, no. 5, pp. 347–358, 1983.

[20] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," Comput. Sci. Eng., vol. 13, no. 2, pp. 22–30, Mar. 2011, http://dx.doi.org/10.1109/MCSE.2011.37

[21] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, http://dx.doi.org/10.1023/A:1010933404324

[22] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2012.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 101–105
DOI: 10.15439/2018F286
ISSN 2300-5963 ACSIS, Vol. 15

Barley Variety Recognition with Viewpoint-aware Double-stream Convolutional Neural Networks

Przemysław Dolata
Wrocław University of Science and Technology
ul. Wyb. Wyspiańskiego 27,
50-370 Wrocław, Poland
Email: przemyslaw.dolata@pwr.edu.pl

Jacek Reiner
Wrocław University of Science and Technology
ul. Wyb. Wyspiańskiego 27,
50-370 Wrocław, Poland
Email: jacek.reiner@pwr.edu.pl

***Abstract*—Varietal homogeneity is an important factor in quality of malting barley, but its inspection is difficult. Biochemical methods are expensive and inefficient, while machine vision suffers due to high variability of the grains' features. In our previous work, we have shown a convolutional neural network for simultaneous feature extraction and classification of image data basing on multiple views. It was suggested that for machine vision inspection, the observed side of a grain should be taken into account – dorsal and ventral sides of each kernel exhibit different features. In this study we present a viewpoint-aware convolutional neural network, which learns to extract specialized features from images of dorsal and ventral sides of barley grains. We show that it increases the average classification accuracy by 0.6% and sensitivity by 2.3% with respect to the viewpoint-ignorant architecture on our dataset.**

I. Introduction

In the case of food products, the quality of ingredients which they are produced from plays a significant role, hence their comprehensive inspection is necessary. The sooner the potential fault can be detected, the lower is the actual production cost including wasted materials. Therefore, effective and quick quality assessment requires automation. Among non-invasive methods, machine vision has a special significance. However, the difficulty in identifying quantitative features and their considerable variability recently draws attention to artificial intelligence methods, especially deep learning.

An example of such a natural product is barley, especially its malting varieties, which is a key ingredient in the production of beer and whiskey. Any deficiencies in its quality immediately affect quality and therefore value of the finished product. Hence, the examination of purchased barley includes detection of impurities or damage as well as moisture content and protein content measurement. Typically, such assessment is performed visually by sampled statistical process control (SPC) as it is technically infeasible to inspect all individual grains in a shipment weighing several tons. This process is however very tedious and, due to the difficulty of the task as well as the human's fatigue, error-prone.

This work was supported from the internal budget of Mechanical Faculty of WUST. The source material (barley grain) for the dataset preparation was supplied from a project financed by National Centre for Research and Development - Project PBS3/A8/38/2015.

The subtle flavors of beer and whiskey are determined, inter alia, by enzymes associated with varieties of barley. However, control of varietal homogeneity without expensive bio-chemical tests is still an unresolved problem. One possible approach, using machine vision methods, seems particularly promising due to its potential speed and no need for direct interaction with the grains. The aforementioned difficulty of feature identification becomes even more challenging in the case of barley grains, because they exhibit different features on dorsal and ventral sides.

In this study, we present a machine vision approach to recognition of barley varieties using convolutional neural networks. We propose a neural network architecture with two feature extraction streams, each specialized to process images of a specific side of the grain. This architecture is complemented with a preprocessing recognition step, in order to identify the dorsoventral orientation of each grain. The paper is arranged as follows: in section II we reintroduce a double-stream convolutional neural network from our previous work, in section III we describe the novel architecture, as well as the dataset and training methods, and in section IV we experimentally evaluate performance of the models and compare them.

II. Related works

There are several known approaches to barley grain varietal recognition. All of them rely on digital image processing and feature extraction. The features are usually hand-engineered, e.g. edges, texture descriptors, and low dimensional or reduced to a low dimension. Most works also employ some form of machine learning to perform recognition. Zapotoczny *et al.* [1] explore possibilities of classifying images of barley kernels using principal component analysis (PCA), and linear or non-linear discriminant analysis (LDA/NDA). Nowakowski *et al.* [2] use extracted features as learning vectors for an artificial neural network (a multilayer perceptron). Hailu and Meshesha [3] present a classifying ensemble of k nearest neighbors and an artificial neural network. Those approaches yield promising results, but they are only tested on very small datasets (up to several hundred images in up to 5 classes). The scope of work by Szczypiński *et al.* [4] is significantly larger, their dataset comprising over 13,000 images of 11 varieties.

In all of those works, feature extraction and classification are considered separate parts of a system, where the features remain fixed and the classifier is designed using machine learning (ML). Recent advancements in ML made it possible to learn the feature extraction function and classification as a single system. Convolutional neural networks (CNNs) can be trained on raw images, without the difficulty of manually designing the feature extractor. Despite their applications to other agriculture-related problems (e.g. [5]), there have been no attempts to classify barley varieties with CNNs.

In our previous paper [6] we have presented a CNN for detection of defects and impurities in barley grain. Our approach made use of the double-sided imaging capacity of the acquisition system presented by [7]. The double-stream CNN was able to extract features from images of both sides of the grain. Then it fused the feature vectors together, creating a single representation from both images. This enabled it to utilize the information contained within both views of the object to predict its class.

However, due to the unpredictable nature of the imaging process, it was never known which side of each grain was actually visible on which image. Therefore the network had to be robust to this unpredictability, effectively discarding the information about dorsoventral orientation of the grains.

III. Experiment Setup

A. Reference neural network architecture

As a reference model we reintroduce the double-stream convolutional neural network from our previous work [6]. This architecture consists of two streams – that is, two separate CNNs – each assigned to a specific camera in order to process one image of each grain. At some point, depending on the setup details, representations produced by those streams are merged into a single stream. Classification is performed by a feed-forward fully-connected (FC) neural network, whose input is this merged representation.

Dorsal and ventral sides of a grain may exhibit different features. However, during the imaging process the dorsoventral orientation of the grains cannot be constrained. Therefore it is not known which side of the grained is imaged by which camera - the cameras cannot be assigned to any particular viewpoint (i.e. dorsal or ventral). In order to provide robustness against this unpredictability, the two feature extraction networks have shared parameters. That means that even though these are two separate networks with different data flowing through them, their parameters are shared, i.e. constrained to always be equal (fig. 1). Since the camera viewpoints are irrelevant in this setting, we term this architecture viewpoint-ignorant.

The actual CNN implementation used in this study is derived from AlexNet [8], comprising 5 convolutional layers of decreasing kernel size (respectively, 11x11, 5x5, 3x3, 3x3 and 3x3) with 3 overlapping pooling operations between them, and 3 fully-connected layers: first two followed by dropout layers [9], the last one by a softmax operation. After each convolutional and FC layer, a ReLU nonlinearity is applied [8]. The FC layers originally consist of 4096 neurons each except the last one, which is scaled depending on the number of classes. In order to reduce overfitting we limit the capacity of the network by replacing those layers with significantly smaller ones: 64 and 16 neurons each.

A double-stream network is constructed by instantiating two copies of each convolutional layer, although the

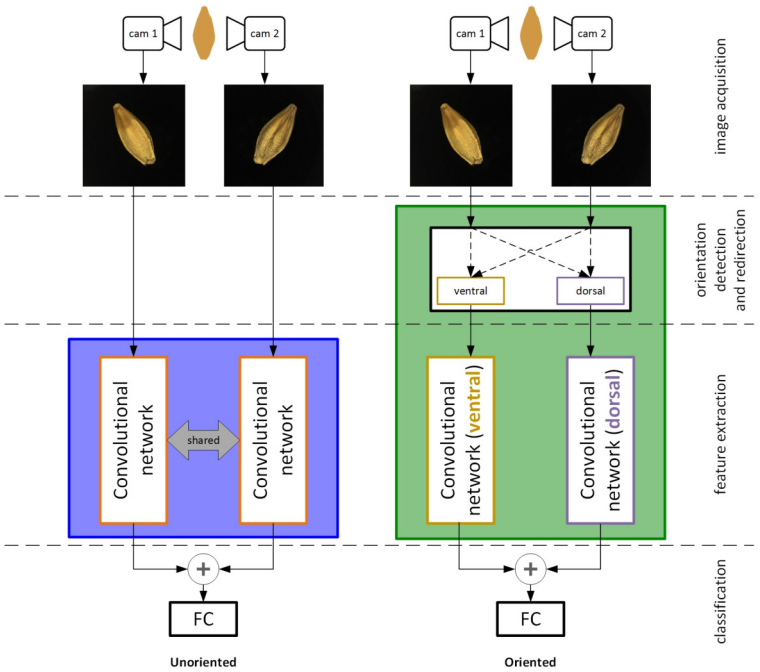


Fig 1. Double-stream CNN architectures: viewpoint-ignorant (on the left) and viewpoint-aware (on the right).

parameters (and gradients thereof) of each pair are constrained to be equal. Each stream will process an image of a different side of the grain, and during training their gradients will be summed. The streams will be merged after the last pooling operation by simple tensor concatenation. The following FC layers will proceed normally, although the input of the first FC layer will be larger to accommodate the concatenated outputs of the streams.

B. Proposed architecture

We propose a network architecture in which the streams are not robust to random dorsoventral orientation of the grains, but instead each learns to extract features specific to each side. In this setting it is assumed that each image the network receives is taken from a particular viewpoint – one of the ventral, another of the dorsal side. We term this network viewpoint-aware.

The novel architecture consists of two streams constructed exactly like in the reference architecture, except for the parameter and gradient equality constraints, which are lifted. This results in two entirely separate feature extraction networks, each associated with a specific view of the object. The rest of the architecture, particularly fusion of the streams by concatenation, and FC layers with softmax, remain structurally identical with the reference architecture.

The assumption that each stream receives a specific view of the object, as opposed to image from a specific camera, requires that the dorsoventral orientation of the object be identified before any processing. In order to achieve this, we introduce a preprocessing step in which the side of a grain is recognized, and the two images are redirected to their associated feature extraction streams according to the result. We solve the task of viewpoint recognition by reducing it to binary classification.

Both the CNN streams and the imaging cameras are ordered, so there is a natural correspondence between them – assume stream 1 is associated with ventral view, then camera 1 can capture either ventral or dorsal side. Therefore, viewpoint recognition becomes a binary classification task: either the images are acquired in the right alignment or not, in which case they need to be switched. For this, we use a CNN of the same structure as the reference architecture with shared streams, except for the final FC layer, which only has 2 outputs (correct alignment or switching needed). The complete system is shown in (fig. 1).

C. Dataset

The data used throughout this research was acquired using a prototype imaging system. The device captured RGB images of individual grains using two cameras located coaxially, opposing to each other, allowing for acquisition of top and bottom views of each grain (details in [7]). Due to the nature of the grain partitioning and transportation subsystem, the dorsoventral orientation of the objects was not predictable. For this reason, no orientation labels could be assigned to the images at the data preparation stage.

Barley was acquired from a research supply. The grains came already separated into 8 varieties, which could be grouped into spring (S) or winter (W), as well as malting (M) and fodder (F) varieties. There were exactly 2 varieties in each of the group combinations (SM, SF, WM, WF).

A total of 3169 pairs of top/bottom images were acquired, ranging between approximately 200 and 500 pairs per variety. During preprocessing, they were cropped so that the grains were visible in the center of the images, and then resized to 256x256. For the purpose of training and cross-validation, the dataset was split into 3 disjoint subsets. For every cross-validation bin, one of those subsets was used as a training set, while the two remaining ones were merged into a validation set. We applied data augmentation on each training set, appending copies of each image rotated 16 times. Table I shows the dataset composition (pre-augmentation).

TABLE I.
DATASET COMPOSITION

Variety	No. training samples	No. validation samples	Percent of total
SM Bordo	140	278	13.2%
SM Kormoran	163	327	15.5%
SF Mercada	133	266	12.6%
SF Skarb	129	257	12.2%
WM Vanessa	116	232	11.0%
WM Vincenta	166	334	15.8%
WF Kobuz	138	274	13.0%
WF Zenek	72	144	6.8%
Total	1057	2112	100%

D. Training procedure

Neural network training was performed using the Caffe framework [10], with Nvidia DIGITS front-end for task management, using a Nvidia GTX TITAN Z graphics processing unit (GPU) with 2 banks of 6 GB VRAM and 2880 CUDA cores each. To reduce the possibility of overfitting the data, the transfer learning technique was used: each of the convolutional layers was initialized from an AlexNet model pre-trained on ImageNet, a dataset of 1.5 million natural images of various origin in 1000 classes (pre-training, performed independently by Jeff Donahue, BVLC, was not a part of this study). The remaining layers' weights were initialized with Gaussian noise of mean 0 and standard deviation of 0.01, while biases were initialized with a constant of 0.1 each.

Networks were trained using multinomial logistic loss function and Nesterov Accelerated Gradient (NAG) optimization method [11], which is a variation of stochastic gradient descent with momentum. Major training hyper-parameters were: momentum $\mu = 0.9$, batch size 128 (as large as could fit in the GPU memory), initial learning rate $\alpha = 0.01$ (as high as the training could still converge at).

For variety recognition, the reference double-stream viewpoint-ignorant network was compared with the proposed viewpoint-aware network. Both networks were trained for 25 epochs: the first 2 epochs at learning rate 0.01, then until epoch 20 at rate 0.001 and for the remaining time at 0.0001. Each process was repeated 3 times for each of the cross-validation folds. Results (F1 measure and confusion

matrices) are reported by averaging of each 3 cross-validation models.

A naïve setup for the viewpoint-aware network would consist of a preprocessing network embedded into the architecture. However, since this sub-network is not being trained at this stage, its presence would only increase the memory and computation power requirement of the entire system, making the training process significantly slower. For this reason, the dorsoventral orientation recognition network was trained separately.

First, a subset of 500 images was selected from the main dataset and annotated manually (with another 500 images selected and annotated for the purpose of validation). The network was trained on this dataset for 20 epochs. After the first 8 epochs learning rate was reduced to 0.001, and after another 8 to 0.0001.

Then, the preprocessing network was queried once over the entire dataset to generate the auxiliary labels containing the information about dorsoventral orientation of each grain. The double-stream viewpoint-aware variety recognition network only read those labels at training time, reducing the preprocessing step only to redirecting images to the feature extraction streams as needed. When using such network in production environment, both the preprocessing sub-network and the variety recognition network would have to be instantiated at the same time.

IV. Results

A. Viewpoint recognition

The dorsoventral orientation recognition network reached 99.8% accuracy after 20 epochs of training. There was little to no overfitting, as both training and validation loss were equal to about -4 in logarithm. The 0.2% error rate was caused by a single image, in which the grain was not imaged from neither the dorsal nor ventral side, but from the sides (fig. 2). This is a rare occurrence which the imaging system allows, but due to an insignificant fraction of such images in the dataset, we decided to ignore their influence – those cases were not handled in any particular way.

B. Variety recognition

Training of a single variety recognition network took approximately 80 minutes (compared to less than 2 minutes for the preprocessing network). All of the models displayed a satisfactory fit – validation loss was actually lower than training, and accuracy was higher (fig. 3). Explanation for this counter-intuitive phenomenon is in dropout. During training, 50% of the fully-connected neurons are randomly deactivated, artificially increasing prediction difficulty, when during validation, no neurons are disabled (their activations are scaled down by a factor of 0.5 to preserve the total magnitude of the activation). This has a significant anti-overfitting effect.

Classification results comparing the viewpoint-ignorant and viewpoint-aware networks, averaged over cross-validation folds, are shown in Table II. Sensitivity and specificity are defined for binary classification, so the table

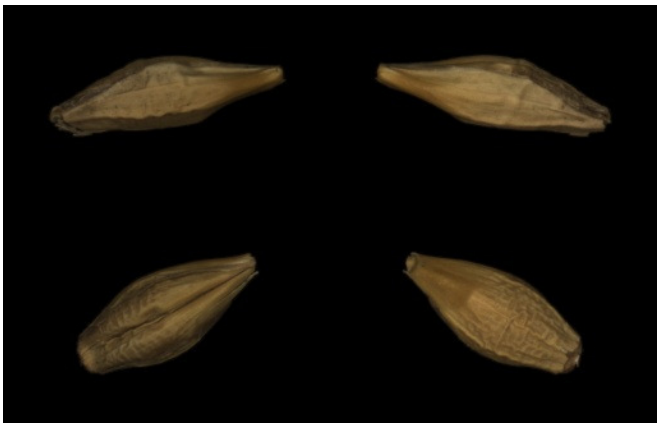


Fig 3. Difficult case of a sidewise grain orientation (top) versus normal grain exhibiting its ventral (bottom left) and dorsal sides (bottom right)

contains averages of values obtained for each class as a one-versus-all classification.

In fig. 4 we compare confusion matrices for viewpoint-ignorant and viewpoint-aware models. The matrices are normalized row-wise, so a percentage on each tile corresponds to a fraction of images from a given row that were recognized as belonging to the given column (true positive ratio on diagonal). In most cases, the TPR is higher for the viewpoint-aware network – most notably for SF Zenek, an increase from 51.9% to 69.7%. With two classes (SM Bordo, WM Vanessa) the viewpoint-aware network performed worse in terms of TPR. However, in those cases the classification precision (ratio of correct predictions to all predictions as this class, interpreted as probability that a prediction is correct) was significantly higher: 92.80% vs 91.92% for Bordo and 81.43% vs 77.60% for Vanessa.

This confirms that the viewpoint-aware approach is more powerful on average, but the scale of the difference would depend on weights assigned to errors of each kind.

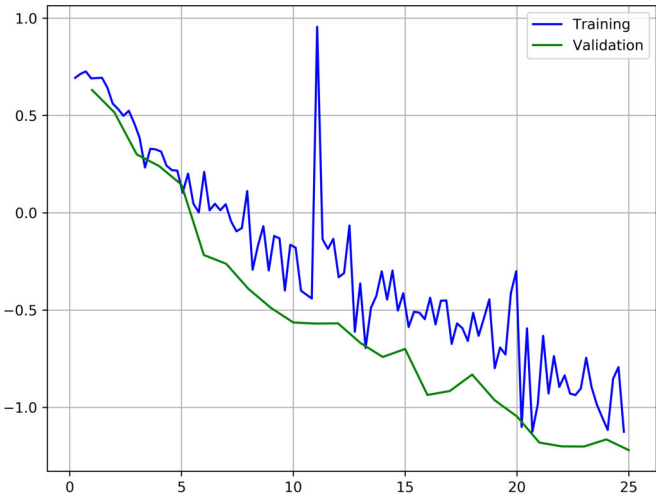


Fig 2. Loss function logarithm on training and validation sets throughout training

TABLE II.
VARIETY CLASSIFICATION RESULTS

Measure	Viewpoint-ignorant	Viewpoint-aware
Accuracy	96.65%	97.24%
Sensitivity	86.63%	88.97%
Specificity	98.09%	98.42%

V. CONCLUSIONS

We have presented a viewpoint-aware double-stream convolutional neural network and proved its superior performance at classification of barley grain varieties. The system performed better on average than a previously shown viewpoint-ignorant network, with slight variations in performance at particular classes.

Those differences might depend on the actual properties of the grains themselves. A detailed study into grain classification would have to account for many such factors, for example phenotypic variability of barley across vegetation seasons.

The system could in principle be trained in an end-to-end setup, if only the dorsal/ventral annotations were available. Due to the image acquisition technique as well as the nature of the imaged objects, obtaining those annotations during data acquisition is not trivial. This is however a limitation of the data imaging system, not our proposed machine learning system.

ACKNOWLEDGMENT

We wish to thank Piotr Lampa and Krzysztof Wall for performing data acquisition.

REFERENCES

[1] P. Zapotoczny, M. Zielinska, and Z. Nita, "Application of image analysis for the varietal classification of barley:," *Journal of Cereal Science*, vol. 48, no. 1, pp. 104–110, Jul. 2008. doi: 10.1016/j.jcs.2007.08.006

[2] K. Nowakowski, P. Boniecki, R. J. Tomczak, S. Kujawa, and B. Raba, "Identification of malting barley varieties using computer image analysis and artificial neural networks," presented at the *Fourth International Conference on Digital Image Processing (ICDIP 2012)*, 2012, vol. 8334, p. 833425. doi: 10.1117/12.954155

[3] B. Hailu and M. Meshesha, "Applying Image Processing for Malt-barley Seed Identification," presented at the *Conference: Ethiopian the 9th ICT Annual Conference 2016 (EICTAC 2016)*, Addis Ababa, 2016.

[4] P. M. Szczypiński, A. Klepaczko, and P. Zapotoczny, "Identifying barley varieties by computer vision," *Computers and Electronics in Agriculture*, vol. 110, pp. 1–8, Jan. 2015. doi: 10.1016/j.compag.2014.09.016

[5] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, "Deep-plant: Plant identification with convolutional neural networks," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 452–456. doi: 10.1109/ICIP.2015.7350839

[6] P. Dolata, M. Mrzygłód, and J. Reiner, "Double-stream Convolutional Neural Networks for Machine Vision Inspection of Natural Products," *Applied Artificial Intelligence*, vol. 31, no. 7–8, pp. 643–659, Sep. 2017. doi: 10.1080/08839514.2018.1428491

[7] P. Lampa, M. Mrzygłód, and J. Reiner, "Methods of manipulation and image acquisition of natural products on the example of cereal grains.," *Control & Cybernetics*, vol. 45, no. 3, 2016.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[10] Y. Jia et al., "Caffe: Convolutional Architecture for Fast Feature Embedding," 2014, pp. 675–678. doi: 10.1145/2647868.2654889

[11] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the Importance of Initialization and Momentum in Deep Learning," in Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, Atlanta, GA, USA, 2013, pp. III–1139–III–1147.

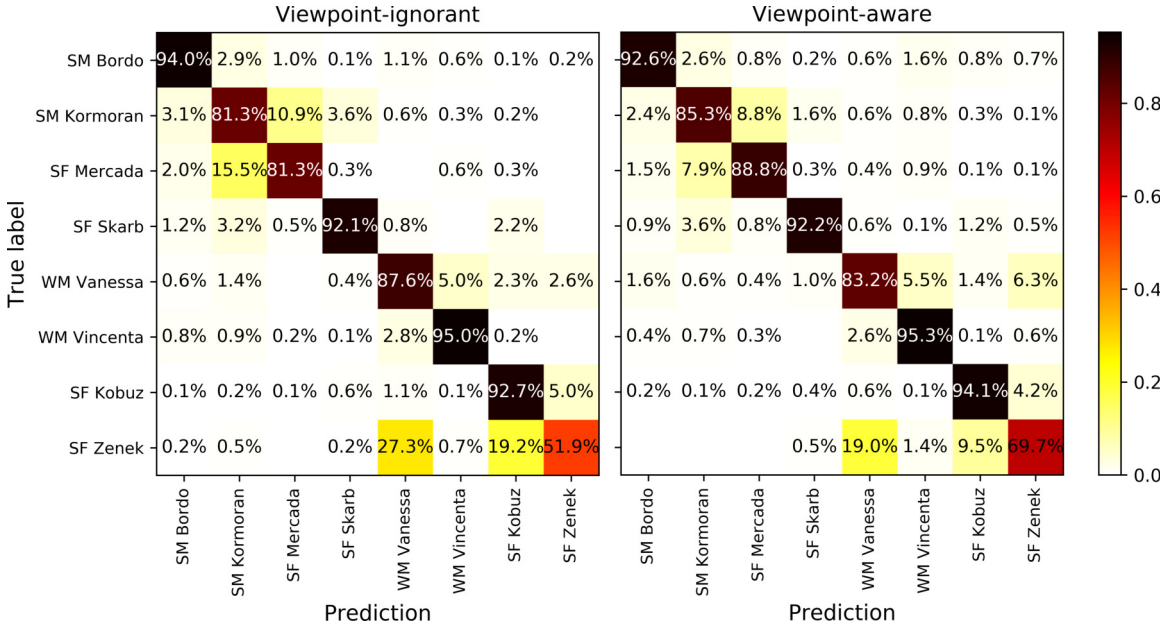


Fig 4. Confusion matrices for viewpoint-ignorant and viewpoint-aware models

8th International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2018—provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogenious field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

TOPICS

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

EVENT CHAIRS

- **Paja, Wiesław,** University of Rzeszów, Poland
- **Pancerz, Krzysztof,** University of Rzeszów, Poland
- **Stoean, Catalin,** University of Craiova, Romania

PROGRAM COMMITTEE

- **Belciug, Smaranda,** University of Craiova, Romania
- **Iantovics, Barna,** Petru Maior University, Romania
- **Lasek, Piotr,** University of Rzeszow, Poland
- **Leniowska, Lucyna,** University of Rzeszow, Poland
- **Lichtblau, Daniel,** Wolfram Research, United States
- **Majernik, Jaroslav,** Pavol Jozef Safarik University in Kosice, Slovakia
- **Mapayi, Temitope,** University of KwaZulu-Natal, Durban, South Africa, South Africa
- **Olszewska, Joanna Isabelle,** University of Gloucestershire, United Kingdom
- **Perner, Petra,** IBaI Leipzig, Germany
- **Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt
- **Stańczyk, Urszula,** Silesian University of Technology, Poland
- **Stoean, Ruxandra,** University of Craiova, Romania
- **Zaitseva, Elena,** University of Zilina, Slovakia

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 109–116
DOI: 10.15439/2018F163
ISSN 2300-5963 ACSIS, Vol. 15

Pitfalls in users' evaluation of algorithms for text-based similarity detection in medical education

Jakub Ščavnický, Matěj Karolyi, Petra Růžičková, Andrea Pokorná,
Hana Harazim, Petr Štourač, Martin Komenda
Faculty of Medicine, Masaryk University, Kamenice 5, 625 00, Czech Republic
Email: {scavnicky, karolyi, ruzickova, pokorna, komenda} @iba.muni.cz, hana.harazim@gmail.com, stourac@med.muni.cz

Abstract—**This paper introduces a user evaluation of several approaches for an automated similarity detection between study materials and curriculum description in the field of medical and healthcare education. Our objective is to present an effective methodology of getting relevant feedback from medical students and teachers. Two various data sets (electronic study materials represented by interactive educational algorithms on the AKUTNE.CZ platform and the curriculum of the General Medicine study programme) are processed. For the purposes of this work, text similarity between two data sets is expressed lexically, i.e. character-based (n-gram) similarity as well as term-based similarity methods are used. We present the comparison of five selected approaches to similarity calculation as well as an objective discussion covering our experience with and pitfalls of user evaluation.**

I. Introduction

Medical and healthcare studies cover a variety of useful information and sources used for learning and teaching leading to professional development. In general, any high-quality education requires that materials guaranteed by experts are available; these materials then constitute the curricula of individual study programmes. In the period between matriculation and graduation, students face a large amount of knowledge and skills to be acquired, which is repetitively emphasised in lectures, seminars and clinical practices. By way of illustration, the General Medicine master's degree programme at the Faculty of Medicine of the Masaryk University contains around 150 obligatory courses which are described by approximately 1,200 events (learning units) and 7,000 competency objects (learning outcomes); in total, this makes up more than 2,500 pages of text. Moreover, each of above-mentioned courses has a set of recommended study materials which are available either in the printed form (scripts/textbooks, atlases, monographs etc.) or in the electronic form (presentations, virtual patients/interactive educational algorithms, educational websites, etc.). With respect to human cognitive abilities, it is virtually impossible to carefully read and remember every single detail of all learning units and book chapters, including their linkages and co-dependencies [1]. This paper picks up the threads of the authors' previously published work, where the development and implementation of modern interactive tools [2], [3], as well as a complex analysis and mapping of medical and healthcare curricula [4]–[7], were introduced. There is also given a proposal of several approaches for an automated similarity detection between study materials and curriculum description in the field of medical education, including the evaluation of achieved results by users. The authors strived to get relevant feedback from medical students and teachers in terms of a systematic and objective evaluation of links between a given virtual patient and particular building blocks (learning units) of the curriculum. The following research questions were formulated in order to define and subsequently solve a particular research problem: What is the relation between the achieved results in a form of detected similarities done by computer and an evaluation by users (medical student and teachers)? Which approach of similarity detection can be effectively implemented in a particular domain of medical education?

II. Methods

A. Input data set

For the purposes of similarity detection between medical education data, we decided to process two various data sets: (i) electronic study materials represented by interactive educational algorithms on the AKUTNE.CZ platform[1] (77 virtual patients described by approximately 550 standard pages of text in total) and (ii) the curriculum of the General Medicine study programme taught at the Faculty of Medicine of the Masaryk University, represented by a full metadata description on the OPTIMED curriculum management system[2] [3] (1,232 learning units described by approximately 2,600 standard pages of text in total). Both input data sets were prepared in English language in order to eliminate problems related to a rather complicated morphology of the Czech language. As for the evaluation by users, we chose a subset of 16 learning units of a course entitled "Diagnostic imaging methods", which provides an introduction to the study of nuclear medicine, more specifically the study of radiology and imaging methods, including CT, MR, X-ray, angiography and ultrasound. All of these units are fully described by all mandatory and optional metadata, covering one complete topic and one complete course in the fourth year of study of the General Medicine. This choice of this particular course was consulted with senior experts in a field of medical education because some general overlapping

This work was not supported by any organization

[1] http://www.akutne.cz/index-en.php

[2] http://opti.med.muni.cz/en/

topics and areas with interactive algorithms were expected here. One of the main motivation given by senior teachers to select this special area was the fact that imaging methods presuppose sufficient image documentation to be used in interactive educational algorithms. Moreover, the quality and length of metadata description of all above-mentioned learning units were sufficient in terms of text-based analysis.

B. Similarity calculation

The similarity of text documents can be understood in two different ways – either semantical or lexical. The former refers to similarity in meaning and used context, whereas the latter represents similarity of character sequences. In this pilot study, we understand text similarity lexically. According to [8], we can classify lexical or string-based text similarity methods into character-based groups and term-based groups. The n-gram method is one of the character-based methods introduced in this pilot study. On the other hand, the term-based methods were implemented using several string measures – the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient. Each of used methods is briefly described in the following paragraphs.

C. Character based (n-gram) similarity

The comparison through a pg_trgm module (a PostgreSQL database engine that can be installed into an existing database using a simple SQL command, namely CREATE EXTENSION IF NOT EXISTS pg_trgm) is one of the approaches we used to calculate the measure of similarity between selected learning units at the OPTIMED portal and interactive educational algorithms at AKUTNE.CZ. This extension of the PostgreSQL standard database provides functions and operators to compare the similarities and distances of input text strings. Generally speaking, the pg_trgm is an n-gram (character-based) algorithm for similarity measurement [8]. In this case, N is equal to three and therefore, the measuring unit is called a trigram. In other words, a trigram is a group of three consecutive characters taken from an input string.

We are able to measure the similarity of two strings by counting the number of shared trigrams (there is a similarity to an ASCII alphanumeric text based on trigram matching). This simple idea turns out to be very effective for measuring the similarity of words in many natural languages (e.g. English) [9]. For example, the set of trigrams in the word "pet" is following: "p", " pe", "pet", "et " (the algorithm takes the input word with two spaces prefixed and one space suffixed). We expect that also in professional terminology, these similarities would be quite easily identifiable.

The pg_trgm module provides four functions and two operators. For our purposes, the function called similarity is the most interesting one, taking two strings to be compared. The function *similarity(text, text)* returns a number between 0 and 1 which indicates how similar the two inputs strings are: zero means that the two strings are completely different, whereas one indicates that the strings are completely identical. In the next step, the operator *text* <-> *text* returns the distance between two strings; it is defined as one minus the similarity of strings.

We computed all possible combinations of learning units and virtual patients/interactive algorithms using the similarity functions and stored the result in a database table (see Table 1) for further analysis and comparison with other algorithms. Similarity column represents computed trgm similarity between a learning unit and an algorithm. Correctness in interpretation of similarity results depends on our experts' expectations. If the two subjects are similar, we want to measure high similarity value.

D. Term-based similarity

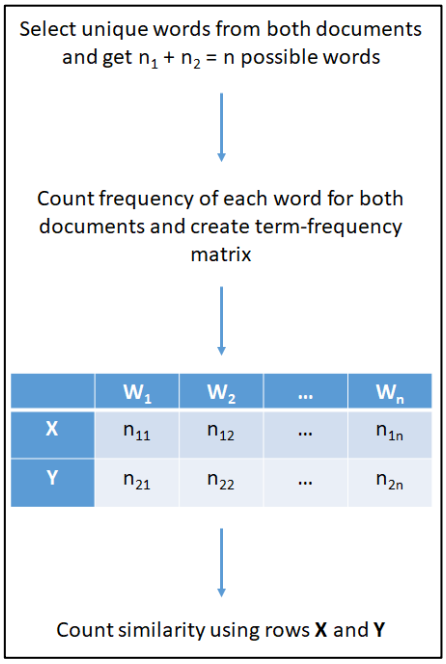
Term-based similarity approaches require that a similarity measure is chosen. A similarity measure quantifies the similarity between two numeric vectors of the same length. When using this approach, text documents are represented as bags of words, and a term-frequency matrix containing the counts of a word occurrence is computed. Figure 1 represents the process of computing text similarity between two documents.

As mentioned above, four similarity measures (the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient) were used to compute similarities between virtual patients/interactive algorithms from the AKUTNE.CZ portal and learning units form the OPTIMED platform. Each similarity measure is computed in a slightly different way and might have a correspondingly different interpretation. The normalised Pearson correlation is a centred correlation similarity measure. The cosine similarity is a measure of similarity between two vectors of an inner product space that

TABLE I.
EXAMPLE OF DATABASE TABLE INCLUDING SIMILARITIES.

id_learning_unit	id_algorithm	title_learning_unit	title_algorithm	similarity
890	77	Protection against radiation, principle of skiagraphy and skiascopy	Car accident	0.3278
891	77	Principle of computer tomography (CT), magnetic resonance (MR) and ultrasound, new horizons	Car accident	0.1798
894	77	Abdominal radiology	Car accident	0.2433
895	77	Uroradiology	Car accident	0.1857

measures the cosine of angle between them. The extended Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both vectors. And finally, the Dice coefficient is defined as twice the number of common terms in the compared vectors divided by the total number of terms in both vectors [8], [10]. These similarity measures lie between 0 and 1. Zero means that two vectors are completely different, whereas one indicates that they are completely identical.



	W_1	W_2	...	W_n
X	n_{11}	n_{12}	...	n_{1n}
Y	n_{21}	n_{22}	...	n_{2n}

Fig. 1 The process of similarity computing.

In our work, all term-based computations were conducted in the R software using the *dplyr*, *tm* and *proxy* packages. First of all, documents from both sources were preprocessed, i.e. HTML tags, punctuation, digits, other special characters and words shorter than three characters were systematically removed. Afterwards, specific words using both a Google stop-word list and a customised stop-word list were removed. In the second step, word occurrence frequencies for each single document were counted. And thirdly, similarities between individual frequency vectors of documents (virtual patients/interactive algorithms) from AKUTNE.CZ and OPTIMED educational texts/learning units were computed using all of the selected similarity measures.

E. Online evaluation tool

A web-based tool has been designed and developed in order to obtain an effective evaluation of achieved results (in a form of detected similarities) by users (medical teachers and students). Using this tool, objective opinions critically assessing the relevance between virtual patients/interactive algorithms and a particular learning unit can be systematically organised and processed. The evaluation module is integrated into the OPTIMED curriculum management system and offers the possibility to view the underlying data from both systems including an easy collection of the users' evaluation via an online form (see Fig. 2). A group of twelve evaluators (fifth and sixth year medical students, young and senior teachers) was involved for the purposes of our pilot evaluation. First of all, they needed to get acquainted with the name of the learning unit and with its brief description. Furthermore, they were invited to view the completed content of an evaluated learning unit, which was available via a direct link to the OPTIMED platform. Afterwards, the users started to evaluate the relevance of available virtual patients/interactive algorithms. Each individual Akutne.cz interactive algorithm was described by a title, a short description and keywords. The users' opinions were expressed using a marking system (grading scale) similar to that used in schools (i.e. the Likert scale from 1 to 5), where 1 meant that the interactive algorithm was very relevant to the learning unit and 5 meant that the interactive algorithm was not relevant to the learning unit at all.

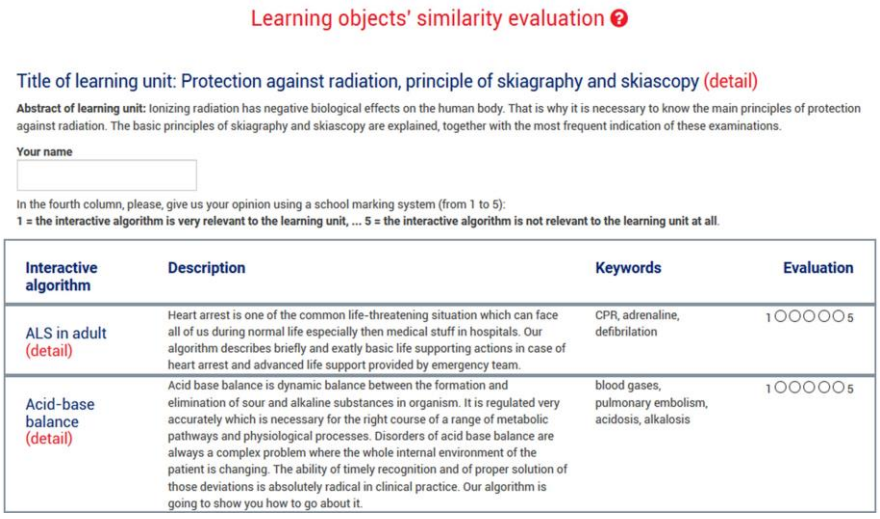


Fig. 2 Online form allowing evaluation of the teaching materials (relevancy of Akutne.cz interactive algorithms to a particular learning unit).

III. RESULTS

A. Overview of calculated similarities

The general overview (see Fig. 3 and Table 2) shows the comparison of five chosen approaches to similarity calculation in the form of a box plot chart, where the similarity measurements between all 77 interactive algorithms and 1232 learning units were taken into account. From our point of view, it is obvious that the pg_trgm module and its similarity function are very useful for cases where we expect to determine whether or not the original document and its copy are modified. It will very precisely and quickly find out whether there are differences in documents or whether the documents are identical.

On the other hand, this approach is not very appropriate for the comparison of two completely different documents, especially because it is dependent on the volume of the documents' content. Therefore, for the purposes of our pilot study, pg_trgm has been eliminated and the attention was only paid to four similarity measures (the normalised Pearson correlation, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient), as described above.

After an in-depth analysis of the results, we discovered that in many cases, the cosine similarity, the extended Jaccard coefficient and the Dice coefficient indicated zero similarity because particular pairs of documents had no words in common. Nevertheless, the normalised Pearson correlation coefficient returned a non-zero value. That might be due to the fact that the correlation is a coefficient of linear dependency of two vectors. For example, let us compare a short text document (namely „What similarity measure value do we measure") with another short text document (namely „if correlation is used?)". The computed frequency vectors have the following form:

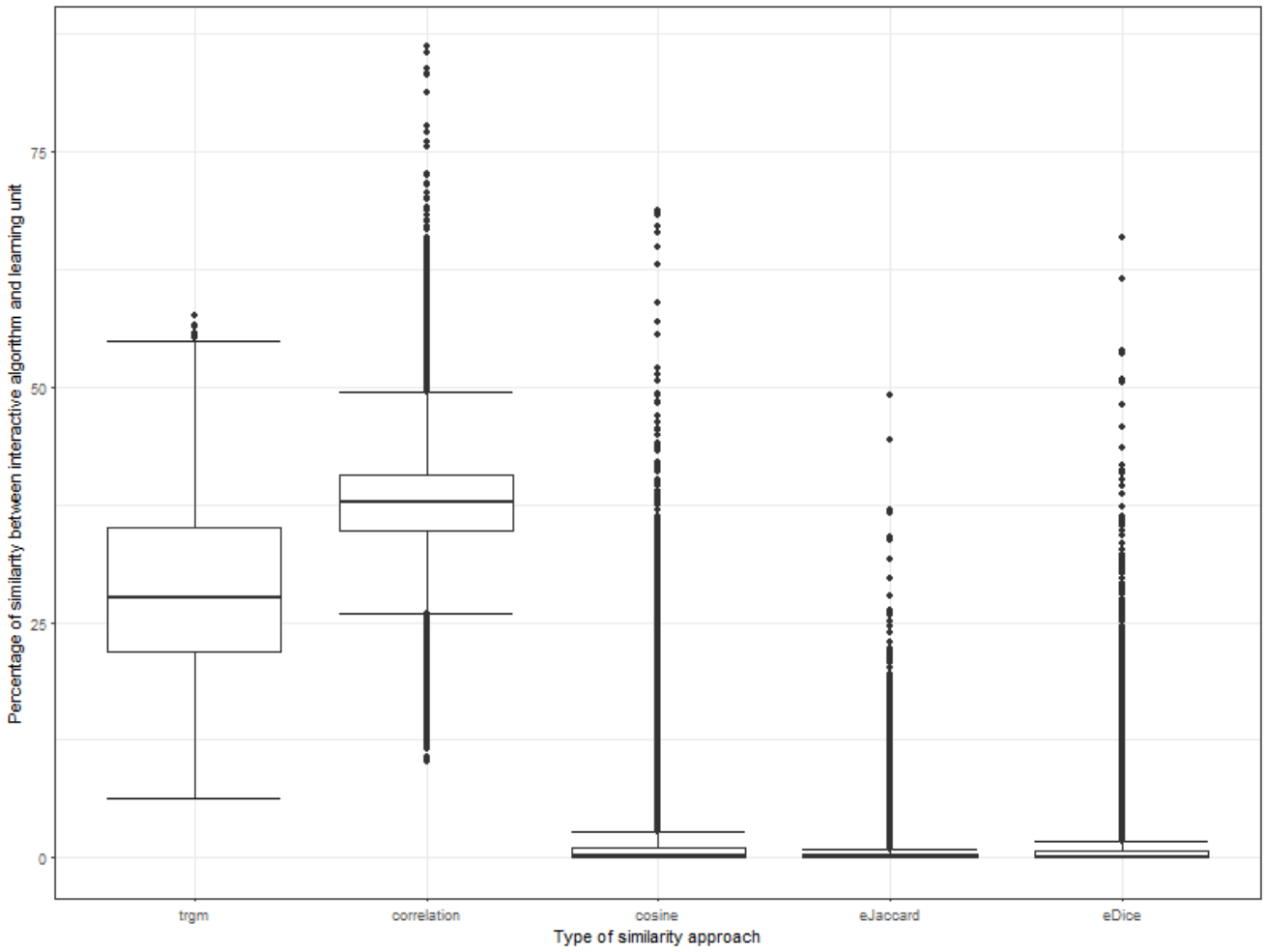


Fig. 3 Box plot representing five approaches for similarity calculation (minimum, maximum, median, average, upper and lower quartiles).

```
terms         freqs.x   freqs.y
correlation   0         1
measure       2         0
similarity    1         0
used          0         1
value         1         0
what          1         0
```

and the corresponding values of similarity measure are as follows:

```
correlation   0.07125354
cosine        0.00000000
eJaccard      0.00000000
Dice          0.00000000
```

This result shows that the correlation takes into account the whole vectors and its value depends on the vector structure rather than just intersection positions. Considering the above-described issue, we assume that the normalised Pearson correlation is not suitable for our text comparisons as out lexical understanding of term frequency in documents.
Therefore, all of our following analyses are conducted on cosine, extended Jaccard and Dice similarity measure outputs.

Figure 4 shows cosine, extended Jaccard and Dice similarity measures based on the normalised Pearson correlation coefficient, which calculates the linear correlation between two variables. The values between measures (0.923. 0.936, 0.996) imply that a linear equation describes the relationship between these measures perfectly, i.e. high positive correlation. Generally, cosine tends to return the highest values, whereas extended Jaccard tends to return the lowest ones. Nevertheless, all three measures provided very similar results.

B. Overview of calculated similarities

In terms of the pilot evaluation of achieved results (calculated similarity measures using various approaches), a set of learning units describing a complete course entitled "Diagnostic imaging methods" were used. Our users (medical students and teachers) used an online form to evaluate the relevance between learning units (OPTIMED) and interactive algorithms (AKUTNE.cz). Figure 5 represents the comparison of similarities (based on the normalised Pearson correlation coefficient) between three measures and the evaluation by users. It is immediately obvious that there is no linear correlation between any similarity measure and the evaluation by users. All algorithms used in a term-based process of similarity calculation provide very similar results, but the user evaluation of content similarity indicates no relationship or dependency between them.

TABLE II.
EXAMPLE OF SIMILARITY SUMMARY TABLE.

Approach	Minimum (%)	Maximum (%)	Median (%)	Average (%)	Upper quartile (%)	Lower quartile (%)
trgm	6.28	57.61	27.7	28.74	35.15	21.93
correlation	10.24	86.15	37.8	37.44	40.62	34.72
cosine	0	68.78	0.19	1.11	1.11	0
extended Jaccard	0	49.22	0.05	0.38	0.34	0
Dice	0	65.97	0.1	0.74	0.68	0
trgm	6.28	57.61	27.7	28.74	35.15	21.93

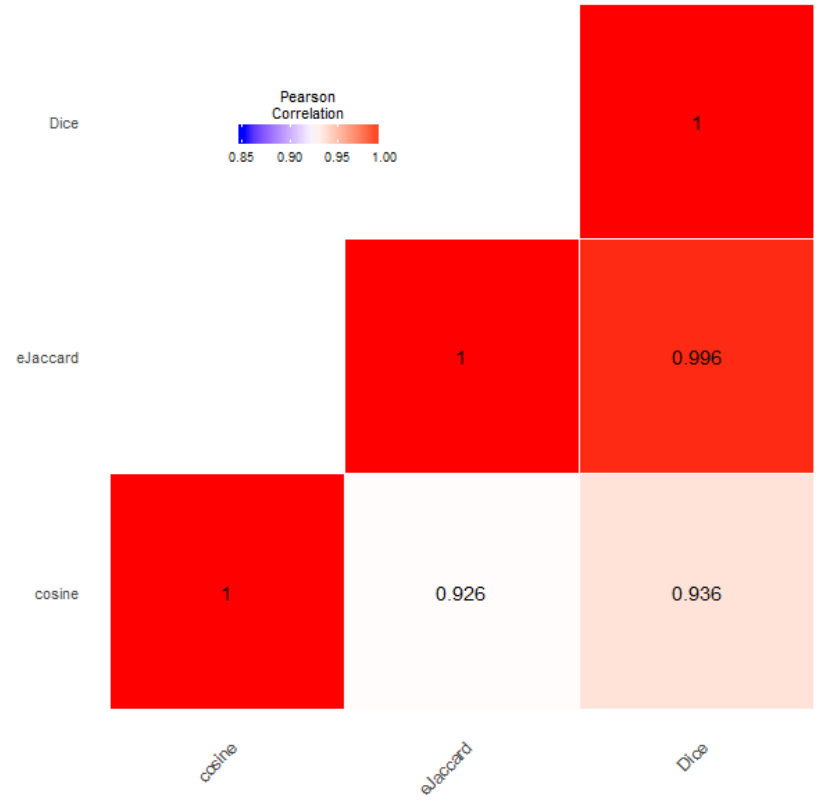


Fig. 4. Comparison of three chosen similarity measures.

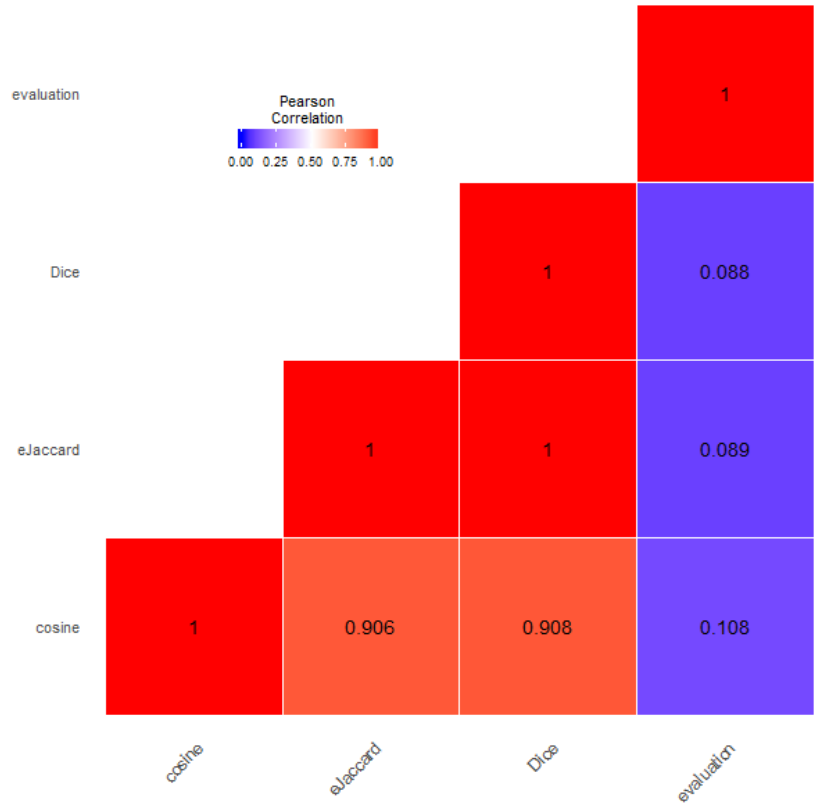


Fig. 5 Comparison of three chosen similarity measures together with users' evaluation.

IV. DISCUSSION

Based on statistical calculations, similarities and common features / terms were found between interactive algorithms in the AKUTNE.cz platform and learning units in the OPTIMED platform. Contrary to our expectations, this study did not find significant similarity between our calculated results of cosine, extended Jaccard and Dice similarity and ratings of users (medical students and teachers). We concluded that a number of factors play a role in determining the results.

Perhaps the most serious limitation of this analysis is an inappropriate choice of the course "Diagnostic imaging methods" with its very specific content and the subsequent search for similarities with a set of interactive educational algorithms in the AKUTNE.CZ platform, which were designed as teaching aids for other courses of the Medical Faculty curriculum: First Aid, Intensive Medicine, Anesthesiology and Pain Management. The fact that X-ray and other imaging methods were used in almost every algorithm only demonstrates that these are frequently used imaging techniques in acute medicine, not that X-ray should be the learning outcome of individual algorithms.

Another important confounding factor is the keywords selection. We believe that mechanically chosen keywords lead to an overinterpretation of search results; the most frequent ones do not represent the most important words, i.e. the keywords. It became obvious that keywords chosen by the machine might have not agreed and in some cases really disagreed with the algorithm's keywords. Undoubtedly, if keywords of the algorithms as defined by the creators had been used, such similarity would have not occurred and it would have increased the accuracy. Moreover, there is definitely space for a systematic improvement of a customised stop-word list. We will need to eliminate terms that do not bring the required information value.

Human evaluators may have contributed to misleading results rather significantly in several ways. Some evaluators had been involved in the design of interactive educational algorithms, which inevitably led to a bias. Some evaluators might have provided an incorrect evaluation due to a misapprehension and/or an unclear assignment. It is important to point out that human evaluators tended to focus on similarities in the meaning of concepts and terms, unlike the machine-based and statistical evaluation of similarities. One improvement to be possibly considered in future might be an optimisation of the evaluation process itself, which should be focused and implemented as a two-stage analysis in the follow-up to this pilot study. First of all, appraisers/evaluators/users would identify similarities according to established keywords, followed by their own analysis of content (abstracts and then full texts). From the methodological point of view, this process would be adopted from the process of assessing professional resources in literature reviews [11], [12]. There is also the possibility to carry out the evaluation as a three-stage process (the third stage would be a peer discussion among evaluators), but it is clear that such a process would be very time-consuming.

Yet another challenge lies in the subjective rating of significance for evaluators and users, which stems from reasoning of both practical and scholarly significance of the teaching problem as well as the scope or the respective teaching topic and issue. In other words, students' and teachers' views could differ when evaluating the learning units and interactive algorithms. Furthermore, the specialised orientation of evaluators could be the explanation for results achieved from their qualitative evaluation: most of our evaluators/users in the pilot study were professionals most familiar with acute care. In their daily practice, they are much more focused on acute and rescue interventions with the goal of saving lives rather than focusing on examination methods, especially not on radiology and imaging methods. Another possible explanation of our findings from qualitative evaluations is that users / evaluators could not see a clear link between the two assessed contents of study materials and interactive algorithms, and their views were reflected in the evaluation. What should be highlighted is that even this finding could help us improve future development of interactive algorithms: there is more space for visual documentation of a clinical condition in intensive care because there is strong evidence that imaging documentation is helpful in the education of healthcare professionals [13]–[15].

We must also emphasise that an important role was played by the fact that the volume of evaluated study materials and interactive algorithms was relatively large (n = 77 virtual patients/interactive algorithms) and that all evaluators carried out their evaluations independently, without the opportunity to communicate with others.

Despite the fact that inconsistencies were identified in our "quantitative/mathematical" and "qualitative – user view" evaluation, we are still convinced that the chosen procedure was appropriate to the above-mentioned set of objectives. We have repeatedly verified that an automated statistical evaluation must always be accompanied by an expert judgment and by an evaluation provided by target users of teaching materials [1]. At least we have verified that our methodology can reveal potential gaps as well as new possibilities of linking study materials to improve the learning process and to increase the students' preparedness for clinical practice. In the follow-up work, we would also like to approach other specialists who might provide their feedback as evaluators; as we have already mentioned, the feedback in this case was mostly provided by intensive care specialists and anaesthetists.

ACKNOWLEDGMENT

The authors were supported from the following grant projects: (i) MERGER project – Reg. No. MUNI/A/1339/2016 funded from the Grant Agency of the Masaryk University; (ii) Masaryk University Strategic Investments in Education SIMU+

(CZ.02.2.67/0.0/0.0/16_016/0002416) funded from the European Regional Development Fund; (iii) Masaryk University 4.0 (CZ.02.2.67/0.0/0.0/16_015/0002418) funded from the European Social Fund. We are also thankful to the team of medical students and teachers, who evaluated achieved results, namely Tereza Prokopová, Daniel Barvík, Václav Vafek, Tereza Ondráčková, Jiří Libra, Matěj Anton, Lucia Macková, Klára Vataha and Martina Žižlavská.

References

[1] M. Komenda, M. Karolyi, R. Vyškovský, K. Ježová, and J. Ščavnický, 'Towards a keyword extraction in medical and healthcare education', in 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), 2017, pp. 173–176.

[2] D. Schwarz et al., 'Interactive Algorithms for Teaching and Learning Acute Medicine in the Network of Medical Faculties MEFANET', J. Med. Internet Res., vol. 15, no. 7, Jul. 2013.

[3] M. Komenda, D. Schwarz, C. Vaitsis, N. Zary, J. Štěrba, and L. Dušek, 'OPTIMED Platform: Curriculum Harmonisation System for Medical and Healthcare Education', Stud. Health Technol. Inform., vol. 210, pp. 511–515, 2015.

[4] M. Komenda et al., 'Curriculum Mapping with Academic Analytics in Medical and Healthcare Education', PloS One, vol. 10, no. 12, 2015.

[5] M. Víta, M. Komenda, and A. Pokorná, 'Exploring Medical Curricula Using Social Network Analysis Methods', Jul. 2015.

[6] M. Karolyi, M. Komenda, R. Janoušová, M. Víta, and D. Schwarz, 'Finding overlapping terms in medical and health care curriculum using text mining methods: reha', MEFANET J., vol. 4, no. 2, pp. 71–77, Jan. 2017.

[7] R. Randell, R. Cornet, and C. McCowan, Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press, 2017.

[8] W. H. Gomaa and A. A. Fahmy, 'A Survey of Text Similarity Approaches', Int. J. Comput. Appl., vol. 68, no. 13, pp. 13–18, Apr. 2013.

[9] 'PostgreSQL: Documentation: 9.1: pg_trgm'. [Online]. Available: https://www.postgresql.org/docs/9.1/static/pgtrgm.html. [Accessed: 15-May-2018].

[10] H. Liu, J. He, D. Zhu, C. X. Ling, and X. Du, 'Measuring Similarity Based on Link Information: A Comparative Study', IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2823–2840, Dec. 2013.

[11] M. Newman and D. Elbourne, 'Improving the Usability of Educational Research: Guidelines for the REPOrting of Primary Empirical Research Studies in Education (The REPOSE Guidelines)', Eval. Res. Educ., vol. 18, no. 4, pp. 201–212, Nov. 2004.

[12] J. P. Gall, M. D. Gall, and W. R. Borg, Applying educational research: A practical guide. Longman Publishing Group, 1999.

[13] B. F. Branstetter, L. E. Faix, A. L. Humphrey, and J. B. Schumann, 'Preclinical Medical Student Training in Radiology: The Effect of Early Exposure', Am. J. Roentgenol., vol. 188, no. 1, pp. W9–W14, Jan. 2007.

[14] R. B. Gunderman, A. R. Siddiqui, D. E. Heitkamp, and H. D. Kipfer, 'The Vital Role of Radiology in the Medical School Curriculum', Am. J. Roentgenol., vol. 180, no. 5, pp. 1239–1242, May 2003.

[15] K. Soyebi, 'Changing students' performance in and perception of radiology', Med. Educ., vol. 42, no. 5, pp. 522–522, May 2008.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 117–123
DOI: 10.15439/2018F127
ISSN 2300-5963 ACSIS, Vol. 15

Retinal Blood Vessel Segmentation Based on Multi-Scale Deep Learning

Ming Li
Information Science and Technology College,
Dalian Maritime University
Dalian, 116026, China
Email: limingdlmu@outlook.com

Qingbo Yin, Mingyu Lu
Information Science and Technology College,
Dalian Maritime University
Dalian, 116026, China,
Email: {qingbo, lumingyu}@dlmu.edu.com

***Abstract*—Fundus images are one of the main methods for diagnosing eye diseases in modern medicine. The vascular segmentation of fundus images is an essential step in quantitative disease analysis. Based on the previous studies, we found that the category imbalance is one of the main reasons that restrict the improvement of segmentation accuracy. This paper presents a new method for supervised retinal vessel segmentation that can effectively solve the above problems. In recent years, it is a popular method that using deep learning to solve retinal vessel segmentation. We have improved the loss function for deep learning in order to better handle category imbalances. By using a multi-scale convolutional neural network structure and label processing approach, our results have reached the most advanced level. Our approach is a meaningful attempt to improve blood vessel segmentation and further improve the diagnostic level of eye diseases.**

I. Introduction

RETINAL fundus images have been widely used for diagnosis, screening and treatment of cardiovascular and ophthalmologic diseases[1], including age-related macular degeneration(AMD), diabetic retinopathy(DR), glaucoma, hypertension, arteriosclerosis and choroidal neovascularization, among which AMD and DR have been considered as two leading causes of blindness[2]. Vessel segmentation is a basic step for the quantitative analysis of retinal fundus images[3]. The segmented vascular tree can be used to extract the morphological attributes of blood vessels, such as length, width, branching and angles.

Moreover, the vascular tree has been adopted in multimodal retinal image registration [4]and retinal mosaic [5]as the most stable feature in the images. In [6], the vascular tree is also used for biometric identification due to its uniqueness. Manual segmentation of the vascular tree in retinal images is a tedious task that requires experience and skill. In the development of a computer-assisted diagnostic system for ophthalmic disorders, automatic segmentation of retinal vessels has been accepted as a vital and challenging step. The size, shape and intensity level of retinal vessels can vary hugely in different local areas. The width of a vessel often ranges from 1 to 20 pixels, depending on both the anatomical width of the vessel and the image resolution. The existence of vessel crossing, branching and centerline re?ex makes it difficult to segment the vessels accurately using artificially designed features. Pathologies in the form of lesions and exudates can further complicate the automatic segmentation. In the past decades, several methods have been proposed for the segmentation of vessels in retinal images, and they can be divided into two categories: unsupervised and supervised methods.

Both classic one-stage object detection methods, like boosted detectors [5]and DPMs(Deformable Parts Model) , and more recent methods, like SSD(Single Shot Multi-Box Detector) , face a large problem of class imbalance during training. These detectors evaluate huge candidate locations per image but only a few locations contain objects. This imbalance causes two problems: (1) training is inef?cient as most locations are easy negatives that contribute no useful learning signal; (2) Simultaneously, the negatives can overwhelm training and lead to degenerate models. A common solution is to perform some form of hard negative mining [6] that samples hard examples during training or more complex sampling/reweighing schemes[7]. In contrast, we show that our proposed focal loss naturally handles the class imbalance faced by a one-stage detector and allows us to ef?ciently train on all examples without sampling and without easy negatives overwhelming the loss and computed gradients.

This paper presents a segmentation method that is suitable for class imbalance. This paper proposes a multi-scale convolutional neural network and improves the traditional loss function, and improves the class labels. Our method outweighs the most advanced methods reported in terms of sensitivity, specificity and accuracy. (1) The proposed method solves the problem of class imbalance in the traditional segmentation method, so that deep learning can better handle the task of image segmentation of the fundus. (2) A series of procedures proposed in the article can be used not only in segmentation tasks but also in various task types such as packet classification detection, and have a wide range of versatility.

II. THE PROPOSED METHOD

A. Motivation

We carefully combed the work of related work and found that most of the previous methods used to perform the two-class task were not evenly sampled. We have found that class imbalances lead to submerged samples, which is usually not

what we want. We explore a solution to the problem from three parts: loss function, network structure and category labels.

The others have done much work in designing robust loss functions (e.g., Huber loss) that reduce the contribution of outliers by down-weighting the loss of examples with large errors (hard examples). So our focal loss functions is designed to address class imbalance by down-weighting inliers (easy examples) such that their contribution to the total loss is small even if their number is large, rather than outliers. In other words, the focal loss performs the opposite role of a robust loss: it just trains on a sparse set of hard examples.

The multi-scale network structure is mainly composed of two deep convolutional network stacks, according to Jarrett . Thus a good multi-level network model to achieve effective target recognition is an important part in our work. The network model in this paper is shown in Figure 1. However, the traditional multi-scale method scales the image segments to different scales and cannot express the boundary of the target region accurately. Thus the method in this paper is based on the selection of each pixel in the image as the center. We can see in Figure 1 two different scale image segments generate input of two deep convolutional networks. The low-level feature extraction network structure first extracts low-level feature information of large-scale image segments, and then uses local refined network structure to capture local region feature information from small-scale image segments. To compare with the traditional multi-scale method, the difference is not that multiple scales separately extract features of different size images. In this paper, the low-level features are combined with the image features extracted from the first layer of the local refinement network. After subsequent network operations, a dense and complete feature vector is obtained, which greatly improves the accuracy of image pixel category prediction.

B. Loss function improvement

The usual method in the objective function of network optimization is the cross-entropy loss function, as followed:

$$E = -\sum_{x} p(x) log(q(x)) \tag{1}$$

Where $p(x)$ is the true distribution of the sample and $q(x)$ is the estimated probability obtained through training. When the cross-entropy loss function is used for a two-category task, its form is:

$$E = \begin{cases} -loh(p), & \text{if } y = 1 \\ -log(1-p), & \text{otherwise} \end{cases} \tag{2}$$

The terrestrial truth category is specified in $y \in \{\pm 1\}$ above and $p \in [0, 1]$ is the estimated probability of the model for the category of label $y = 1$ For symbol convenience, we define p :

$$p = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases} \tag{3}$$

However, in the category imbalance problem, sometimes the difference between positive and negative sample ratios is very different. Taking a fundus image as an example, blood vessel pixels in a single image are only one-fifth that of non-vascular pixels. This imbalance leads to two kinds of situations: (1) Excessive samples of negative examples cause the information of positive examples to be difficult to be effectively learned and even concealed. (2) Simple negative factors lead to training shifts and make the model degenerate. Therefore, we propose an improvement to the cross-entropy loss function that can solve this kind imbalance problem. More crucially, this improvement can make the loss function able to calculate the difficulty in judging every sample, then we can give higher weight to those samples that are more difficult to distinguish. We call the new loss function LCE (Le Cross Entropy), LCE is defined as:

$$E = -\alpha * cos(\beta * p) * log(p) \tag{4}$$

$$\alpha = \frac{\alpha^*}{\beta} \tag{5}$$

Among them, the value ofβ is $\frac{\pi}{2}$. The reason for the value of β is that the range of p is $(0, 1)$, so that $\beta * p$can have the same mapping range as p . The weight α is the balance coefficient, usually $\alpha \in [0, 1]$. This means that the real balance coefficient is actually α^* , but for the sake of clearly, we uses α to describe it. When using LCE as a loss function, we noticed that it has the following two characteristics: (1) When an example is misclassified and p is small, the loss will not be affected. When p is larger, then α goes to zero, and the loss of well-classified examples is reduced. (2) The balance factor ęÁ effectively adjusts the weight of the instance. With the change of ęÁ, LCE can adapt to different degrees of class imbalance. In our experiment, the best results are obtained when ęÁ is taken as 0.25. About the first feature, we take a specific example to illustrate. In the case of $\alpha = 0.2$, and $p = 0.9$, the results of LCE is 33.3 times lower than CE, and the value of loss function when $p \simeq 0.2$ is 50 times lower. From the data shown in the above examples, we can find that correcting misclassification examples is necessary.

C. The Proposed Architecture

In this study, we designed a multi-scale convolutional neural network structure to segment blood vessels from fundus images. In the figure 1, we will show the details of this network in this work. The network consists of two consecutive convolution structures with input sizes of 13 * 13 and 17 * 17 respectively. The convolution kernel we have chosen on each convolutional layer is 5*5 in size and each time we move kernel one pixel. Networks of different scales have a similar convolutional layer structure. The first convolutional layer has 64 feature plots and the second convolutional layer has 128 feature plots, and the third convolutional layer has 256 feature plots. After each convolutional layer, we use a rectifying linear unit (ReLU) excitation as an activation function. As shown in this article[8], using ReLU as an activation function for the convolutional layer can speed up the training of the network. After the features are extracted from the convolution layer, we add the feature maps and connect them with the fully

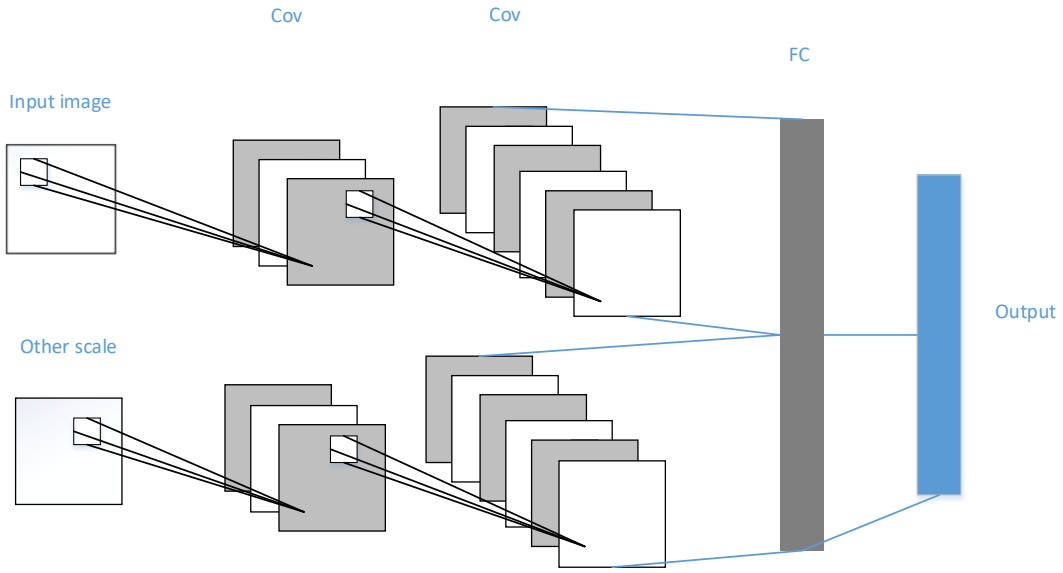


Fig. 1. The Structure of Multi-Scale Convolutional Neural Network

connected layer. It should be noted that the full-connected layer is not only one layer. Only one layer is drawn on the graph for the sake of simplicity of the view. In fact, there are three full-connected layers. We used dropout learn more powerful skills and reduce overfitting. The technique sets the output of each neuron to zero with a probability of 0.5. Finally, we use the softmax function to classify and use our improved cross-entropy function as an energy function.

D. Label Processing

In deep learning, the handling of sample tags is a part that is easily overlooked, especially in binary tasks. The tag y_- is simply designed as $y_- \in \{\pm 1\}$ or $y_- \in (0,1)$ .However, in the traditional machine method, the processing of the label has been proved to be an effective method to improve the accuracy of the classifier, and this method does not have any additional requirements for the computing power. Therefore, we try to introduce sparse variables into the label design of deep convolutional neural networks. The new tag y^* is defined as:

$$y^* = \alpha * y_- + \varepsilon \tag{6}$$

Where ęÁ is called the proportional coefficient, and its size depends on the proportion of the positive and negative samples in the training set. This coefficient was specially designed to solve the imbalance problem. ęÅ is called a sparse variable. It is to bias the same kind of sample. We usually convert segmentation tasks into classification tasks, and convolutional neural networks as a classifier, its role can be simplified to find the optimal classification plane between different types of samples. By adding sparse variables to the category labels, increasing the distance between the classes' classification planes has been used in traditional machine learning. We refer to this approach to deep convolutional neural networks.

TABLE I
PERFORMANCE OF ε-DRAGGING ON DATA POINT IN TWO CLASSES

	class	y	y after ε-Dragging	
x1	1	[1,0]	$[1+\varepsilon_{11}, -\varepsilon_{12}]$	$\varepsilon_{11}, \varepsilon_{12} > 0$
x2	1	[1,0]	$[1+\varepsilon_{21}, -\varepsilon_{22}]$	$\varepsilon_{21}, \varepsilon_{22} > 0$
x3	2	[0,1]	$[-\varepsilon_{31}, 1+\varepsilon_{32}]$	$\varepsilon_{31}, \varepsilon_{32} > 0$
x4	2	[0.1]	$[-\varepsilon_{41}, 1+\varepsilon_{42}]$	$\varepsilon_{41}, \varepsilon_{42} > 0$

Table 1 further explains our method, which reports the four data points in the two categories. Their class label vector is listed in the third column. Now, if we group together the first element of the class labeling vector, we get "1,1,0,0". In this way, a binary class partition can be obtained in which the first two is divided into one class, and the latter two classified data points are classified into another class. After label processing is performed, their image will change from "1,1,0,0" to"$1+\varepsilon_{11}$,$1+\varepsilon_{12}$,ε_{13},ε_{14}" science all are non-negative, this processing can help expand the distance between classes after data point mapping.

E. Model training

Our method does not require image preprocessing including image enhancement, which greatly simplifies the difficulty of segmentation tasks and improves the versatility of the method. According to the mask image, the fundus image part of the original image is centered on each pixel, and we can get the classification of the preset scale, and after we delete the sample of the edge part sample beyond the mask range, remain about 150 samples. Millions of sample drawings. One million of

them are training sets and 500,000 are test sets. However, our method does not use all the training set samples for training. We used a special selection method to obtain samples that was only one-tenth of the original training set. This greatly speeds up the training time without losing too much classification accuracy.

We trained and tested the network on an Intel core computer and implemented it using Anconda+TensorFlow. The processor we used was the Intel Xeon(R) CPU E5-2680V3. The training on these lasted 18 hours. Between the limits of our experimental equipment, we chose a batch size of 64.During training, the weights were updated by stochastic gradient descent algorithm with a momentum of 0.9 and a weight decay of .The biases in convolutional layers and fully-connected layer were initialized to 1. The number of epochs was tuned on a validation set consisting of patches from one randomly selected subject in the training set. The learning rate was set to initially.

III. Result

A. Implementation Details

We have evaluated the nature of our proposed method on a very popular DRIVE dataset, which consists of 40 retinal images. The dataset is divided into two subsets, training sets and test sets, each set contains 20 images. The image is 565 x 584 pixels,8 bits per color channel. Only the images of training set can be used for the training of the network. First, we clip the training image into 23x23 tiles and mark each tile by the label image. Then add a slack variable to the tag based on the method above and use the mask contained in the DRIVE database to identify the FOV. All pixels of each patch should be in the FOV area, and a total of 100000 patches are obtained for 20 training images. But not all small pieces will be trained. In fact, we only used 50,000 small pieces, which accounted for only about 2

Test the performance of the segmentation algorithm in the test set. In 20 images of the test set, four images are pathological and the other are normal images. During the testing phase, we also use the first expert's tag as a basic fact. The images in all test sets are used for testing to evaluate the performance of the algorithm.

B. Evaluation Criterion

In vessel segmentation, there are two class labels: vessel and non-vessel. By comparing the segmentation results with the manual ground truth, we obtain four measures: the vessel pixels that are predicted as vessels are denoted as true positives (TP), the vessel pixels that are predicted as non-vessels are denoted as false negatives(FN), the non-vessel pixels that are predicted as non-vessels are denoted as true negatives (TN), and the non-vessel pixels that are predicted as vessels are denoted as false positives (FP).

Usually we use three criteria to compare the performance of the proposed method with other state-of-the-art methods: sensitivity (Se), specificity (Sp) and accuracy (Acc). Evaluation indicators are only calculated for pixels within the FOV.

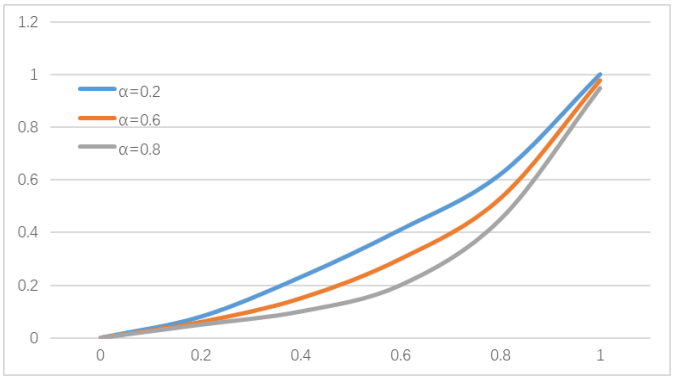


Fig. 2. Results of different segmentation methods

These metrics are defined as

$$Se = \frac{TP}{TP + FN} \tag{7}$$

$$Sp = \frac{TN}{TN + FP} \tag{8}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{9}$$

Because true positive score (Se) and false positive score (Sp) are sensitive to the number of sample categories. That is, when the number of samples of the binary classification problem is not balanced, these indicators cannot accurately reflect the performance of the classifier, so we also use the performance of the area under the ROC curve (AUC) evaluation method. The AUC is equal to 1 when the classifier can perfectly classify the sample. In addition, we have additionally introduced the concept of interclass accuracy.

This indicator better reflects the performance of the classifier when dealing with unbalanced categories.

C. Performance of the proposed method

Figure 2 is a comparison of our segmentation results with the traditional CNN segmentation results. Among them are (a) the original image, (b) the ground truth, (c) the segmentation result of the traditional CNN, and (d) the segmentation result of the proposed method. It can be clearly seen that our segmentation method is more delicate, and the segmentation accuracy of blood vessel details is higher and closer to the truth on the ground. The reason why our performance better is our method considers the effect of category imbalance on the segmentation of blood vessels. After modify the loss function and use multi-scale convolutional neural networks, we obviously reduce the impact of this problem on the results. In this method, the importance of the loss function is highlighted. Table III shows the comparison between the traditional CE and our proposed LCE. When using our structure at the same time, the accuracy of LCE is significantly higher than that of CE. Figure 3 shows the influence of different parameters ęÁ on the results. And through the experiments, we can get the best results when ęÁ=0.2. The proposed method is evaluated

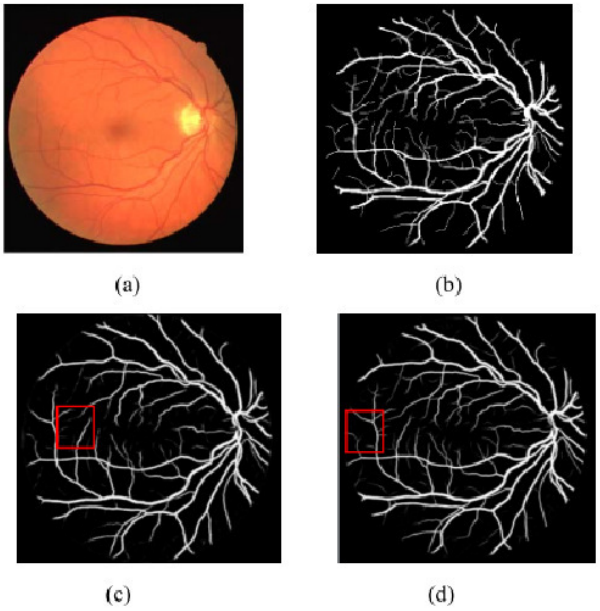


Fig. 3. The Effect of Different ęÁ Values on Loss in LCE Functions

TABLE II
COMPARISON OF CORRECT RATIOS OBTAINED BY DIFFERENT SAMPLE RATIOS ON DRIVE DATA SETS

Num	1:1	1:2	1:3	1:4
Acc	0.917	0.932	0.944	0.956

on DRIVE and STARE database images with available real ground images. The performance results are shown in TablesII and III. The performance index is calculated based on the first human observer. The accuracy, sensitivity, specificity, and AUC of the DRIVE database were 0.8347, 0.9796, 0.951, and 0.9792, respectively. The accuracy of the segmentation results of the STARE database is 0.956; the sensitivity, specificity, and AUC are 0.94471, 0.99432, and 0.988388, respectively. Figure 4 shows the performance of ROC curves on the DRIVE and STARE datasets. The average AUC of the ROC curves on the two datasets is 0.9792, 0.9743. As we know, the closer the AUC value is to 1, the better the performance of the classifier. So our retinal vessel segmentation results are excellent.

Then, we compare the proposed method with several advanced retinal vessel segmentation methods. In general, supervised learning methods have better classification accuracy than unsupervised learning methods. The methods in the reference document achieve better results than other methods because of

TABLE III
COMPARISON OF CORRECT RATE OF CE AND LCE ON DIFFERENT DATA SETS

	DRIVE	**STARE**
CE	0.937	0.932
LCE	0.951	0.956

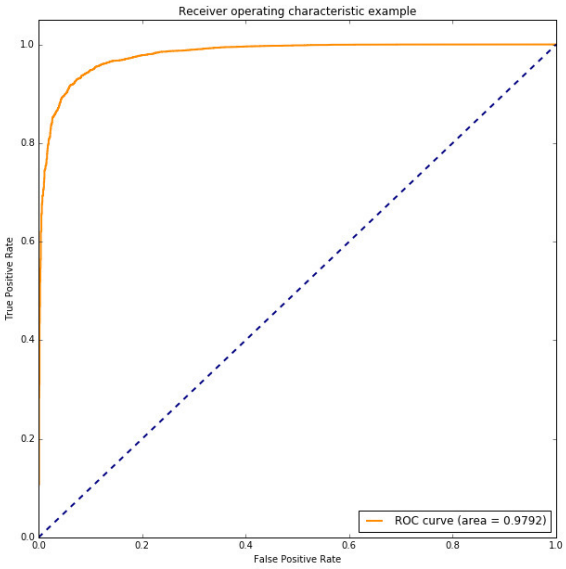


Fig. 4. ROC curve of proposed method

the use of ensemble learning methods. In a single classifier, our method has better average accuracy than other methods.

IV. DISCUSSIONS

Table II shows the accuracy of the classifier at different sample rates. On the DRIVE data set, we use the same network structure and evaluation indicators, different positive and negative sample ratios. From the above figure, we can clearly see that as the proportion of the sample much closer to the true proportion of the data set, the segmentation accuracy of our network becomes more accurate. Although it is not pursuit the unbalanced proportion on purpose, when randomly selecting small blocks from the sample set, it is actually obtained an imbalanced sample. This method can quickly classify the network, but it has no practical meaning. Because the network learning is the distribution state of the sample set itself, rather than the real characteristics of the sample.

After realizing that the above approach is not rigorous, we consider how to solve this problem. The modification of the loss function and network structure is the method that first enters our mind. Table III shows the performance of different loss functions on two data sets. After we used the improved loss function, the performance of the network has improved significantly. It is worth noting that our proposed improvement of the loss function not only has a broad prospect in this network but also in the broader field of deep learning. We are conducting experiments in this area and will soon have results. According to the definition of LCE, ęÁ is a very important hyper parameter, which determines the performance of this loss function on a specific problem. Figure 3 shows the impact of different ęÁ on network performance when we use LCE.

TABLE IV
PERFORMANCE COMPARISON OF VESSEL SEGMENTATION METHODS ON DRIVE IMAGES

No	Methods	Se	Sp	Acc	Auc
1	Fraz[1]	0.7302	0.9472	0.9422	N.A
2	Fraz[9]	0.7406	0.9807	0.9480	0.9747
3	Soares[10]	0.7283	0.9788	0.9466	0.9616
4	George[3]	0.7655	0.9704	0.9442	0.9614
5	Nicola[11]	0.7731	0.9724	0.9467	0.9588
6	Aslani[12]	0.7545	0.9801	0.9513	0.9682
7	Maji[13]	N.A	N.A	0.9470	0.9283
8	Lahiri[14]	0.7500	0.9800	0.9480	0.9500
9	Martina[15]	0.7276	0.9785	0.9466	0.9749
10	Avijit[16]	0.7691	0.9801	0.9533	0.9744
11	Fu[17]	0.7294	N.A	0.9470	N.A
12	Proposed method	0.8347	0.9796	0.9510	0.9792

TABLE V
PERFORMANCE COMPARISON OF VESSEL SEGMENTATION METHODS ON STARE IMAGES

No	Methods	Se	Sp	Acc	Auc
1	Hoover[18]	0.6747	0.9384	0.9348	N.A
2	Jiang[19]	N.A	N.A	0.9009	N.A
3	Mendonca[20]	0.6996	0.9730	0.9440	N.A
4	Lam[21]	N.A	N.A	0.9567	0.9739
5	You[22]	0.7260	0.9756	0.9479	N.A
6	Marin[15]	0.6944	0.9819	0.9526	0.9769
7	Fraz[1]	0.7548	9763	0.9534	0.9768
8	Proposed method	0.8231	0.9782	0.9560	0.9743

After our experiments, we found that when ęÁ=0.2, the best effect was obtained.

After using the method, we mentioned above, we have achieved very good results on the two data sets. The accuracy, sensitivity, specificity, and AUC of the DRIVE database were 0.8347, 0.9796, 0.951, and 0.9792, respectively. The accuracy of the segmentation results of the STARE database was 0.956; the sensitivity, specificity, and AUC were 0.94471, 0.99432, and 0.988388, respectively. In particular, except that our experiment is based on the results of a balanced sample size, other experimental results can be obtained when the sample is not balanced.

For blood vessel segmentation tasks, we are more likely to get blood vessel pixels than non-vascular pixels because blood vessel pixels are very rare and their value is much higher than non-vascular pixels. Therefore, in TableIV and Table V, our results showed a significantly higher specificity than other results. That means, in the case of a balanced sample, we only lost a little bit of accuracy, but we improved our specificity significantly. This is what we are happy to get. Our experiment has embarked on a new direction for the next study, which is not to regard accuracy as the first criterion, but rather to focus on specificity.

V. CONCLUSION

By comparing the differences of the experimental results, we found that there are unbalanced samples in the fundus image segmentation task, and we hope to improve the segmentation accuracy of blood vessels. Further we propose three solutions. By using these three methods together, we can get a better result. The deep neural network can learn hierarchically preprocessed images from it. It has a great potential in medical image processing and can help doctors easily diagnose accurately. In this paper, firstly, we use the slack variable method to increase the distance among different categories, thereby improving the performance of the classifier. Secondly, we propose a multi-scale convolutional neural network to extract the difference in information among different views, so that we can make accurate judgments. Finally, we solve the problem of unbalanced quantity among different types of samples by modifying the loss function. Our proposed method has performed well on two common data sets.

As we said above, we are conducting more tests and improvements on LCE so that it can perform well when dealing with unbalanced tasks for example target detection. The other limitation of our method is that it requires more training time than the previous method. Obviously, multi-scale networks have more parameters and the introduction of slack variables, which slows down the training speed of the network. Although the loss of time seems unavoidable, it can be acceptable to improve the accuracy. In the future, we hope to continue to improve the network structure so that it can be trained and tested more quickly.

REFERENCES

[1] M. M. Fraz, A. Basit, and S. A. Barman, "Application of morphological bit planes in retinal blood vessel extraction." *Journal of Digital Imaging*, vol. 26, no. 2, pp. 274–286, 2013.

[2] M. D. AbrÍd'moff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, and L. Tang, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *Jama Ophthalmology*, vol. 131, no. 3, p. 351, 2013.

[3] G. Azzopardi, N. Strisciuglio, M. Vento, and N. Petkov, "Trainable cosfire filters for vessel delineation with application to retinal images," *Medical image analysis*, vol. 19, no. 1, pp. 46–57, 2015.

[4] F. Zana and J.-C. Klein, "A multimodal registration algorithm of eye fundus images using vessels detection and hough transform," *IEEE transactions on Medical Imaging*, vol. 18, no. 5, pp. 419–428, 1999.

[5] P. DollÍćr, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*, 2009.

[6] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 2241–2248.

[7] S. R. Bulo, G. Neuhold, and P. Kontschieder, "Loss maxpooling for semantic image segmentation," *CVPR), July*, vol. 7, 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, and S. A. Barman, "An ensemble classification-based approach applied to retinal blood vessel segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2538–2548, 2012.

[10] J. V. Soares, J. J. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification," *IEEE Transactions on medical Imaging*, vol. 25, no. 9, pp. 1214–1222, 2006.

[11] N. Strisciuglio, G. Azzopardi, M. Vento, and N. Petkov, "Supervised vessel delineation in retinal fundus images with the automatic selection of b-cosfire filters," *Machine Vision and Applications*, vol. 27, no. 8, pp. 1137–1149, 2016.

[12] S. Aslani and H. Sarnel, "A new supervised retinal vessel segmentation method based on robust hybrid features," *Biomedical Signal Processing and Control*, vol. 30, pp. 1–12, 2016.

[13] D. Maji, A. Santara, P. Mitra, and D. Sheet, "Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images," *arXiv preprint arXiv:1603.04833*, 2016.

[14] A. Lahiri, A. G. Roy, D. Sheet, and P. K. Biswas, "Deep neural ensemble for retinal vessel segmentation in fundus images towards achieving label-free angiography," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 1340–1343.

[15] M. Melinščak, P. Prentašić, and S. Lončarić, "Retinal vessel segmentation using deep neural networks," in *VISAPP 2015 (10th International Conference on Computer Vision Theory and Applications)*, 2015.

[16] A. Dasgupta and S. Singh, "A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 248–251.

[17] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 698–701.

[18] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical imaging*, vol. 19, no. 3, pp. 203–210, 2000.

[19] X. Jiang and D. Mojon, "Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 1, pp. 131–137, 2003.

[20] A. M. Mendonca and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *IEEE transactions on medical imaging*, vol. 25, no. 9, pp. 1200–1213, 2006.

[21] B. S. Lam, Y. Gao, and A. W.-C. Liew, "General retinal vessel segmentation using regularization-based multiconcavity modeling," *IEEE Transactions on Medical Imaging*, vol. 29, no. 7, pp. 1369–1381, 2010.

[22] X. You, Q. Peng, Y. Yuan, Y.-m. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2314–2324, 2011.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 125–128
DOI: 10.15439/2018F111
ISSN 2300-5963 ACSIS, Vol. 15

Experiments with Classification of MMPI Profiles using Fuzzy Decision Trees

Krzysztof Pancerz
University of Rzeszów, Poland
Email: kpancerz@ur.edu.pl

Vitaly Levashenko, Elena Zaitseva
University of Zilina, Slovakia
Email: Vitaly.Levashenko@fri.uniza.sk
elena.zaitseva@fri.uniza.sk

Jerzy Gomuła
Cardinal Stefan Wyszyński University in Warsaw, Poland
Email: jerzy.gomula@wp.pl

***Abstract*—The paper is devoted to classification of MMPI (Minnesota Multiphasic Personality Inventory) profiles using fuzzy decision trees generated by means of the algorithm that uses cumulative information estimations of the initial data proposed by V. Levashenko et al. All of the stages of the classification process (i.e., fuzzification of the input data, generation of the classifier, testing the classifier) are presented and the results are discussed. A special attention is focused on determination of the center points on the MMPI scales for the fuzzification process.**

I. Introduction

OUR research, conducted for over eight years, is devoted to analysis and classification of data coming from the MMPI (Minnesota Multiphasic Personality Inventory) test (see some overview given in [1]). This test delivers psychometric data in the form of the so-called profiles (thirteen descriptive attributes corresponding to scales) used to assess patients in terms of personality traits and psychopathology. We have used several methodologies for classification of MMPI profiles which can be roughly grouped into the following categories: dissimilarity measure based classifiers, index based classifiers, classification functions, rule based classifiers, and decision tree based classifiers. A palette of classifiers has been extended by a classifier based on fuzzy decision trees. Firstly, preliminary results of research on application of fuzzy decision tress for classification of psychometric data presented in [2] are very promising in relation to results obtained for other kinds of classifiers (cf. [3]). Secondly, the character of MMPI data matches the idea of fuzzy set based approaches. For each scale included in the MMPI profile, we can define linguistic values (e.g., extremely low, very low, average, raised, high, very high, extremely high) which can be described by fuzzy sets. Fuzzy decision trees for creation of classifiers are generated by means of the algorithm based on cumulative information estimations of the initial data proposed by V. Levashenko et al. (see [4]). This algorithm is recalled in Section III-B. The main research problem, touched upon in this paper, is the determination of the intervals and/or the center points (shortly called the centers) on the MMPI scales for the fuzzification process. Therefore, we have tested different approaches which can be grouped into two categories. The first category includes approaches based on the expert knowledge used to determine intervals/centers for the fuzzification process. The second category includes approaches in which intervals/centers are induced from data.

II. Input Data

MMPI is a standardized psychometric test of adult personality and psychopathology (cf. [5]). The MMPI test delivers psychometric data in a form of the so-called profiles. Formally, a profile for the patient is a data vector consisting of values of thirteen descriptive attributes (corresponding to scales). The set of scales can be divided into two parts: the validity part (three scales: L - lying, F - infrequency, K - correction) and the clinical part (scales: $1.Hs$ - Hypochondriasis, $2.D$ - Depression, $3.Hy$ - Hysteria, $4.Pd$ - Psychopathic deviate, $5.Mf$ - Masculinity/Femininity, $6.Pa$ - Paranoia, $7.Pt$ - Psychasthenia, $8.Sc$ - Schizophrenia, $9.Ma$ - Hypomania, $0.It$ - Social introversion). In our research, we have used data coming from the WISKAD-MMPI test that is a Polish adaptation of the American test. The test originally was translated by M. Choynowski (as WIO) [6] and elaborated by Z. Płużek (as WISKAD) in 1950 [7]. The data set was collected for research by T. Kucharski and J. Gomuła in the Psychological Outpatient Clinic. It includes profiles of 1710 women. Before the profiles of women screened with the WISKAD-MMPI test formed a database for further experiments, first they had been sorted by the competent judges method - five specialists with many years of experience in the application and interpretation of the MMPI results/profiles. On the basis of these items, scores are calculated for both validity scales and clinical scales. Hence, values of descriptive attributes describing patients are expressed by the so-called T-scores [T]. The T-scores scale, which is traditionally attributed to MMPI, represents the following parameters: offset ranging from 0 to 100 T-scores, average equal to 50 T-scores, standard deviation equal to 10 T-scores. The scores are expressed as K-corrected T-Scores. The scales $1.Hs$, $4.Pd$, $7.Pt$, $8.Sc$, and $9.Ma$ are corrected by adding multiples of the scale K to them.

In our experiments, the patients' profiles are recorded in a tabular form which is formally called a decision table $DT = (U, Attr, Dec)$, where U - the set of cases (patients), $Attr = \{A_1, A_2, \ldots, A_{13}\}$ - the set of descriptive (condition) attributes corresponding to scales, $Dec = \{D\}$ - the set of decision attributes consisting of the attribute D assigning each patient from U to one of 20 classes such as the reference (*norm*) class and nosological types: neurosis (*neur*), psychopathy (*psych*), organic (*org*), schizophrenia

(*schiz*), delusion syndrome (*del.s*), reactive psychosis (*re.psy*), paranoia (*paran*), (sub)manic state (*man.st*), criminality (*crim*), alcoholism (*alcoh*), drug addiction (*drug*), simulation (*simu*), dissimulation (*dissimu*), and six deviational answering styles ($dev1$, $dev2$, $dev3$, $dev4$, $dev5$, $dev6$).

III. Methods and Tools

In this section, we present methods and tools used in experiments with classification of MMPI profiles by means of classifiers built on the basis of fuzzy decision trees.

A. Fuzzification

Fuzzification is the process that transforms the continuous value variables into linguistic variables whose domains contain linguistic values which can be described by fuzzy sets (their membership functions). Fuzzification is an important stage of the process of creation of a fuzzy decision tree based classifier. Many types of membership functions can be used to describe linguistic values, but triangular or trapezoidal shaped membership functions are the most common. In our approach, the fuzzification process consists of three stages. In the first stage, we determine intervals/centers (within the range $[0, 120]$) for each linguistic value assigned to a given descriptive attribute (scale). In the second stage, we define membership functions on the basis of centers determined in Stage 1 for each linguistic value assigned to a given descriptive attribute (scale). In the third stage, we calculate values of fuzzified descriptive attributes on the basis of membership functions defined in Stage 2. Determination of intervals/centers for the fuzzification process is one of the main research problems. In experiments, we have tested different approaches which can be grouped into two categories: approaches based on the expert knowledge used to determine intervals/centers (further, such approaches are called expert approaches) and approaches in which intervals/centers are induced from data (further, such approaches are called inductive approaches). The centers for four tested expert approaches are as follows:

- the Welsh's approach
  - all scales: 15.0, 35.0, 45.0, 55.0, 62.5, 67.5, 75.0, 85.0, 95.0, 110.0,
- the Płużek's (original) approach
  - L: 38.0, 43.0, 55.5, 75.5, 88.0,
  - F: 45.5, 60.5, 80.5, 100.5,
  - K: 36.0, 55.5, 74.5,
  - clinical scales: 34.5, 60.0, 75.5, 90.5, 105.5, 115.5,
- the Gomuła's (modified Płużek's) approach
  - validity scales: 15.0, 35.0, 42.5, 47.5, 55.0, 62.5, 67.5, 75.0, 82.5, 87.5, 100.0, 115.0,
  - clinical scales: 15.0, 37.5, 47.5, 52.5, 60.0, 67.5, 75.0, 90.0, 105.0, 115.0,
- the Gomuła's (original) approach:
  - all scales: 15.0, 35.0, 45.0, 55.0, 62.5, 67.5, 72.5, 77.5, 82.5, 87.5, 95.0, 105.0, 115.0.

The calculated centers for four tested inductive approaches are as follows:

- the K-means based approach [8]
  - L: 46.85, 55.69, 63.29, 79.58,
  - F: 56.43, 68.64, 80.82, 100.08,
  - K: 35.44, 48.60, 54.74, 66.47,
  - $1.Hs$: 53.14, 61.85, 69.29, 81.68,
  - $2.D$: 58.70, 70.22, 80.32, 92.67,
  - $3.Hy$: 55.97, 63.23, 69.61, 79.76,
  - $4.Pd$: 56.39, 65.07, 73.68, 87.68,
  - $6.Pa$: 57.45, 70.39, 83.23, 101.11,
  - $7.Pt$: 56.44, 67.28, 75.87, 95.50,
  - $8.Sc$: 58.06, 73.38, 85.22, 105.08,
  - $9.Ma$: 49.01, 58.11, 69.42, 85.47,
  - $0.It$: 51.26, 58.06, 62.39, 66.76,
- the equipotent interval approach
  - L: 46.25, 59.50, 76.25,
  - F: 55.75, 73.50, 94.75,
  - K: 38.75, 52.50, 68.75,
  - $1.Hs$: 41.25, 62.50, 89.25,
  - $2.D$: 48.75, 74.00, 99.25,
  - $3.Hy$: 42.25, 64.00, 89.75,
  - $4.Pd$: 42.25, 68.50, 95.75,
  - $6.Pa$: 44.25, 67.00, 96.25,
  - $7.Pt$: 41.75, 67.00, 88.75,
  - $8.Sc$: 45.75, 73.00, 98.75,
  - $9.Ma$: 37.25, 56.50, 83.75,
  - $0.It$: 42.25, 61.50, 75.25,
- the MDL based discretization approach (10 intervals)
  - L: 38.70, 44.10, 49.50, 54.90, 0.30, 65.70, 71.10, 76.50, 81.90, 87.30,
  - F: 47.65, 54.45, 60.75, 67.05, 73.35, 79.65, 85.95, 92.25, 98.55, 105.85,
  - K: 31.00, 37.50, 42.50, 47.50, 52.50, 57.50, 62.50, 67.50, 72.50, 79.00,
  - $1.Hs$: 36.50, 53.50, 60.50, 67.50, 74.50, 81.50, 88.50, 95.50, 102.50, 112.00,
  - $2.D$: 40.70, 57.10, 64.50, 71.90, 79.30, 86.70, 94.10, 101.50, 108.90, 116.30,
  - $3.Hy$: 37.35, 53.05, 57.75, 62.45, 67.15, 71.85, 76.55, 81.25, 85.95, 100.15,
  - $4.Pd$: 33.00, 49.50, 56.50, 63.50, 70.50, 77.50, 91.50, 84.50, 98.50, 110.50,
  - $6.Pa$: 37.65, 51.95, 59.25, 66.55, 73.85, 81.15, 88.45, 95.75, 103.05, 113.35,
  - $7.Pt$: 33.80, 50.90, 57.50, 64.10, 70.70, 77.30, 83.90, 90.50, 97.10, 103.70,
  - $8.Sc$: 35.85, 52.55, 60.25, 67.95, 75.65, 83.35, 91.05, 98.75, 106.45, 115.15,
  - $9.Ma$: 31.25, 44.75, 51.25, 57.75, 64.25, 70.75, 77.25, 83.75, 90.25, 100.75,
  - $0.It$: 31.75, 40.75, 45.25, 49.75, 54.25, 58.75, 63.25, 67.75, 72.25, 80.75,
- the MDL based discretization approach (min. 5 intervals)
  - L: 40.25, 49.50, 56.50, 62.00, 75.50, 87.75,
  - F: 52.75, 64.00, 68.00, 72.00, 78.00, 85.00, 89.50, 96.50, 106.25, 120.00,
  - K: 30.25, 36.50, 45.00, 52.00, 56.50, 64.25, 76.00, 120.00,
  - $1.Hs$: 38.25, 55.50, 58.50, 61.00, 63.50, 66.00, 70.50, 93.25, 120.00,
  - $2.D$: 44.75, 64.00, 67.50, 73.00, 85.00, 96.00, 109.75, 120.00,
  - $3.Hy$: 40.75, 58.50, 61.50, 66.00, 69.50, 73.50, 94.25, 120.00,
  - $4.Pd$: 39.75, 61.50, 65.00, 70.00, 74.50, 81.50, 103.25, 120.00,
  - $6.Pa$: 40.25, 57.50, 65.00, 70.00, 74.50, 78.50, 83.50, 97.00, 113.25, 120.00,
  - $7.Pt$: 37.75, 57.50, 61.00, 65.50, 70.50, 76.00, 85.50, 94.00, 101.75, 120.00,
  - $8.Sc$: 42.25, 62.50, 66.50, 70.00, 74.50, 81.50, 90.00, 102.50, 114.75, 120.00,
  - $9.Ma$: 33.25, 47.50, 52.50, 57.50, 61.50, 66.00, 70.50, 79.75, 97.50, 120.00,
  - $0.It$: 35.25, 53.50, 58.00, 61.00, 74.75.

In case of the K-means based approach, centers were generated for $k = 4$. In case of the equipotent interval approach, an unsupervised attribute discretization (the equal-frequency binning method) implemented in WEKA [9] was used. In this method, the same number of cases falls into each interval. Intervals are of different sizes. In case of the MDL based discretization approach, a supervised attribute discretization (the minimum length description method) implemented in WEKA was used. For more information, we refer readers to [10] and [11].

In each case, we have obtained a set (sequence) of centers located within the range $[0, 120]$. Formally, let $\{c_1, c_2, \ldots, c_{k_i}\}$ be a set of centers obtained for the i-th descriptive attribute. Triangular shaped membership functions are defined as follows.

1) For $j = 1$:

$$\mu_{c_j}(x) = \begin{cases} 1 & \text{if } x \geq 0 \text{ and } x \leq c_j, \\ 1 - \frac{x - c_j}{c_{j+1} - c_j} & \text{if } x > c_j \text{ and } x \leq c_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

2) For $j > 1$ and $j < c_{k_i}$:

$$\mu_{c_j}(x) = \begin{cases} \frac{x-c_{j-1}}{c_j-c_{j-1}} & \text{if } x \geq c_{j-1} \text{ and } x \leq c_j, \\ 1 - \frac{x-c_j}{c_{j+1}-c_j} & \text{if } x > c_j \text{ and } x \leq c_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

3) For $j = c_{k_i}$:

$$\mu_{c_j}(x) = \begin{cases} \frac{x-c_{j-1}}{c_j-c_{j-1}} & \text{if } x \geq c_{j-1} \text{ and } x \leq c_j, \\ 1 & \text{if } x > c_j \text{ and } x \leq 120, \\ 0 & \text{otherwise.} \end{cases}$$

Let $ls = c_j - 0.2(c_{j+1} - c_{j-1})$ and $rs = c_j + 0.2(c_{j+1} - c_{j-1})$, trapezoidal shaped membership functions are defined as follows.

1) For $j = 1$:

$$\mu_{c_j}(x) = \begin{cases} 1 & \text{if } x \geq 0 \text{ and } x \leq rs, \\ 1 - \frac{x-rs}{c_{j+1}-rs} & \text{if } x > rs \text{ and } x \leq c_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

2) For $j > 1$ and $j < c_{k_i}$:

$$\mu_{c_j}(x) = \begin{cases} \frac{x-c_{j-1}}{ls-c_{j-1}} & \text{if } x \geq c_{j-1} \text{ and } x \leq ls, \\ 1 & \text{if } x > ls \text{ and } x < rs, \\ 1 - \frac{x-rs}{c_{j+1}-rs} & \text{if } x \geq rs \text{ and } x \leq c_{j+1}, \\ 0 & \text{otherwise.} \end{cases}$$

3) For $j = c_{k_i}$:

$$\mu_{c_j}(x) = \begin{cases} \frac{x-c_{j-1}}{ls-c_{j-1}} & \text{if } x \geq c_{j-1} \text{ and } x \leq ls, \\ 1 & \text{if } x > ls \text{ and } x \leq 120, \\ 0 & \text{otherwise.} \end{cases}$$

It is worth noting that, in each approach, intervals are disjoint. However, one can see that membership functions overlap.

Let $DT = (U, Attr, Dec)$ be a decision table containing MMPI data, where $Attr = \{A_1, A_2, \ldots, A_{13}\}$ and $Dec = \{D\}$. After fuzzification, for each descriptive attribute A_i, where $i = 1, 2, \ldots, 13$, we obtain k_i fuzzified attributes $A_i^1, A_i^2, \ldots, A_i^{k_i}$, where k_i is a number of linguistic values used for fuzzification of A_i. In case of the decision attribute D, we obtain 20 fuzzified attributes, each for one decision class, i.e., $D^{norm}, D^{neur}, \ldots, D^{dev6}$. However, values of the attributes $D^{norm}, D^{neur}, \ldots, D^{dev6}$ are binary. For example, $D^{norm}(u) = 1$ if $D(u) = norm$, and 0 otherwise, where $u \in U$.

B. Fuzzy Decision Trees

To build the classifier, we have used the algorithm for generation of fuzzy decision trees that uses cumulative information estimations of the initial data proposed by V. Levashenko et al. (see [4]). This algorithm was used by us in our preliminary experiments with the MMPI data (see [2]). The obtained results were very promising in relation to results obtained for other kinds of classifiers (cf. [3]). In general, the cumulative mutual information for a given attribute A_i, a sequence of attributes SFA, and the decision attribute D reflects the influence of the attribute A_i on the attribute D when the sequence SFA of attributes is known.

In this section, we briefly recall the algorithm used in our experiments. Let us assume the following notation: U - the set of cases, n - the number of cases, $lval(A)$ - the set of all linguistic values used for the fuzzification process of the attribute A, $cer(D^v)$ - the certainty of the decision class D^v of the attribute D, RA - the set of the remaining descriptive attributes, SFA - the set of the selected fuzzified attributes. The algorithm is recursive (see Procedure 1). There are two tuning parameters θ_{freq} and θ_{cer} used in the algorithm as the stop conditions. Expanding a tree branch is stopped when either the frequency of the branch is below θ_{freq} or when more than or equal to θ_{cer} percent of cases left in the branch has the same decision class label. Moreover, the natural stop condition is fulfilled if the set of the remaining descriptive attributes is empty (i.e., $RA = \emptyset$).

The cardinality measure of the set B of fuzzified attributes is defined as $card(B) = \sum_{u \in U} \prod_{B_i \in B} B_i(u)$. The certainty $cer(D^v)$ of the decision class D^v is calculated as $cer(D^v) = card(SFA \cup \{D^v\})$.

Procedure FDT

Data: A decision table $DT = (U, Attr, Dec)$
$RA \leftarrow Attr$; $SFA \leftarrow \emptyset$;
$E(D) \leftarrow n \log(n) - \sum_{v \in lval(D)} card(\{D^v\}) \log(card(\{D^v\}))$;
foreach $A_i \in RA$ **do**
  $E(A_i) \leftarrow n \log(n) - \sum_{v \in lval(A_i)} card(SFA \cup \{A_i^v\}) \log(card(SFA \cup \{A_i^v\}))$;
  $E(D, A_i) \leftarrow n \log(n) - \sum_{v \in lval(D), w \in lval(A_i)} card(SFA \cup \{D^v\} \cup \{A_i^w\}) \log(card(SFA \cup \{D^v\} \cup \{A_i^w\}))$;
  $CMI(A_i) \leftarrow E(D) + E(A_i) - E(D, A_i)$;
Select A_i from RA with the greatest $CMI(A_i)$;
$RA \leftarrow RA - \{A_i\}$;
foreach $v \in lval(A_i)$ **do**
  $SFA \leftarrow SFA \cup \{A_i^v\}$;
  if $\max_{cer(D_v)} < \theta_{cer}$ *and* $\frac{card(SFA)}{n} \geq \theta_{freq}$ *and* $RA \neq \emptyset$ **then**
    call FDT with $DT = (U, RA, Dec)$;
  else
    create a decision node;

C. The CLAPSS System

All of the stages of the classification process (fuzzification of the input data, generation of the classifier, testing the classifier) were performed using our software tool called CLAPSS (Classification and Prediction Software System) [12] that is a tool developed for solving different classification and prediction problems using, among others, some specialized approaches based mainly on fuzzy sets and rough sets. A new module added to CLAPSS consists of implementation of the selected methods based on fuzzy sets (especially to creation of classifiers based on fuzzy decision trees generated by means of the algorithm described in Section III-B). The user that creates classifiers based on fuzzy decision trees can select among others: a fuzzification process (*triangular*, *trapezoidal*, *Gaussian*), thresholds (*certainty* and *frequency*) to stop the process of fuzzy decision tree creation, t-norm

(*minimum*, *algebraic product*, *Lukasiewicz product*, *Einstein product*, *Hamacher product*, *drastic product*) for calculation of the certainty of rule antecedents, and a number of folds for the cross-validation procedure.

IV. Results

Our experiments were performed on real-life data described in Section II using the CLAPSS software tool (see Section III-C). In each experiment, the $5.Mf$ scale was excluded. This scale is assumed by the experts to be the weakest one. In each case, the stratified 10-cross-validation approach was used to test the classifier. In the experiments, the following settings have been used: a shape of membership functions: *triangular* and *trapezoidal*, the certainty threshold: 0.999, the frequency threshold: 0.001, the t-norm for calculation of the certainty of rule antecedents: *algebraic product*. The results of the stratified 10-cross-validation tests for all approaches (both expert and inductive) are presented in Figure 1 (for triangular shaped membership functions) and Figure 2 (for trapezoidal shaped membership functions). The results showed that expert approaches like Gomuła's (original), Gomuła's (modified Płużek's), and Welsh's are suitable for the fuzzy decision tree based classification. Among inductive approaches, good results were obtained for MDL based multi-interval discretization. It is worth noting that easy and natural diagnostic interpretation of the obtained intervals becomes the advantage of the expert approaches. Weak results obtained for the original Płużek's intervals do not seem to be surprising because this approach is recognized by the experts as rough. The approach based on equipotent intervals turned out to be inappropriate. On the basis of the results, one can see that classifiers based on fuzzy decision trees show a high effectiveness (accuracy noticeably greater than 0.9) in classification of the MMPI data. If we take into consideration solely the MMPI scales (without any additional indexes), only a few previously tested approaches are found to be such effective (cf. [3]).

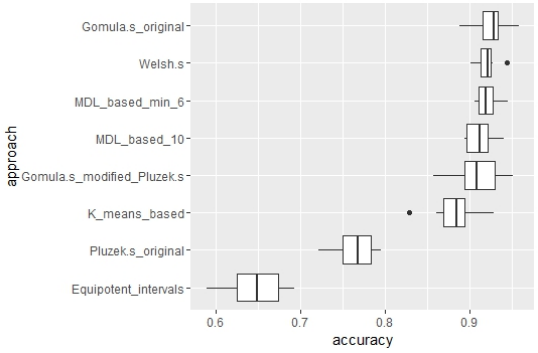


Fig. 1. Results for triangular shaped membership functions.

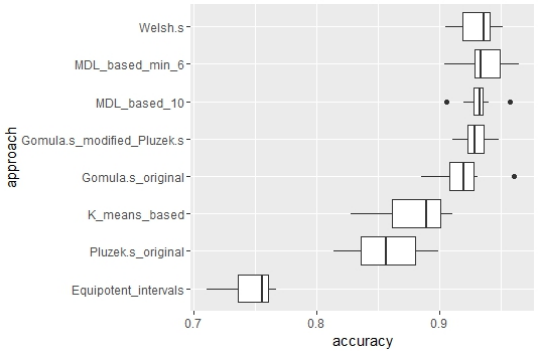


Fig. 2. Results for trapezoidal shaped membership functions.

V. Conclusions and Further Work

In general, classifiers based on fuzzy decision trees showed a high effectiveness in classification of the MMPI data. The main challenge in the future is to propose the method for searching for optimal intervals used in the fuzzification process. Simultaneously, we need to take care of diagnostic interpretation of the obtained intervals. Therefore, automated searching for optimal intervals should be aided with the expert knowledge. This fact determines the main direction of our further research. Moreover, we plan to test application of some other shapes of membership functions and some other t-norms for calculation of the certainty of rule antecedents.

References

[1] K. Pancerz, O. Mich, A. Burda, and J. Gomuła, "A tool for computer-aided diagnosis of psychological disorders based on the MMPI test: An overview," in *Applications of Computational Intelligence in Biomedical Technology*, ser. Studies in Computational Intelligence, R. Bris, J. Majernik, K. Pancerz, and E. Zaitseva, Eds. Cham: Springer International Publishing, 2016, vol. 606, pp. 201–213.

[2] V. Levashenko, E. Zaitseva, K. Pancerz, and J. Gomuła, "Fuzzy decision tree based classification of psychometric data," in *Position Papers of FedCSIS'2014*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 3. Warsaw, Poland: PTI, 2014, pp. 37–41.

[3] O. Mich, A. Burda, K. Pancerz, and J. Gomuła, "The knowledge base for computer-aided diagnosis of mental disorders based on psychometric tests," in *Proceedings of DT'2014*, Zilina, Slovakia, 2014. doi: 10.1109/DT.2014.6868724 pp. 255–261.

[4] V. Levashenko and E. Zaitseva, "Usage of new information estimations for induction of fuzzy decision trees," in *Proceedings of IDEAL 2002*, ser. LNCS, H. Yin, N. Allinson, R. Freeman, J. Keane, and S. Hubbard, Eds., vol. 2412. Springer Berlin Heidelberg, 2002. doi: 10.1007/3-540-45675-9_74 pp. 493–499.

[5] D. Lachar, *The MMPI: Clinical assessment and automated interpretations*. Fate Angeles: Western Psychological Services, 1974.

[6] M. Choynowski, *Wielowymiarowy inwentarz osobowości (in Polish)*, Psychometry Laboratory, Polish Academy of Sciences, Warsaw, 1964.

[7] Z. Płużek, "Wartość diagnostyczna testu WISKAD-MMPI w zakresie nozologii psychiatrycznej (in Polish)," *Roczniki Filozoficzne / Annales de Philosophie / Annals of Philosophy*, vol. 17, no. 4, pp. 125–143, 1969.

[8] H.-M. Lee, C.-M. Chen, J.-M. Chen, and Y.-L. Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 3, pp. 426–432, 2001. doi: 10.1109/3477.931536

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[10] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of IJCAI'1993*, Chambery, France, 1993, pp. 1022–1029.

[11] I. Kononenko, "On biases in estimating multi-valued attributes," in *Proceedings of IJCAI'1995*, Montreal, Quebec, Canada, 1995, pp. 1034–1040.

[12] K. Pancerz, "On selected functionality of the classification and prediction software system (CLAPSS)," in *Proceedings of IDT'2015*, Zilina, Slovakia, 2015. doi: 10.1109/DT.2015.7222984 pp. 278–285.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 129–135
DOI: 10.15439/2018F196
ISSN 2300-5963 ACSIS, Vol. 15

Imputing Missing Values for Improved Statistical Inference Applied to Intrauterine Growth Restriction Problem

Agnieszka Wosiak, Kinga Glinka
Lodz University of Technology
Institute of Information Technology
ul. Wolczanska 215
90-924 Lodz, Poland
Email: agnieszka.wosiak@p.lodz.pl, kinga.glinka@edu.p.lodz.pl

Agata Zamecznik, Katarzyna Niewiadomska-Jarosik
Department of Pediatric
Cardiology and Rheumatology
2nd Chair of Pediatrics
Medical University of Lodz, Poland
Email: agazamek@gmail.com, kasiajarosik@wp.pl

***Abstract*—The paper describes the study on the problem of missing values in medical data collected to discover new dependencies between parameters in children born with intrauterine growth restriction disorder. The aim of the research is to propose a procedure that may be taken to improve the medical inference in the presence of missing data. The approach with use of unconditional mean and k-nearest neighbor imputation has been applied. The experiments proved that application of missing data imputation in original dataset yields more valuable dependencies when compared to original data, maintaining the confidence interval for goodness of fit with the original distribution above 90%. The discovered dependencies in data may establish the basis for new treatment procedures of children with intrauterine growth restriction disorder.**

***Index Terms*—missing values, imputation, medical data analysis, intrauterine growth restriction disorder**

I. Introduction

THE IMPROVEMENT of medical diagnostics and health care is based on scientific studies, which are often based on observations gathered from patients. The reliable analysis of medical dataset usually assumes that subjects of the research were chosen randomly from a greater population at the beginning of the trial. Such an approach is called a randomized controlled trial (RTC) and the analysis of its data is referred to as intention-to-treat (ITT) principle [1].

According to the ITT strategy, all the participants should be included in the analysis regardless whether their outcomes were actually collected [2]. At the same time, the ITT principle requires a complete set of data [3]. The "ideal" ITT analysis is usually not possible to perform, as the problem of missing values commonly occurs [4]. The lacking entries are basically caused by the fact that patient's data are usually gathered as a product of care actions, rather than an organized research protocol [5], [6]. Moreover, in many medical studies, the patients may withdraw or drop out from the trials, which is almost unavoidable and their data may be incomplete [1].

Therefore, appropriate procedures have been created in order to overcome the problem of missing data and estimate a treatment effect.

In most situations, a complete case analysis is considered, and the data of patient with missing vales are discarded. However, in some areas of medical research, more than 50% of missing entries may be encountered [5], [7]. Removing some instances leads to smaller datasets and as a result, to loss of statistical power of the analysis.

As an alternative to a complete case analysis, dropping variables with missing values from the analysis may be also applied [4], [8]. This approach, in turn, neglects valuable observed data and causes less beneficial data analysis.

Another common procedure, that may be a kind of compromise between a complete case analysis and dropping variables, is an available case analysis, where only a piece of patient's data, where no values are provided, is neglected. Despite the complications in analyzing such data due to differences in numbers of instances for various parameters, the method may produce biased estimates of associations [9].

The approach with use of imputation methods is of increasing importance nowadays. Many studies have been conducted on the topic of data imputation techniques [4], [10]–[12], but due to the complexity of the problem of missing data and its close relationship to inner data characteristics, no universal procedure has been discovered and researchers still strive to find standards in data imputation [13].

The aim of this paper is to verify, if appropriate imputation techniques can improve medical inference applied to the problem of intrauterine growth restriction and its relationship with metabolic disorders. There is no universal statistical method that deals with missing data as each study has its own design, measurement characteristics and different assumptions about missing data mechanisms [13]. Therefore, the research constitutes an independent contribution to the relevant literature and also attempts to find a successful way to perform accurate statistical analysis of IUGR in terms of missing data.

The rest of the paper is organized as follows. Section II corresponds to missing values imputation techniques. Section III explains the medical problem of IUGR and is followed by the description of medical data used in the research.

Next, section IV is dedicated to the experiments conducted on sample data and the results. Finally, in Section V, the concluding remarks are discussed.

II. Missing Values Imputation

The imputation methods can be divided into two categories:

- single imputation algorithms,
- multiple imputation algorithms.

A. Single Imputation Methods

In single imputation approach, missing data are imputed by single values. The most popular technique is the mean imputation (MI). The method uses mean of the values of an attribute that contains missing data. The modification of MI technique is using the mode instead of the mean, i.e. the most frequent value in the case of categorical attributes.

Two variations of MI can be distinguished: conditional and unconditional. The unconditional mean imputation (UMI) is not conditioned on the values of other parameters that describe the patient's data. Conditional mean approach (CMI) imputes a mean value that depends on the complete attributes for the analyzed record.

The widely applied single imputation technique is the hot deck [5], [14], [15]. The procedure finds the most similar object for the record that contains missing data, and the missing values are imputed from that object. If the most similar object also contains missing data for the same parameters as in the imputed record, then another closest object is found, until all the missing values are successfully imputed. To find the closest object, several distance functions can be used [16].

One of the hot deck techniques used to compensate for missing data is called kNN imputation (kNNI) [17]. It uses k closest complete instances in the dataset for imputing a missing value, assuming that the k most relevant complete objects are the k nearest neighbors of the incomplete instance in the dataset.

Another approach is based on regression (RI) of the missing values using complete data for a given record [18]. Different regression models can be used, usually depending on the types of imputed parameters, e.g. linear for numerical attributes, logistic for binary features or polytomous for discrete values.

B. Multiple Imputation Methods

Multiple imputation methods use several ordered choices for imputing the missing values [9]. The procedure is performed by creating several complete datasets, in which different imputations are based on a random draw from separately estimated underlying distributions.

One of the most popular approach to multiple imputation is multivariate imputation by chained equations (MICE) described in [19]. It provides a full spectrum of conditional distributions and related regression based methods (linear regression, logistic regression and polytomous regression). To make the application of MICE available, a missing data imputation software package was developed [20].

Multiple imputation algorithms also include:

- Markov chains [21],
- machine learning algorithms (e.g EM algorithm) [22],
- genetic algorithms [23].

Results based on those complex methods are increasingly reported, but their use needs to be applied carefully to avoid misleading conclusions. The multiple imputation procedures require modeling the distribution of each attribute with missing values based on the observed data. Therefore, the validity of results performed on the modified datasets depends on the correctness of such modelling [24].

C. Selection of Imputation Methods

The selection of imputation techniques was determined by the assumption that they should be simple and comprehensive, so that human expert could understand the underlying mechanisms. Moreover, the availability of the methods in statistical program packages such as StatSoft Statistica and SPSS facilitates their use [25]. It was also reported that unsupervised imputation methods may provide more accurate imputation for large amounts of missing data [5]. Therefore, in the experimental studies three single imputation methods were applied: unconditional mean imputation (UMI), conditional mean imputation (CMI) and k nearest neighbor with $k = 5$ (5NNI).

III. Data Description

Intrauterine growth restriction (IUGR) is a fetal disorder of growing. It is often related to fetal hypoxia and higher percentage of perinatal mortality. IUGR is a risk factor for many cardiovascular, metabolic, and pulmonologic diseases in adult life [26]. It occurs in about 3-10% of live-born newborns, but in developing countries it concerns up to 20-30% of newborn infants [27]. The comparisons of absolute measurements of the fetuses with reference values, as well as birth weight percentiles, allow detection of deviations between expected and actual fetal growth and identification of newborns being possibly at risk for adverse health events [28]. However, the diagnosis of IUGR is based on non-consistent definitions [29].

The world-wide research studies report that IUGR makes a risk factor for metabolic syndrome [30], [31], however more environmental studies are still needed to put additional treatment in practice [32]–[34].

The research was based on a study group (SG) of 113 children aged 5-10 years (average 8.1 $\pm$ 1.5) born on term with IUGR and birth weight below 10 percentile according to gestational age for the Polish population [35] and a control group (CG) of 39 children aged 4.5 - 12 (average 7.6 $\pm$ 1.2). All patients were selected during prospective studies at the Pediatric Cardiology and Rheumatology Department of Medical University of Lodz in 2010-2013. The study was approved by Medical Ethical Committee of the Health Sciences Faculty of Lodz University (No: RNN/760/10/KB).

The characteristics of all parameters subjected to further analysis included general attributes, cardiovascular parameters, lipids levels and adipocytokines values. Most of the parameters had missing values. The characteristics of the dataset is

presented in Table I, where the first column refers to the type of an attribute, the second is the name, next three columns include the range of values, the mean and standard deviation and the last column holds the percentage of missing values.

IV. Results and Discussion

The purpose of experiments was to find how the missing values imputation methods improve medical inference for the intrauterine growth restriction problem by discovering new significant correlations between attributes.

The experiments were conducted according to the methods introduced in Section II on the dataset described in Section III. Three main procedures were performed:

A. The experimental procedure that includes analysis with original but incomplete data.
B. The experimental procedure that results in choosing the best imputation technique.
C. The experimental procedure that performs the analysis with the imputed data.

A. Experimental Procedure that Includes Analysis with Original but Incomplete Data

The procedure was performed to discover the characteristics of all parameters and to perform their comparison between the control and study groups. The intention was to confirm by the epidemiological studies the hypothesis that:

- IUGR enhances the susceptibility to metabolic syndrome, and
- there is a correlation between levels of lipids and adipocytokines in IUGR group.

The results of the statistical analysis for the original dataset are presented in Table II.

The first hypothesis was successfully confirmed only for total cholesterol and triglicerides. The significant differences for the rest of parameters were not possible to obtain, mostly due to the numerous missing values and interrelated low significance level, as the level of missing values for adipocytokines was almost 50% in the study group and over 80% in the control group. Therefore, the presence of relationship between lipids and adipocytokines was not possible to be confirmed by statistical analysis as well.

B. Experimental Procedure that Results in Choosing the Best Imputation Technique

In literature the imputation methods are usually related to machine learning problems, mainly to the classification [11], [36]–[38]. Then, the validation can be based on comparisons of imputed datasets to the results obtained for the complete datasets with use of the standard classification metrics, e.g. accuracy or TP rate.

The IUGR problem described in the paper did not refer to the classification, and no class labels were available. Therefore, the choice of the best imputation technique was based on the differences between distributions for the complete sets of data and the sets with randomly dropped and artificially imputed values. The procedure involved five steps:

1) Choose the parameters that were originally complete.
2) Randomly introduce missing data into each parameter in the amounts of: 5%, 10%, 20%, 30%, 40% and 50%.
3) Impute the missing values in each dataset using three imputation methods: mean, conditional mean and kNN with k=5.
4) Compare the distributions of original and modified data for each parameter.
5) For each amount of data imputed, choose the method that built the distribution closest to the original distribution.

In our dataset only 5 parameters out of 18 were originally complete and those data were further used for verification the best suitable imputation technique.

We used missing completely at random (MCAR) approach to drop the data. Values were dropped in the amounts of 5%, 10%, 20%, 30%, 40% and 50%. Each type of missing values' generation was repeated 10 times. As a result we performed 150 experiments (5 attributes x 10 draws x 3 imputation methods). The percentage of cases, where the particular imputation technique was the closest to original distribution, taking into account the amounts of imputed data, are presented in Table III.

The results of comparison for distributions revealed that either the simplest imputation technique by mean values, or more complex with 5NN, can be used for imputation in term of our IUGR datasets. However, it can be also noticed that the imputation by mean gives better results when only small amounts of data are missing. Moreover, the experiments revealed that for smaller amount of missing data, the confidence intervals for goodness of fit with original distribution were above 95%, and at least 80% for the highest amounts of missing values.

C. Experimental procedure that performs the analysis with the imputed data

The final procedure was performed in three steps:

1) Impute the missing data with the method that resulted best for a specified amount of missing values.
2) Perform comparison of characteristics of all parameters between datasets with original and imputed values.
3) Verify the dependencies between lipids and adipocytokines with use of correlation analysis.

As the indication of the best imputation method was not clear enough (although there was a slightly higher recommendation of 5NNI), we decided to use two approaches for further analysis: unconditioned mean imputation and 5NN imputation.

The results of the analysis that includes unconditional mean imputation were presented in Table IV, whereas Table V presents the results for 5-nearest neighbor imputation.

When comparing results of statistical analysis for original dataset (Table II) and for the dataset imputed by unconditional mean (Table IV), one can notice that the statistically significant differences were attained for the parameters with rather small amount of missing data (12% for glucose - Fig. 1 and HDL - Fig. 2), whereas no additional medical conclusions could be drown for the parameters with higher levels of missing values.

TABLE I: Characteristics of attributes for the dataset

Type	Name	Range	Mean	Standard Deviation	Missing values
General	Age (years)	4.5 - 12.0	7.97	1.68	0%
	Body mass (kg)	14.5 - 73.0	25.08	7.73	0%
	BMI	10.8 - 31.6	15.72	2.60	0%
	Birth mass (g)	1800 - 4700	2808	472	0%
	Gestational age	38 - 42	39	0.89	0%
Cardiovascular	Average heart rate	70 - 120	88	10	14%
	SBP	11 - 129	103	13	14%
	DBP	40 - 85	62	8	14%
	SBP load	0 - 96	21.79	17.33	5%
	DBP load	0 - 60	10.54	10.81	5%
Lipids	Glucose	66 - 133	85.90	8.83	12%
	Total cholesterol (mg/dl)	81 - 214	155.11	25.22	12%
	HDL (mg/dl)	27.9 - 100.6	60.91	15.76	12%
	LDL (mg/dl)	24.0 - 133.7	82.53	19.20	12%
	Triglicerides (mg/dl)	24 - 236	70.93	32.99	12%
Adipocytokines	Leptin (ng/ml)	0.48 - 30.79	6.52	6.61	60%
	Adiponectin (μg/dl)	7.33 - 36.70	19.92	6.22	60%
	Resistin (ng/ml)	1.23 - 9.73	2.45	1.50	60%

TABLE II: Characteristics of lipids and adipocytokines levels in the original dataset

	Study group (SG)	Control group (CG)	
Parameter	**Mean $\pm$SD (*)**	**Mean $\pm$SD (*)**	**p-value (**)**
Glucose (mg/dl)	86.50 $\pm$ 9.55	84.46 $\pm$ 6.67	0.228
Total cholesterol (mg/dl)	159.08 $\pm$ 25.37	145.44 $\pm$ 22.32	0.004
HDL (mg/dl)	62.14 $\pm$ 14.27	57.93 $\pm$ 18.80	0.160
LDL (mg/dl)	81.77 $\pm$ 20.23	84.39 $\pm$ 16.57	0.475
Triglicerides (mg/dl)	75.99 $\pm$ 36.54	58.62 $\pm$ 16.97	0.005
Leptin (ng/ml)	6.68 $\pm$ 6.78	4.35 $\pm$ 3.56	0.500
Adiponectin (μg/dl)	19.94 $\pm$ 6.21	19.83 $\pm$ 7.42	0.974
Resistin (ng/ml)	2.48 $\pm$ 1.55	2.01 $\pm$ 0.37	0.551

(*) described as average values $\pm$ standard deviations
(**) p-value $<$0.05 defines statistical significance

TABLE III: Summary of evaluation for imputation techniques

% of missing values	UMI	CMI	5NNI
5%	80%	10%	10%
10%	40%	20%	40%
20%	50%	10%	40%
30%	40%	0%	60%
40%	40%	20%	40%
50%	30%	20%	50%

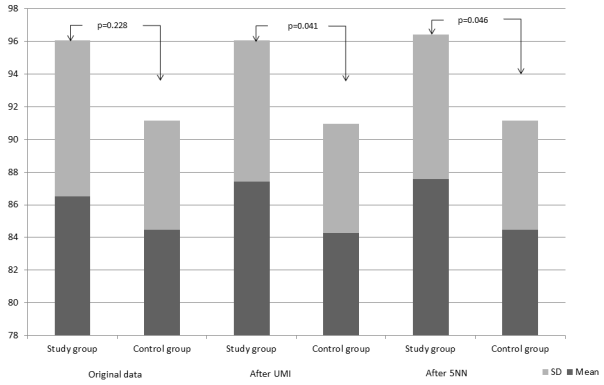


Fig. 1: Differences for glucose in study and control groups before and after performing imputation

When data were imputed with 5NN (Table V), new differences were discovered between levels of leptin and resistin, for which the percentage of missing values equaled 60%.

The correlations between lipids and adipocytokines are presented in Tables VI and VII, for datasets after imputation by unconditioned mean and 5NN respectively.

Unconditioned mean imputation enabled discovering statistically significant correlations between glucose and resistin, total cholesterol and leptin, HDL and leptin, and triglicerids and resistin. The 5NN imputation additionally revealed depen-

TABLE IV: Characteristics of lipids and adipocytokines levels in the dataset after unconditional mean imputation

	Study group (SG)	**Control group (CG)**	
Parameter	**Mean ±SD (*)**	**Mean ±SD (*)**	**p-value (**)**
Glucose (mg/dl)	84.26 ± 6.68	84.40 ± 8.65	0.041
Total cholesterol (mg/dl)	145.44 ± 22.32	158.45 ± 23.29	0.003
HDL (mg/dl)	57.53 ± 16.20	62.95 ± 13.08	0.038
LDL (mg/dl)	84.39 ± 16.57	81.89 ± 18.53	0.457
Triglicerides (mg/dl)	58.61 ± 16.97	75.18 ± 13.52	0.004
Leptin (ng/ml)	6.30 ± 1.20	6.60 ± 4.66	0.692
Adiponectin (μg/dl)	19.92 ± 2.09	19.93 ± 4.27	0.984
Resistin (ng/ml)	2.41 ± 0.17	2.46 ± 1.07	0.726

(*) described as average values ± standard deviations
(**) p-value <0.05 defines statistical significance

TABLE V: Characteristics of lipids and adipocytokines levels in the dataset after imputation with use of 5-nearest neighbor

	Study group (SG)	**Control group (CG)**	
Parameter	**Mean ±SD (*)**	**Mean ±SD (*)**	**p-value (**)**
Glucose (mg/dl)	84.46 ± 6.68	87.58 ± 8.83	0.046
Total cholesterol (mg/dl)	145.44 ± 22.32	160.62 ± 23.60	0.001
HDL (mg/dl)	57.93 ± 18.80	63.38 ± 13.17	0.049
LDL (mg/dl)	84.39 ± 16.57	83.67 ± 19.12	0.834
Triglicerides (mg/dl)	58.62 ± 16.97	76.31 ± 13.61	0.001
Leptin (ng/ml)	3.67 ± 1.25	5.55 ± 4.82	0.017
Adiponectin (μg/dl)	20.47 ± 2.32	20.17 ± 4.38	0.691
Resistin (ng/ml)	2.36 ± 0.59	2.74 ± 1.13	0.048

(*) described as average values ± standard deviations
(**) p-value <0.05 defines statistical significance

TABLE VI: Correlations between lipids and adipocytokines in the dataset after unconditional mean imputation

Parameter	**Leptin**		**Adiponectin**		**Resistin**	
	r	**p-value (*)**	**r**	**p-value (*)**	**r**	**p-value (*)**
Glucose	0.1034	0.276	0.0871	0.359	-0.4631	0.013
Total cholesterol	0.3616	0.036	0.0673	0.479	-0.0262	0.783
HDL	0.2405	0.020	0.0531	0.576	0.0036	0.970
LDL	0.1165	0.219	0.0479	0.615	-0.0876	0.356
Triglicerides	0.1104	0.244	-0.0335	0.725	0.2861	0.023

(*) p-value <0.05 defines statistical significance

TABLE VII: Correlations between lipids and adipocytokines in the dataset after 5NN imputation

Parameter	**Leptin**		**Adiponectin**		**Resistin**	
	r	**p-value (*)**	**r**	**p-value (*)**	**r**	**p-value (*)**
Glucose	0.2626	0.046	0.0762	0.423	-0.3647	0.050
Total cholesterol	0.3632	0.042	-0.0013	0.989	-0.0096	0.919
HDL	0.2809	0.037	0.0425	0655.	-0.0014	0.988
LDL	0.1004	0.290	-0.0162	0.865	-0.0624	0.512
Triglicerides	0.0963	0.310	-0.0413	0.664	0.2806	0.037

(*) p-value <0.05 defines statistical significance

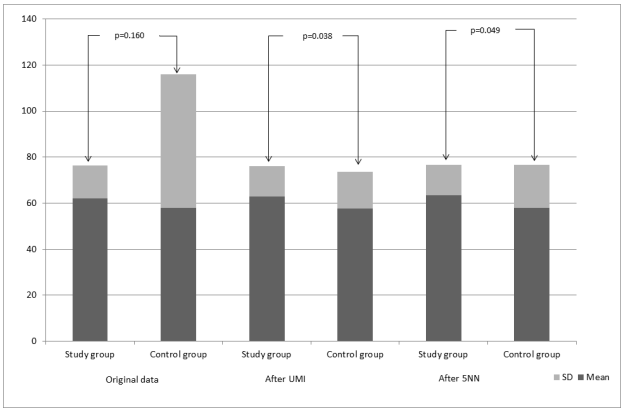


Fig. 2: Differences for HDL in study and control groups before and after performing imputation

dencies between glucose and leptin. These relationships can build the basis for further medical diagnosis and new treatment procedures.

V. Conclusions

Missing values make one of the most common problems for real data collection and extraction in medicine. It is mainly due to the fact, that their presence excludes the intention-to-treat principle and interferes with statistically significant inference. Incomplete data may also refer to other measurements e.g. derived from modern textronic structures used in medicine or protective clothing. Their correct interpretation and analysis ensures reliable operation of intelligent sensors, and as result the entire control system of life signs of the body [41].

Each medical study has its own design, measurement characteristics and different assumptions about missing data mechanisms. Therefore, there is no universal statistical method that deals with missing data, and new investigations should be performed. In this paper, a procedure to improve medical reasoning applied to the problem of discovering new dependencies in the presence of intrauterine growth restriction in children is proposed.

The procedure consists of selecting the imputation technique that results best as applied to the characteristics of data considered and uses the chosen method to impute missing values in data subjected for further analysis. In the empirical test two imputation methods were chosen: unconditional mean and k-nearest neighbor. The statistical analysis of imputed dataset proved to yields more valuable dependencies when compared to original data, maintaining the confidence interval for goodness of fit with the original distribution above 90%. The discovered dependencies in data may establish the basis for new treatment procedures of children with intrauterine growth restriction disorder.

Further studies will involve other medical domains, e.g. monosymptomatic nocturnal enuresis in children where the problem of missing data was encountered [42]. They will also focus on investigating the impact of amounts of missing data on the validity of an imputation technique. Some other methods for dealing with missing values based on rough sets will be used, as proposed by J. Grzymala-Busse et al. [43], [44]. Moreover, the problem of high-dimensional data and feature selection techniques should be considered. More and more data are collected either by interviews, equipment [39] or extraction from text [45], speech [46] or images [47], including medical imaging [48]. In high-dimensional datasets missing data may be more frequent [49] and appropriate feature selection technique [50], [51] may improve the imputation accuracy [10]. Novel solutions of outlier detection based on linguistically quantified statements may be also considered to remove impurities from the data [52].

References

[1] Armijo-Olivo S., Warren S., Magee D. (2009). *Intention to treat analysis, compliance, drop-outs and how to deal with missing data in clinical research: a review*. Physical Therapy Reviews, Vol. 14(1), pp. 36-49, DOI: 10.1179/174328809X405928.

[2] Higgins J. P., Green S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

[3] Lachin J. M. (2000). *Statistical considerations in the intent-to-treat principle*. Contemporary Clinical Trials, Vol. 21(3), pp. 167-189.

[4] Janssen K. J., Donders A. R. T., Harrell F. E., Vergouwe Y., Chen Q., Grobbee D. E., Moons K. G. (2010). *Missing covariate data in medical research: to impute is better than to ignore*. Journal of Clinical Epidemiology, Vol. 63(7), pp. 721-727, DOI: 10.1016/j.jclinepi.2009.12.008.

[5] Farhangfar A., Kurgan L. A., Pedrycz W. (2007). *A novel framework for imputation of missing values in databases*. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 37(5), pp. 692-709, DOI: 10.1109/TSMCA.2007.902631.

[6] Cios K. J., Moore G. W. (2002). *Uniqueness of medical data mining*. Artificial Intelligence in Medicine, Vol. 26(1-2), pp. 1-24.

[7] Kurgan L. A., Cios K. J., Sontag, M., Accurso F. J. (2005). *Mining the cystic fibrosis data*. In: Next generation of data-mining applications, IEEE Press, pp. 415-444.

[8] Klebanoff M. A., Cole S. R. (2008). *Use of multiple imputation in the epidemiologic literature*. American Journal of Epidemiology, Vol. 168(4), pp. 355-357, DOI: 10.1093/aje/kwn071.

[9] Donders A. R. T., Van Der Heijden G. J., Stijnen T., Moons K. G. (2006). *A gentle introduction to imputation of missing values*. Journal of Clinical Epidemiology, Vol. 59(10), pp. 1087-1091, DOI: 10.1016/j.jclinepi.2006.01.014.

[10] Aydilek I. B., Arslan, A. (2013). *A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm*. Information Sciences, Vol. 233, pp. 25-35, DOI: 10.1016/j.ins.2013.01.021.

[11] Farhangfar A., Kurgan L., Dy, J. (2008). *Impact of imputation of missing values on classification error for discrete data. Pattern Recognition*, Vol. 41(12), pp. 3692-3705, DOI: 10.1016/j.patcog.2008.05.019.

[12] Moons K. G., Donders R. A., Stijnen T., Harrell F. E. (2006). *Using the outcome for imputation of missing predictor values was preferred*. Journal of Clinical Epidemiology, Vol. 59(10), pp. 1092-1101, DOI:10.1016/j.jclinepi.2006.01.009.

[13] Li T., Hutfless S., Scharfstein D. O., Daniels M. J., Hogan J. W., Little R. J., Royh J. A., Law A.H., Dickersin K. (2014). *Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus*. Journal of Clinical Epidemiology, Vol. 67(1), pp. 15-32, DOI: 10.1016/j.jclinepi.2013.08.013.

[14] Andridge R. R., Little, R. J. (2010). *A review of hot deck imputation for survey non-response*. International Statistical Review, Vol. 78(1), pp. 40-64, DOI:10.1111/j.1751-5823.2010.00103.x.

[15] Myers T. A. (2011). *Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data*. Communication Methods and Measures, Vol. 5(4), pp. 297-310, DOI:10.1080/19312458.2011.624490.

[16] Joenssen D. W., Bankhofer U. (2012). *Hot deck methods for imputing missing data*. In International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 63-75, DOI:10.1007/978-3-642-31537-4_6.

[17] Zhang S. (2011). *Shell-neighbor method and its application in missing data imputation*. Applied Intelligence, Vol. 35(1), pp. 123-133, DOI:10.1007/s10489-009-0207-6.

[18] Yu Q., Miche Y., Eirola E., Van Heeswijk M., SeVerin E., Lendasse A. (2013). *Regularized extreme learning machine for regression with missing data*. Neurocomputing, Vol. 102, pp. 45-51, DOI:10.1016/j.neucom.2012.02.040.

[19] Van Buuren S., Oudshoorn K. (1999). *Flexible multivariate imputation by MICE*. Leiden, The Netherlands: TNO Prevention Center.

[20] Horton N. J., Lipsitz S. R. (2001). *Multiple imputation in practice: comparison of software packages for regression models with missing variables*. The American Statistician, Vol. 55(3), pp. 244-254.

[21] Zhang P. (2003). *Multiple imputation: theory and method. International Statistical Review*, vol. 71(3), pp. 581-592, DOI:10.1111/j.1751-5823.2003.tb00213.x

[22] Fichman M., Cummings J. N. (2003). *Multiple imputation for missing data: Making the most of what you know*. Organizational Research Methods, vol. 6(3), pp. 282-308.

[23] Zhong M., Sharma S., Lingras P. (2004). *Genetically designed models for accurate imputation of missing traffic counts*. Transportation Research Record: Journal of the Transportation Research Board, vol. 1879, pp. 71-79, DOI:10.3141/1879-09.

[24] Sterne J. A., White I. R., Carlin J. B., Spratt M., Royston P., Kenward M. G., Carpenter J. R. (2009). *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls*. BMJ, vol. 338(b2393), DOI: 10.1136/bmj.b2393.

[25] Gadbury G. L., Coffey C. S., Allison D. B. (2003). *Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF*. Obesity Reviews, Vol. 4(3), pp. 175-184, DOI:10.1046/j.1467-789X.2003.00109.x.

[26] Mahajan, S.D. and Aalinkeel, R. and Singh, S. and Shah, P. and Gupta, N. and Kochupillai, N.: "Endocrine regulation in asymmetric intrauterine fetal growth retardation", Journal of Maternal-Fetal and Neonatal Medicine, 2006, vol. 19(10), pp. 615–623, DOI: 10.1080/14767050600799901

[27] Black, R.E. and Victora, C.G. and Walker, S.P. and Bhutta, Z.A. and Christian, P. and de Onis, M. and et al.: "Maternal and child undernutrition and overweight in low-income and middle-income countries", Lancet, 2013, vol. 382, pp. 427–451, DOI: 10.1016/S0140-6736(13)60937-X

[28] Gürgen, F. and Zeynep, Z. and Füsun, V.: "Intrauterine growth restriction (IUGR) risk decision based on support vector machines", Expert Systems with Applications, 2012, vol.39(3), pp. 2872–2876, DOI: 10.1016/j.eswa.2011.08.147

[29] Bagi, K.S. and Shreedhara, K.S.: "Biometric measurement and classification of IUGR using neural networks", Proceedings of the International Conference on Contemporary Computing and Informatics (IC3I 2014), 2014, pp. 157–161, DOI: 10.1109/IC3I.2014.7019613

[30] Dessi A., Atzori L., Noto A., Visser A. G. H., Gazzolo D., Zanardo V., Magistris A. D. (2011). *Metabolomics in newborns with intrauterine growth retardation (IUGR): urine reveals markers of metabolic syndrome*. The Journal of Maternal-Fetal & Neonatal Medicine, Vol. 24(sup2), pp. 35-39 DOI:10.3109/14767058.2011.605868.

[31] Neitzke U. T. A., Harder T., Plagemann A. (2011). *Intrauterine growth restriction and developmental programming of the metabolic syndrome: a critical appraisal*. Microcirculation, Vol. 18(4), pp. 304-311, DOI:10.1111/j.1549-8719.2011.00089.x .

[32] Zamecznik, A. and Niewiadomska-Jarosik, K. and Wosiak, A. and Zamojska, J. and Moll, J. and Stańczyk, J.: *Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old*, Cardiovascular Journal of Africa, 2014, pp.73–77, DOI: 10.5830/CVJA-2014-009

[33] Niewiadomska-Jarosik K., Zamojska J., Zamecznik A., Stańczyk J., Wosiak A., Jarosik P. (2017). *Myocardial dysfunction in children with intrauterine growth restriction: an echocardiographic study*. Cardiovascular Journal of Africa, Vol. 28(1), pp. 36-39, DOI:10.5830/CVJA-2016-053.

[34] Zamecznik A., Stańczyk J., Wosiak A., Niewiadomska-Jarosik K. (2017). *Time domain parameters of heart rate variability in children born as small-for-gestational age*. Cardiology in the Young, Vol. 27(4), pp. 663-670, DOI:10.1017/S1047951116001001.

[35] Malinowski, A. and Chlebna-Sokół, D.: "Dziecko łódzkie-metody badań i normy rozwoju biologicznego", Ankal, 1998, (In Polish)

[36] Baneshi M. R., Talei A. R. (2010). *Impact of imputation of missing data on estimation of survival rates: an example in breast cancer*. Iranian Journal of Cancer Prevention, Vol 3(3), pp. 127-131.

[37] Luengo J., Garcia S., Herrera F. (2012). *On the choice of the best imputation methods for missing values considering three groups of classification methods*. Knowledge and information systems, Vol. 32(1), pp. 77-108, DOI: 10.1007/s10115-011-0424-2.

[38] Tran C. T., Andreae P., Zhang M. (2015). *Impact of imputation of missing values on genetic programming based multiple feature construction for classification*. In Evolutionary Computation (CEC), 2015 IEEE Congress on, pp. 2398-2405, DOI: 10.1109/CEC.2015.7257182.

[39] Ridgway G. R., Lehmann M., Barnes J., Rohrer J. D., Warren J. D., Crutch S. J., Fox N. C. (2012). *Early-onset Alzheimer disease clinical variants multivariate analyses of cortical thickness*. Neurology, vol. 79(1), pp. 80-84, DOI:10.1212/WNL.0b013e31825dce28.

[40] Pawlak R., Korzeniewska E., Koneczny C., Halgas, B. (2017). *Properties Of Thin Metal Layers Deposited On Textile Composites By Using The Pvd Method For Textronic Applications*. Autex Research Journal. Vol. 17(3). pp. 229-237 DOI: 10.1515/aut-2017-0015.

[41] Korzeniewska E., Walczak M., Rymaszewski J. (2017). *Elements of elastic electronics created on textile substrate*. Proceedings of The 24th International Conference Mixed Design of Integrated Circuits and Systems - MIXDES 2017. pp. 447-450.

[42] Tkaczyk M., Maternik M., Krakowska A., Wosiak A., Miklaszewska M., Zachwieja K., Runowski D., Jander A., Ratajczak D., Korzeniecka-Kozyrska A., Mader-Wolynska I., Kilis-Pstrusinska K. (2017). *Evaluation of the effect of 3-month bladder basic advice in children with monosymptomatic nocturnal enuresis*. Journal of Pediatric Urology. Vol. 13. pp. 615.e1-e615.e6. DOI: 10.1016/j.jpurol.2017.03.039.

[43] Grzymala-Busse J. W., Clark P. G., Kuehnhausen M. (2014). *Generalized probabilistic approximations of incomplete data*. International Journal of Approximate Reasoning. Vol. 55(1). pp. 180-196. DOI: 10.1016/j.ijar.2013.04.007.

[44] Clark P. G., Grzymala-Busse J. W., Rzasa W. (2014). *Mining incomplete data with singleton, subset and concept probabilistic approximations*. Information Sciences. Vol. 280. pp. 368-384. DOI: 10.1016/j.ins.2014.05.007.

[45] Komenda M., Karolyi M., Vyskovsky R., Jezova K., Scavnicky J.(2017). *Towards a Keyword Extraction in Medical and Healthcare Education*. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, pp. 173-176, DOI: 10.15439/2017F351.

[46] Bhaskar J., Sruthi K., Nedungadi P. (2015). *Hybrid approach for emotion classification of audio conversation based on text and speech mining*. Procedia Computer Science, vol. 46, pp. 635-643, DOI:10.1016/j.procs.2015.02.112.

[47] Wojciechowski A., Staniucha R. (2016). *Mouth features extraction for emotion classification*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pp. 1685-1692, DOI: 10.15439/2016F390.

[48] Tomczyk, A. (2014). Detection of line segments. Journal of Applied Computer Science. Vol. 22 No. 2 (2014), pp. 81-90, URL: http://it.p.lodz.pl/file.php/12/2014-2/jacs-2014-2-Tomczyk.pdf

[49] Zaitseva E., Levashenko V., Kvassay M., Deserno T.M. (2016). *Reliability Estimation of Healthcare Systems using Fuzzy Decision Trees*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pp. 331-340, DOI:10.15439/2016F150.

[50] Paja W. (2015). *Medical diagnosis support and accuracy improvement by application of total scoring from feature selection approach*. Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 5, pp. 281-286, DOI: 10.15439/2015F361.

[51] Paja W, Pancerz K. (2017). *Feature Selection Methods Applied to Severe Brain Damages Data*. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, pp. 199-202, DOI: 10.15439/2017F382.

[52] Duraj A., Niewiadomski A., Szczepaniak P. S. (2018) *Outlier detection using linguistically quantified statements*. International Journal of Intelligent Systems. DOI: 10.1002/int.21924

3rd International Workshop on AI aspects of Reasoning, Information, and Memory

THERE is general realization that computational models of human reasoning can be improved by integration of heterogeneous resources of information, e.g., multidimensional diagrams, images, language, syntax, semantics, memory. While the event targets promotion of integrated computational approaches, we invite contributions from any individual areas related to information, language, memory, reasoning.

TOPICS

We welcome submissions of papers on the following topics, without limiting to them, across approaches, methods, theories, and applications:

- Reasoning systems — theories and applications
- Proof systems and model checkers
- Theories of computation and information
- Interactive computation and reasoning
- Computation and reasoning with heterogeneous information
- Space and time in information, language, memory, and reasoning
- Partiality, underspecification, vagueness, and possibilities
- Detection of and reasoning with inconsistency
- Logic and language — approaches, theories, methods
- Computational morphology, syntax, semantics, and interfaces between these
- Constraint-based and type-theoretic approaches and grammars
- Logical approaches to multilingual processing
- Logical and computational foundations in machine learning and information retrieval
- Mathematics for linguistics and cognitive science
- Reasoning, information, and memory in computational neuroscience and life sciences
- Interdisciplinary approaches to information, language, memory, and reasoning

EVENT CHAIRS

- **Grabowski, Adam,** Institute of Informatics, University of Bialystok, Bialystok, Poland
- **Ishihara, Hajime,** Japan Advanced Institute of Science and Technology, Japan
- **Loukanova, Roussanka,** Stockholm University, Sweden
- **Schwarzweller, Christoph,** Institute of Informatics, University of Gdansk, Poland
- **van den Herik, Jaap,** Leiden University, The Netherlands

PROGRAM COMMITTEE

- **Akman, Varol,** Ihsan Dogramaci Bilkent University, Turkey
- **Becerra, Leonor,** Jean Monnet University, France
- **Bekki, Daisuke,** Ochanomizu University / JST CREST, Japan
- **Borgefors, Gunilla,** Uppsala University, Sweden
- **Buszkowski, Wojciech,** Adam Mickiewicz University, Poland
- **Cooper, Robin,** University of Gothenburg, Sweden
- **Hellan, Lars,** Norwegian University of Science and Technology, Trondheim, Norway
- **Jiménez López, M. Dolores,** Universitat Rovira i Virgili, Spain
- **Kerber, Manfred,** University of Birmingham, United Kingdom
- **Kornilowicz, Artur,** Institute of Informatics, University of Bialystok, Poland
- **Litak, Tadeusz,** Informatik 8, FAU Erlangen-Nuremberg, Germany
- **Nemoto, Takako,** School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Japan
- **Nilsson, Jørgen Fischer,** Technical University of Denmark, Denmark
- **Penn, Gerald,** University of Toronto, Canada
- **Retoré, Christian,** Université de Montpellier & LIRMM-CNRS, France
- **Rocha, Ana Paula,** University of Porto, LIACC / FEUP, Portugal
- **Sailer, Manfred,** Goethe-Universität Frankfurt am Main, Germany
- **Salvati, Sylvain,** Laboratoire Bordelais de Recherche en Informatique, Unité Mixte de Recherche CNRS (UMR 5800), France
- **Schwichtenberg, Helmut,** Mathematisches Institut der Universitaet Muenchen, Germany
- **Villadsen, Jørgen,** Technical University of Denmark, Denmark

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 139–143
DOI: 10.15439/2018F236
ISSN 2300-5963 ACSIS, Vol. 15

Adaptive Supervisor: Method of Reinforcement Learning Fault Elimination by Application of Supervised Learning

Mateusz Krzysztoń
Research and Academic Computer Network (NASK)
ul. Kolska 12, 01-045 Warsaw, Poland
Email: mateusz.krzyszton@nask.pl

***Abstract*—Reinforcement Learning (RL) is a popular approach for solving increasing number of problems. However, standard RL approach has many deficiencies. In this paper multiple approaches for addressing those deficiencies by incorporating Supervised Learning are discussed and a new approach, Reinforcement Learning with Adaptive Supervisor, is proposed. In this model, actions chosen by the RL method are rated by the supervisor and may be replaced with safer ones. The supervisor observes the results of each action and on that basis it learns the knowledge about safety of actions in various states. It helps to overcome one of the Reinforcement Learning deficiencies – risk of wrong action execution. The new approach is designed for domains, where failures are very expensive. The architecture was evaluated on a car intersection model. The proposed method eliminated around 50% of failures.**

I. Introduction

REINFORCEMENT Learning (RL) is a popular approach for solving increasing number of problems. In contrast to Supervised Learning (SL) this type of learning does not require any training data or teacher with prior knowledge. Instead, experimenting with the environemnt is performed to generate knowledge. RL has been successfully used to solve problems in multiple domains: robotics and control [1], game playinng [2] and power systems [3], just to name a few.

However, RL has many deficiencies that hinder applying it to the complex real word problems (e.g. unscalability [4], small data efficiency [5] and low human-readability of generated knowledge). Another defficiency is a risk of failures while searching optimal solution by experimenting with the environment, which requires taking random decisions from time to time. For many domains such risk is justified. However, in some domains any failure can be expensive (e.g. robot control). Thus, chance of failure should be minimized, even at the expense of the exploration. Multiple safe exploration techniques for RL were already proposed in literature [6], [7]. Most of these approaches assumes, that some prior external knowledge exists and can be used in early steps of exploration to avoid failures. However, this assumption is not always valid. Hence, need for techniques that limit risk of failures with no prior knowledge arises. It should be emphasized, however, that all failures in the exploration phase can be eliminated only if a prior knowledge is incorporeted [8].

In the literature multiple successful approaches for combining RL with Supervised Learning (SL) in form of *hybrid methods* were proposed to address various deficiencies of RL [9]–[11]. However, to the best to the Author's knowledge, no hybrid method dedicated to increasing safety of exploration has been proposed yet. In this work such approach by introducing the Adaptive Supervisor to support RL method is proposed. Adaptive Supervisor use SL approach to create knowledge about risky actions and observes states and actions that led to failures during exploration to create training set. The supervisor learns online so it can support RL and limit failures in an early phase of exploration.

The article is organized as follows. Firstly, various concepts for increasing safety of exploration are discussed. Then the novel Reinforcement Learning with Adaptive Supervisor architecture is introduced and the realisation of this architecture is proposed. The approach was verified in SInC domain [12]. Finally, results are presented and discussed. Additionally the knowledge generated by SL is verified.

II. Related Research

The comprehensive survey on Safe Reinforcement Learning can be found in [6]. The survey was recently extended in work [7]. Safe Reinforcement Learning methods can be divided into two groups. The first one involves modifying the risk-neutral optimization criterion to address possibility of failures. The modification can involve adding constraints (based on the external knowledge), optimizing performance for the worst scenario (in case the process is stochastic) or adding factor that makes safe policies more preferable over risky ones (e.g. ones with smaller variability of observed rewards). In the second group the optimization criterion remains risk-neutral, instead the exploration process is modified to avoid failures. Methods in this group can be further divided into methods that incorporates external knowledge (in the form of constraints, a set of demonstrations, a teacher that guides or supervise learning process, initial policy, etc.) and those in which exploration is directed to less risky areas by additional mechanism.

However, none of the presented works verifies possibility to increase safety of exploring process with on-line Supervisor (with no initial knowledge).

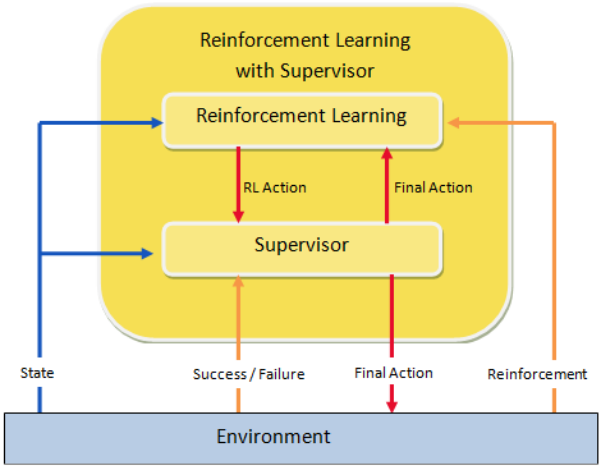


Fig. 1: General scheme of RL with Supervisor (RLS) architecture

Supervisor can be defined as additional mechanism, which role is to support RL in making better and safer decisions or in learning faster. The idea was firstly introduced in [14], where the supervisor is a set of tips, defined by expert. An idea for incorporating the supervisor (this time implemented with SL) to RL was described in [16]. The proposed solution is based on the assumption, that in some domains significant number of available actions in the given state are useless, so there is no sense in performing them. The role of the supervisor is to suggest small set of relevant actions for the given state.

In works [12], [13] SL is used to reduce state space of RL, thus increases efficiency of learning. This *hybrid method* (RLSR) is based on the assumption that two states are similar, if achieving them produces similar reward. The state is described by a set of attributes. The part of the state that is to be reduced is passed to a classifier. The classifier reduces that part of state to a single attribute. That single attribute and not reduced part of state are passed to RL module together as *reduced state*. Based on the *reduced state* the RL module chooses next action to perform. After the action a reinforcement is delivered both to the RL module and classifier for the learning purpose. The approach was tested in two domains: "Hunt the Wumpus" and "Hunter, Preys and Predators", which is the extended version of "Predators and Prey" problem. In the most of cases the conducted experiments proved the approach to be successful in shortening time necessary for learning the best solution, comparing with Q-learning. The proposed method performs well with noisy data.

III. Reinforcement Learning with Adaptive Supervisor – dealing with a risk of failures

The lack of research on applying supervised learning to implement the Supervisor concept was inspiration for developing a novel approach. The Reinforcement Learning with Adaptive Supervisor combines RL with supervisor implemented according to the SL approach.

A. General idea

To address the risk of failures the approach for combining RL with the adaptive supervisor is proposed (RLS). The architecture of the approach is presented in Fig. 1. The concept of the supervisor in this approach is similar to the one proposed in [15], where guard with explicit constraints is introduced. However, in the RLS the supervisor's knowledge is being created simultaneously with RL component. Each time the RL component chooses an action to perform, the selected action is rated by the supervisor. If the supervisor concludes, that the action chosen by RL is not safe enough (may lead to a failure) the supervisor overrides the action with the safest one (according to its current knowledge). The Supervisor observes the result of each action and on that basis it creates the knowledge about the safety of each action in each state in which that action can be preformed.

B. Realization

The proposed architecture was applied for the case where RL is implemented with the hybrid version (RLSR). This version accelerates learning, but the trade off is potentially worse quality of decision. Hence, a supervisor is introduced to minimize number of failures. Scheme of the method is presented in Fig. 2.

The supervisor is implemented as a classifier. In the process of learning the classifier receives training examples in the form: $\langle s, a, e \rangle$, where s is not reduced state, a is performed action and e is evaluation of performing the action a in the state s. e takes one of the following values: *good* or *bad*.

Asking the supervisor if the action is correct corresponds to classifying pair $\langle s, a \rangle$ as *good* or *bad*. If the action is classified as *bad* the supervisor iterates over all possible actions in state s and chooses the action with the highest certainty of being classified as *good* (the safest action). The chosen action is performed and sent back to the RL component as feedback.

To teach the supervisor rating actions, the supervisor has to store all examples gathered during learning process. As the

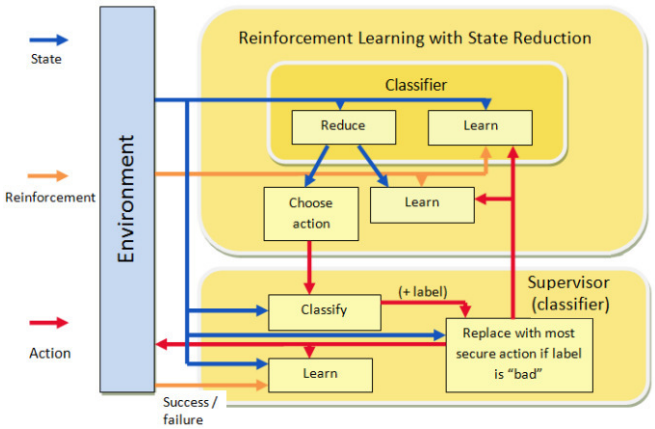


Fig. 2: Realization of RL with Supervisor architecture. First action is chosen by RL with State Reduction (RLSR). Then Supervisor rates action and replace it only if rate is *bad*.

supervisor gets information from the environment that success or failure happened, configured number of last stored examples should be labeled as *good* or *bad* respectively and added to the training set. Additionally weights of examples can be introduced — the closer example is to failure/success state, the bigger weight it should have.

The classifier used in this hybrid method should provide certainties of labels. Hence, rigor r of the supervisor can be configured. To classify action as *good* the certainty of label *good* has to be higher than r. Otherwise, action is classified as *bad*, because there were enough cases (according to the given rigor) which led to failure immediately or few steps after taking this action. It is worth noting, that too high rigor may led to rejecting action that is potentially good (there exists the list of actions following it that lead to success), but in the past that action preceded the incorrect actions that resulted in failure and hence is considered wrong.

C. Experimental domain

The approach was examined in the domain of crossing intersection by autonomous vehicles (SInC) [12]. Crossing intersection is simultaneous, therefore collisions can occur. To avoid them vehicles have to adjust their speed. In the same time vehicles should cross intersection as fast as possible. Hence, in every step of simulation an agent steering car chooses action $a_{\mathrm{i}} \in A$, which corresponds to changing speed by i[3]. The set of actions A is defined as $A := \{-a_{max}, -a_{max}+1, \ldots, a_{max}-1, a_{max}\}$, where a_{max} is the maximal speed change. In case as a result of taking action a_{i} the value of speed should be smaller than 0 or greater than v_{max}, then the speed value is set to 0 or v_{max}, respectively.

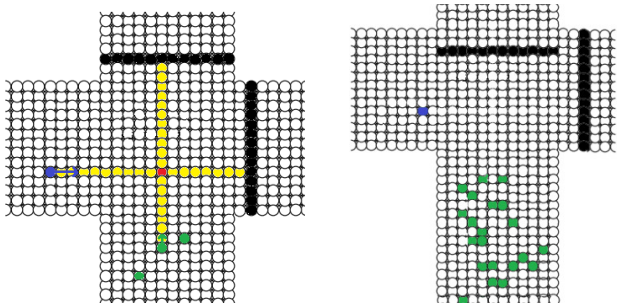
To choose an action the agent can use following information, updated in every step:

- distance to the end of intersection (d_t);
- speed of the car (v_c);
- distance between car and the nearest collision point with a car coming from the crossing road (collision car) (d_{cp});
- distance between the collision car and the collision point (d_{ccp});
- speed of the collision car (v_{cc}).

Below state s is defined as $s = \langle d_t, v_c, d_{cp}, d_{ccp}, v_{cc}\rangle$. In Fig. 3a an exemplary $s = \langle 18, 2, 10, 8, 1\rangle$ is presented.

After each step the agent observes the result of it's action. If the car reached the end of intersection successfully or collision occurred all stored pairs of states and actions ($s = \langle s, a\rangle$) are labeled by the Adaptive Supervisor as $e = good$ or $e = bad$, accordingly. Otherwise (car is still crossing intersection safely) the Adaptive Supervisor continues storing examples.

[3]the decision to change speed influences the speed of the car in the step following the step in which the decision was taken — e.g. if in the step t the speed of the car was equal to 1 field per step and the decision of the agent in that step t is to reduce speed by 1 field per step, than the car will change its position by one field in the step t and than stop (in the step $t+1$ the car won't change its position regardless the decision in the step $t+1$). Postponing result of the speed change makes the considered domain more challenging and realistic (gap between making observation and changing speed is taken into account).



(a) (b)

Fig. 3: (a) The example state at an intersection. Point where collision between the blue car (intelligent one) and the nearest moving car can happen is marked with a red color. With black line is the end of intersection (target); (b) Initial situation at intersection for experiment [12].

D. Verification scenario

Three approaches for taking decision were compared — standard RL, hybrid method RLSR, and the novel hybrid approach RLS (version described in III-B).

The first method (QRL) was based on the standard RL method (Q-learning with ϵ-greedy policy). The following rewards were introduced:

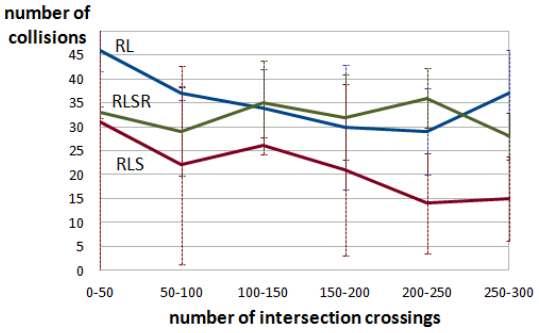
- r_c — negative reward for collision ($r_c = -100$)
- r_t — positive reward for reaching target ($r_t = 100$)
- r_s — negative reward for making move ($r_s = -2$)

Rewards r_t and r_s promote crossing intersection as quickly as possible. Reward r_c teaches the agent to avoid collisions. To implement Q-learning RLPark library was used ($\alpha = 0.25$, $\lambda = 0.4$, $\gamma = 0.55$, $\epsilon = 0.5$). All parameters were chosen with the hill climbing approach.

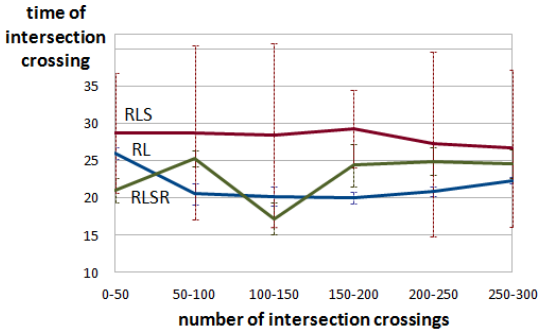
As the second method RL with state reduction with classifier (RLSR) was used. To detect similar states, the state attributes which corresponds to the possibility of collision (v_c, d_{cp}, d_{ccp}, v_{cc}) were reduced to a single bivalent attribute. The value of this attribute can be interpreted as possibility of collision in the given state. Classifier was implemented with C4.5 algorithm supplied by WEKA library. As the RL method the QRL was used.

The third approach was RLS method — to RLSR method supervisor was introduced to increase safety. For the Adaptive Supervisor success was defined as crossing intersection safely (getting to the end of intersection without collision). Any collision during crossing is considered as a failure. To implement the supervisor C4.5 algorithm was used. The rigor r of the supervisor dynamically changes with the development of agent's knowledge and is given for ith simulation of the experiment with formula:

$$r = \begin{cases} 0.5, & i < 10, \\ 0.66, & 10 \leq i < 200, \\ 0.75, & 200 \leq i < 300. \end{cases} \quad (1)$$



(a)



(b)

Fig. 4: A relationship between (a) collision number and (b) intersection crossing time and the agent experience — moving average with window size equal to 50 crossings of intersection.

The performance of methods was compared in the experiment conducted with platform MABICS [12]. The initial situation on the intersection for this experiment is shown in Fig. 3b. On the left side of the intersection a vehicle controlled by the intelligent agent is placed. On the bottom there are 21 vehicles that simulate real traffic. Each car moves with random, constant speed. Maximal speed change $a_{max} = 1$ and maximal speed $v_{max} = 3$. The experiment was repeated three times only, because of efficiency issues of the intersection crossing simulator, integrated with the MABICS. Each experiment consisted of 300 simulations. Between simulations within one experiment all gained experience of the agent was persisted.

E. Experimental results

Obtained average results of crossing time and collision numbers for all three methods are presented in Fig. 4. The RLSR method speeds up learning comparing to RL, but the number of collisions in the last periods is similar. Both RL and RLSR methods have difficulty in exploration of so complex domain, therefore only suboptimal solutions were found. The proposed RLS hybrid method proved to be safer in comparison with both standard RL and RLSR methods ($p < 0.05$ according to t-test[2]). RLS causes about 50% less collisions in the last periods of learning, which is satisfactory in so complex domain.

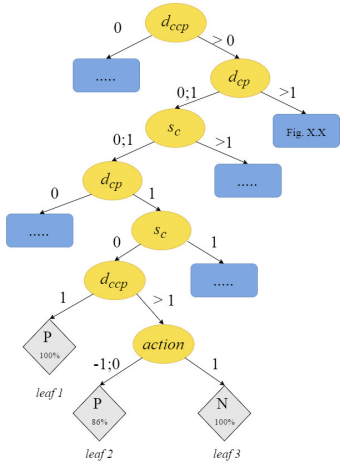
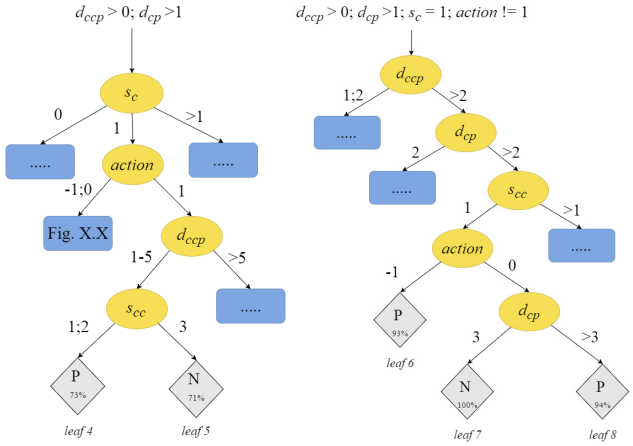


Fig. 5: Decision tree for classifying the given action in the given state as *good* or *bad* — root part



(a) (b)

Fig. 6: Decision tree for classifying the given action in the given state — (a) branch for $d_{ccp} > 0$ and $d_{cp} > 1$ and (b) branch for $d_{ccp} > 0$, $d_{cp} > 1$, $s_c = 1$ and $a < 1$

The decision tree which represents the knowledge created by the supervisor in the end of one of the experiments is presented in Fig. 5 and Fig. 6. Since the obtained tree has about 100 nodes only the most interesting branches are shown in details. With each leaf one rule for accepting or rejecting given action a in the given state(s) can be associated.

In some states any action chosen by the agent is accepted by the supervisor since collision cannot happen after visiting this state (according to the supervisor's experience). The example of such situation is represented by *leaf 1* (Fig. 5). The set of states associated with this leaf is illustrated in Fig. 7a. In this situation the speed of the car controlled by the intelligent

[2]the one sided t-test with equal variance. The equality of variance was verified with the two-sample F-test

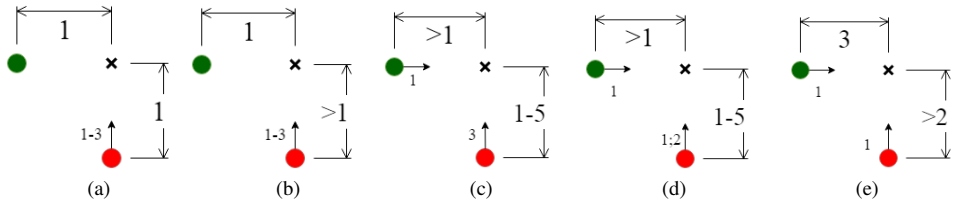


Fig. 7: Illustration of states associated with some leaves of decision tree, respectively (a) *leaf 1*, (b) *leaf 2* and *leaf 3*, (c) *leaf 5*, (d) *leaf 4*, (e) *leaf 7*. As the tree generalize examples each leaf represents set of similar states, e.g. (a) represents states satisfying $s_c = 0$, $s_{cc} \in \{1, 2, 3\}$, $d_{cp} = 1$ and $d_{ccp} = 1$.

agent is equal to zero and distance to collision point of both cars is equal to 1. Accepting each decision by supervisor is correct as the speed of collision car (v_{cc}) is greater than zero (constraint) — even if the agent decides to accelerate the car will move to the collision point in step $t+2$ whereas collision car even in the most pessimistic case ($v_{cc} = 1$) will move to collision point in step $t + 1$ and will leave this point in step $t + 2$.

Similar situation is associated with *leaf 2* and *leaf 3*. (Fig 5), but here the distance of collision car to the collision point (d_{ccp}) is greater than one (Fig. 7b). In this case accelerating is not safe anymore and supervisor will correctly reject decision to increase the speed by one and replace the action with decision to sustain speed or decrease it by one.

leaf 4 and *leaf 5* (Fig. 6a) show situation when the same action can be accepted or rejected depending on only one attribute of the state s, in this case v_{cc}. If $v_{cc} = 3$ (Fig. 7c) the supervisor reject the decision to accelerate. If the value of v_{cc} is smaller (the collision car approach to collision point slower, Fig. 7d) the decision to increase speed will be accepted as long as rigor r is smaller than 73%.

Finally, it is presented how the supervisor selects the best action in case of rejection, taking interesting rule as an example. The rejection rule is associated with *leaf 7* (Fig. 6b) and is illustrated on Fig. 7e. If the selected action is maintaining current velocity ($a = 0$) the supervisor rejects it. For simplification let choose as current state one that match the rule: $d_{ccp} = 3$, $d_{cp} = 3$, $v_c = 1$, $v_{cc} = 1$. Then, to select new action, the tree is searched for $a = 1$ and $a = -1$. For $a = -1$ the *leaf 6* (Fig. 6b) with label P and certainty equal to 93% is found, whereas for $a = 1$ the *leaf 4* (Fig. 6a) with label P and certainty 73%. As both leaves has label P the action with greater certainty is chosen ($action = -1$.). The action will cause that car will stop before reaching collision point, which is the safest option.

IV. Conclusions

This paper presents how Supervised Learning concept can be used to improve safety of the exploration process in RL, when no prior knowledge is available. The proposed method reduces number of failures, that are usually result of the space search, unavoidable part of RL method. Hence, the method should be used in domains where failures are expensive or even intolerable. The conducted experiments made it possible to verify new idea as promising since adding the Adaptive Supervisor eliminated around 50% of failures.

The presented RLS concept should be further examined in various domains, especially ones in which failures are costly and standard RL methods or hybrid methods (e.g. RLSR) are able to find good solution (it is not a case of the SInC domain) to analyze how many failures can be avoided additionally.

References

[1] Fangyi Zhang, Jürgen Leitner, Michael Milford, Ben Upcroft, and Peter Corke. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv:1511.03791*, 2015.

[2] Guillaume Lample and Devendra Singh Chaplot. Playing fps games with deep reinforcement learning. In *AAAI*, pages 2140–2146, 2017.

[3] Mevludin Glavic, Raphaël Fonteneau, and Damien Ernst. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-PapersOnLine*, 50(1):6918–6927, 2017

[4] Reid, M. Ryan, M. R. K.: Using ILP to Improve Planning in Hierarchical Reinforcement Learning. In: Proceedings of the 10th International Conference on Inductive Logic Programming (ILP '00). Springer-Verlag, London, UK, pp. 174-190, 2000. DOI:10.1007/3-540-44960-4_11

[5] Fachantidis, A.,Partalas, I.,Tsoumakas G.,Vlahavas, I.: Transferring task models in Reinforcement Learning agents. Neurocomput. 107, pp.23-32. May 2013. DOI: 10.1016/j.neucom.2012.08.039

[6] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[7] José M Faria. Machine learning safety: An overview. 2018.

[8] Javier Garcia and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564, 2012, DOI: 10.1613/jair.3761

[9] Uther, W. T. B.—Veloso, M. M.: Tree based discretization for continuous state space reinforcement learning. Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence (AAAI '98/IAAI '98), American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 769-774, 1998

[10] Kalyanakrishnan, S.—Stone, P.—Liu, Y.: Model-Based Reinforcement Learning in a Complex Domain, RoboCup 2007: Robot Soccer World Cup XI, Springer-Verlag, Berlin, Heidelberg, 2008

[11] Henderson, J.,Lemon, O.,Georgila, K.: Hybrid reinforcement/supervised learning for dialogue policies from communicator data, IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2005

[12] Krzysztoń, M. Sniezynski, B.: Combining Machine Learning and Multi-Agent Approach for Controlling Traffic at Intersection. Computational Collective Intelligence, Springer, 2015, pp 57-66

[13] Wiatrak, Ł.: Hybrid Learning in agent systems, Master Thesis, AGH University of Science and Technology, Cracow, 2012 (in Polish)

[14] Maclin, R.—Shavlik, J. W.: Creating advicetaking einforcement learners, Machine Learning, 22((13)): pp. 251-281, 1996

[15] Benbrahim, H., Franklin, J. A.: Biped dynamic walking using reinforcement learning. Robotics and Autonomous Systems,22,pp.283-302, 1997.

[16] Cetina, V.U.: Supervised reinforcement learning using behavior models, Machine Learning and Applications, 2007. ICMLA 2007, pp.336-341, 13-15 Dec. 2007. DOI: 10.1109/ICMLA.2007.14

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 145–153
DOI: 10.15439/2018F248
ISSN 2300-5963 ACSIS, Vol. 15

Combining the Syntactic and Semantic Representations of Mizar Proofs

Karol Pąk
University of Białystok,
Ciołkowskiego 1M, 15-245 Białystok, Poland
Email: pakkarol@uwb.edu.pl

***Abstract*—The Mizar system provides two representations of the proofs present in its library. The syntactic representation preserves the human-friendly rich Mizar language, where the meaning of structures and expressions is still influenced by their context. The semantic one, on the other hand, explicitly reflects the meaning of all elements present in the proof scripts, however many features of the Mizar language are eliminated.**

In this article, we overcome the limitations of both representations of proofs, by proposing a method combining them. We show that we can simultaneously maintain the richness of the language and provide access to the derived proof information. We discuss how such combined information closer corresponds to that present in other proof assistant languages, for example that of Isabelle/Isar.

I. Introduction

ONE of the most recognizable features of the Mizar system [1] is its highly human-oriented proof environment. For over four decades the Mizar project has developed an environment that allows to create formal reasoning similar to how it is done in informal mathematical practice. Many actions at various levels have been taken to ensure this environment:

- Linguistically motivated dependent soft type system that closely reflects how most of the mathematicians use mathematical objects and how they categorize them.
- Rich expression language that provides complex symbol overloading and notational constructs, e.g., the meaning of individual symbols is strongly influenced by the context, as in textbooks.
- The Mizar natural-deduction proof that try to approximate the way how informal proofs are written, where the deduction is not steered by an explicate list of rules that indicates how to match a statement to its proof.

These possibilities are reflected in the size of one of the largest formal libraries, the Mizar Mathematical Library (MML) [2] that includes many domains that have not been formalized elsewhere. However, the Mizar system is the only tool that can fully operate on the content of the library written in such rich language.

a) Related works: Obviously, there have been a number of attempts to explore the content of the MML by external tools. Most of them ensure access to the semantic representation of MML where many proof details are easily accessible for a human reader, starting with easy semantic searching tools as MML Query [3], variants of XML format [4], MMT logical framework [5] or provide the extensive theorems database of MPTP [6] that can be served as a tool in automated theorem proving and machine learning [7].

The Mizar system also provides two ways of access to the syntactic representation. These are *Weakly Strict Mizar* (WSX) [8] and *More Strict Mizar* (MSX) [9], where MSX is an extension of WSX. However, both human-friendly syntactic representations are still too far from semantic one and selecting access, we have to choose between these two.

A lot of work has been done to cross-verify MML by external tools that struggle with many problems such as the Mizar logic that goes a little bit beyond the first-order logic and the Mizar type system. Pioneering and the largest translation of MML to the TPTP untyped first-order language has been done by Urban [10], where higher-order problems have been verified with an extension of TPTP language [11]. Kunčar [12] has attempted to translate the MML as transparent higher-order logic theories, however, his concept was not able to cover more advanced features of the Mizar type system. Successful attempt to extract and cross-verify higher-order problems using higher-order automated theorem provers Satallax and LEO-II have been done by C. Brown [13]. However, these translations are fixed from the point of view of a further development of the MML, i.e., they do not allow any modification or further development.

Urban's MizAR system [14] facilitates the search for a given justification using machine learning and automated reasoning. It returns a list of premises needed for the current goal, which in most cases can be directly used in the script, as in the case of justification generated by the Sledgehammer subsystem for Isabelle/HOL [15]. However, in some cases to use obtained list of premises, authors need to carefully modify the

The paper has been supported by the resources of the Polish National Science Center granted by decision n°DEC-2015/19/D/ST6/01473.

Mizar proof script `environ` and/or create manually a few auxiliary steps.

b) Isabelle/Mizar: is a project whose goals are to: (1) specify the Mizar foundations fully formally in the Isabelle logical framework [16] and (2) cross-verify all the proofs in MML in resulting fully specified logic. The work on these can in the longer term lead to (3) the creation of a Mizar like environment in the Isabelle logical framework, which would provide Mizar foundations and its various mechanisms and allow users to further develop proofs in typed set theory using various Isabelle mechanisms.

There has been a lot of progress on the first goal. The environment [17], [18] has an equivalent of the Mizar dependent type system including Mizar-like structures [19], as well as higher-order concepts, such as set comprehensions and schemes [20]. We have recently developed an automated translation of the statements; however, a large gap between the syntactic and semantic Mizar representations significantly hinders work on an automatic export of the MML proofs. The proofs should be as close to the syntactic Mizar representation as possible, but at the same time, we need to use the semantics, e.g., to identify objects and reconstruct their lists of hidden arguments.

c) Contribution: This paper introduces a representation of Mizar proofs that combines the syntactic and semantic ones. Most of the original syntax is preserved, annotated by the information how particular constructs have been interpreted by Mizar, which allows the processing of the proofs by external systems. The particular contributions are:

- An elegant presentation of Mizar proofs using implicit types assigned to frequently used variable names. This combines the syntactic and semantic representation of the Mizar reserve mechanism.
- A simplification of the Mizar proof outlines to the fix–assume–show outlines present in most declarative proof languages, e.g., Isabelle/Isar [21]. This involves eliminating several constructions, such as the take steps that indicates a suitable term for instantiating an existentially quantified thesis. In particular the concept of the take step has no equivalent in most other proof languages [21].
- Lists of dedicated rules that determine the reasoning patterns for each sub-proof with the fix–assume–show proof outline present in most declarative proof languages. We propose Isabelle/Isar rules as an example.
- A human-friendly rendering of our Mizar representation.

The paper is constructed as follows. In Section II we discuss the Mizar proof concepts lost in the Mizar semantic exports. In Section III we briefly describe a method that matches syntactic and semantic representations. Then we present a reconstruction of these concepts based on the matched representations. In Section IV we present a way to simplify Mizar proofs to logical framework formats preserving reconstructed concepts. Finally in Section V we show an example formalization created in the Isabelle/Mizar environment that uses a rebuilt Mizar proof.

II. Mizar Semantic Export

The scope of this paper does not allow to fully explain the details in the Mizar system including the semantic representation (for such details see [22], [4]). We will point out the main problems and solutions using a single example that states a natural basic property of the set inclusion (Fig. 1). Additionally, the deduction used there as the property justification is quite similar to *informal* one and can be analyzed by a reader who does not have experience with the Mizar system.

a) Hidden Types and Quantifiers: According to the Mizar syntax each statement has first-order form where atomic predicative formulas are combined with classical logic connectives and quantifiers. Also note, that each quantifier has to associate a Mizar type (types correspond to first-order predicates) with all bounded variables. However, the Mizar Analyzer can not infer this type automatically based on the context of the variables as is the case for most type-theory based systems, since Mizar disambiguates the meaning of each symbol based on types of its arguments (corresponding to type inference in an intersection-type system). To avoid specifying *explicitly* the types of all bound variables at their quantifiers, the Mizar system provides a *reservation* mechanisms that allows global associating variable names with their types. For example, `object` that is the type of the variable `x` is not mentioned in the sentence ex `x` st in Fig. 1 but is imported from the reservation reserve `x,y` for `object`.

The type of a reserved variable can be skipped not only in the declaration of a quantifier, but also in each Mizar construction which requires it (compare lines 7 and 11). Moreover, for convenience, universal quantifiers that bind reserved variables can be implicit. Additionally, in many systems, free variables that correspond to *implicit* quantifiers in a given statement are automatically introduced to a sub-deduction that justifies the statement. Obviously such user support is welcome in the human-readable export, especially since the equivalent of the Mizar reserve mechanism has already been provided in Isabelle/Mizar. Unfortunately, Mizar reconstructs these quantifiers and corresponding introduction steps very early, between WSX and MSX representations and does not distinguish reconstructed objects in the XML semantic representation[1]. As a consequence, there is no difference in the semantic representation between the theorem presented in Fig. 1 and Fig. 2. On the other hand, we can easily distinguish this information in the WSX representation, but based on this representation we can not provide enough information

[1]The semantic representation of the theorem presented in Fig. 1 where hidden types and quantifiers are fully reconstructed is available in the HTMLization of the current Mizar library http://mizar.uwb.edu.pl/version/current/html/xboole_0.html#t8

```
1  reserve x,y for object,
2          X,Y for set;
3  theorem
4    X c< Y implies ex x st x in Y & X c= Y\{x}
5  proof
6    assume A1: X c< Y;                                  thesis: X c< Y implies ex x st x in Y & X c= Y\{x}
7    then consider x such that                                 thesis: ex x st x in Y & X c= Y\{x}
8    A2: x in Y and A3: not x in X by Def8,TARSKI:def 3;
9    take x;                                                 thesis: x in Y & X c= Y\{x}
10   thus x in Y by A2;                                            thesis: X c= Y\{x}
11   let y;                                                  thesis: y in X implies y in Y\{x}
12   assume A4: y in X;                                            thesis: y in Y\{x}
13   then y<>x by A3;
14   then A5: not y in {x} by TARSKI:def 1;
15   X c= Y by A1;
16   then y in Y by A4;                                                  thesis: verum
17   thus then thesis by Def5,A5;
18 end;
```

Fig. 1. An example Mizar style theorem, originally occurring as XBOOLE_0:8 (eighth theorem in the Mizar proof scripts XBOOLE_0), with *implicit* thesis explicitly shown. The theorem states that if X is a proper subset of Y (X c< Y), then there exists a member x of Y (x in Y) for which X is a subset of the complement of the singleton {x} in Y.

```
1  theorem
2    for X,Y being set st X c< Y holds
3      ex x being set st x in Y & X c= Y\{x}
4  proof
5    let X,Y be set;
6    assume A1: X c< Y;
```

Fig. 2. An example of the semantically equivalent formulation of the theorem presented in Fig. 1 together with a fragment of its justification.

to explore the MML. Note that information about disambiguating symbols and their hidden arguments are missing there and are very hard to reconstruct for any external tool.

b) Normal Form: The reservation system is one of the three main reasons why we are forced to combine information from syntactic and semantic representations. To simplify the grammar of the semantic representation, the Mizar Analyzer also transforms each formula to the Mizar normal form (MNF) which uses only selected logical connectives, such as $\neg$, a generalization of $\wedge$ for n-arguments, the universal quantifier $\forall$ and $\bot$. A lot of work has been done by Urban to minimize the consequences of normalization [4]. He built directly into the Mizar Analyzer a hint system that is visible as an additional attribute pid in selected nodes of the XML semantic representation. This system should allow the reconstruction of the original formula from the normalized one based on pid-s, e.g., every implication $\alpha \rightarrow \beta$ is replaced by the formula $\neg_{-4}(\wedge_{-5}(\alpha, \neg_{-6}(\beta)))$ where subscripts represent pid–values. However, these hints are often lost in the normalization process, since, e.g., the Analyzer eliminates double negation together with the corresponding pid-s. In consequence, there are several cases (a few percent of the library)[2], where the original formulation is different than the HTMLization generated with the pid-support even if we omit reconstructed hidden quantifiers.

It is important to note that equivalent reformulation of statements in a Mizar proof script does not affect its correctness, since most of Mizar verifier's modules are based on the MNF including the Reasoner which check the applicability of *Skeleton steps* – discussed in Section IV that operates on the thesis in a given proof. For comparison, equivalent reformulation of statements is not possible in most declarative proof languages, as all reasoning pattern must precisely correspond to the related statements and proofs.

c) More advanced pid*–problems:* The reconstruction of the original logical conjunctions and quantifiers was one of the additional tasks in the pid system created by Urban. His system mainly focused on solving the conflict between *patterns* and *constructors*, that we only sketch here (for more details see [10]). Note that every Mizar defined object (i.e, a function, a type, a predicate, an attribute or a structure) together with its list of arguments with their types (and a result type if applicable) and positions of each visible arguments constitutes the Mizar pattern. Mizar constructors are just absolute identifiers (in the environment of a given article) of Mizar objects of which the patterns are translated during their full identification by the Mizar Analyzer. Obviously, a given overloaded pattern can be translated into different constructors. Unfortunately, different patterns can be translated into the same constructor, since, e.g., synonyms and antonyms of predicates and adjectives inherit the constructor from their ancestor. To ensure full control over the many-to-many

[2] For example, the ZFMISC_1 article contains 139 theorems and the differences occur in 8 cases, i.e, theorems: 6, 19, 22, 37, 53, 58, 112, 138.

relationship, the number of constructor (nr) and pattern (pid) should be associated with each object. However, many pid-s are lost or arguments of patterns are incorrectly reconstructed. We can observe this in the HTMLization as *technical* constructors rather than the corresponding patterns or as missing values in the lists of arguments[3]. Such defects do not have a significant negative impact for the rendered HTML, but are unacceptable in every cross-verification of the MML.

III. Disambiguated Syntactic Mizar Export

The problems indicated in Section II are typical, if we want to obtain access to the MML based only on the semantic representation. Therefore we develop an application that combines the semantic and syntactic informations. Obviously there are many inconveniences in this approach, since the semantic representation is completely rewritten by the Mizar Analyzer with respect to the syntactic one and contains only the information that are absolutely necessary for the checking proof steps. We combine the information in three stages.

a) Top level: First, we match all the items from the two representations, where often a few semantic steps correspond to a single syntactic one. For example, a step that introduces variables X, Y in Fig. 2 (see 5th line) is hidden in Fig. 1, but for semantic comparison, both variables are introduced in independent steps that are followed by additional steps where the corresponding modified thesis is formulated.

b) Logic connectives and quantifiers: Next we match syntactic and semantic representation of each statement to find corresponding atomic formula. For this purpose we transform every syntactically represented statement imitating the Mizar Analyzer process such as the normalization and then we compare the obtained formula with the corresponding semantic representation of this statement. We transform also all formulas into a system of abstractions and applications in meta logic

```
<logic id=("ball"|"hidden_ball"|"bex"|"iff"|
  "impl"|"or"|"and"|"not"|"False"|"True")\>
```

that should be easy-to-read by external tools, since e.g., our system directly corresponds to logical framework application and abstraction. Note that the constant hidden_ball is semantically equivalent to ball but corresponds to a universal quantifier that is originally hidden. Additionally, we distinguish types of variables that are imported form *reservations*. For example, hidden quantifiers that bind variables with types imported from *reservations* in the statement of theorem presented in Fig. 1 obtained the following our representation:

[3] See for example http://mizar.uwb.edu.pl/version/current/html/pboole.html#CC4 where the expression includes V8 rather than non-empty, V9 rather than empty-yielding, and the argument A is missing in the type ManySortedSet of A.

```
<proposition label="xboole_0_th_8">
 <app>
  <logic id="hidden_ball" type="o" args="2"
     argsType="ty_abs"/>
  <ReservationType id="X">
   <const id="HIDDENM2" type="ty" args="0"
      argsType="set"/>
  </ReservationType>
  <abs id="X" type="set" args="0">
   <app>
    <logic id="hidden_ball" type="o" args="2"
       argsType="ty_abs"/>
    <ReservationType id="Y">
     <const id="HIDDENM2" type="ty" args="0"
        argsType="set"/>
    </ReservationType>
    <abs id="Y" type="set" args="0">
    ...
```

where the constant HIDDENM2 corresponds to the Mizar set that is the second type definition (called mode in Mizar) in article HIDDEN. Note that the name directly corresponds to the absolute constructor name proposed in [10] and the OMDoc node <OMS module="HIDDEN"name="M2"/> (according to the
naming scheme proposed in [5]).

c) Atomic propositions: Matching at the atomic proposition level is quite natural. Generally, we just match predicates and then recursively terms and subterms to disambiguate them. However, we have to take into account Mizar local abbreviations that are fully unfolded in the semantic representation. It is also important to note that most of the Mizar objects have different numbers and order of arguments in compared representations, since the semantic representation contains visible arguments of each Mizar object, but also their hidden arguments calculated by the Mizar Analyzer. We explore these differences to access hidden arguments at the syntactic level, and use them just like visible arguments. Additionally, as in the case of logic connectives, we present every object using an application of meta-constant that corresponds to unique pattern of the object and list of its arguments.

For example, let us consider the statement of theorem SUBSET_1:13 presented in Fig. 3. It states that the set difference of sets A and B is equal to the intersection of A and the complement of B in the given universal set E.

The universal set E does not appear *explicitly* in the statement, but is necessary to determine the complement set B`. Additionally, the difference (represented as \) and the intersection (/\) are originally defined for arbitrary sets, however the Mizar redefinitions (for more detail see [22]) change types of return values in these functors for Subset of E, if both arguments are also Subset of E. Therefore, E is a hidden argument of these three functors if we want to fully reflect the meaning of this statement. The representation of the fact is presented in Fig. 4.

To provide a human-friendly access to our representation we also build an initial system that automatically generate pdf files that visualize our representation as well as HTMLization of the MML

```
reserve E for set,
        A,B for Subset of E;
theorem :: SUBSET_1:13
  A \ B = A /\ B`;
```

Fig. 3. An example Mizar theorem whose statement contains hidden arguments.

```
<app>
 <const id="XBOOLE_0R4" type="o" args="2"
   argsType="set"/>
 <app>
  <const id="SUBSET_1K7" type="set" args="3"
    argsType="set"/>
  <var id="A" type="set" args="0" argsType="set"/>
  <var id="E" type="set" args="0" argsType="set"/>
  <var id="B" type="set" args="0" argsType="set"/>
 </app>
 <app>
  <const id="SUBSET_1K9" type="set" args="3"
    argsType="set"/>
  <var id="A" type="set" args="0" argsType="set"/>
  <var id="E" type="set" args="0" argsType="set"/>
  <app>
   <const id="SUBSET_1K3" type="set" args="2"
     argsType="set"/>
   <var id="B" type="set" args="0" argsType="set"/>
   <var id="E" type="set" args="0" argsType="set"/>
  </app>
 </app>
</app>
```

Fig. 4. The formula `A \ B = A /\ B`` represented in our format. These are necessary to decode the complete information. `SUBSET_1K7` corresponds to the pattern `_ \ _`, defined in the Mizar article `SUBSET_1` to represent the difference of sets in an universe, where the universe is a hidden argument and is calculated by Mizar from types of the sets. Similarly, `SUBSET_1K9` and `SUBSET_1K3` correspond to patterns `_ /\ _` and `_`, respectively.

visualize the semantic representation. We use a presentation inspired by that of Isabelle rendering of its formalizations, in particular applied to Isabelle/Mizar that combines selected components of Isabelle/Isar and Mizar[4]. In particular, the theorem presented in Fig. 3 is expressed as follows:

mtheorem subset_1_th_13:
∀E: ⟨ set HIDDENM2 ⟩ .
∀A: ⟨ Subset SUBSET_1M2 of E ⟩ .
∀B: ⟨ Subset SUBSET_1M2 of E ⟩ .
A \ SUBSET_1K7⟨E⟩ B = XBOOLE_0R4
A /\ SUBSET_1K9⟨E⟩ B ` SUBSET_1K3⟨E⟩

where hidden quantifiers and types imported from reservations are highlighted as well as hidden arguments are visible in subscripts. Additionally, identifiers indicate absolute patterns and links indicate absolute constructors in the HTMLization of the current MML.

[4] Readers can check automatically generated pdf files (generated now for 104 initial Mizar articles) at the author's web site http://alioth.uwb.edu.pl/~pakkarol/fedcsis2018/, stylized for the Isabelle/Isar language.

```
reserve x for object, X,Y for set;
theorem
  X/\Y = X implies X c= Y
proof
  assume that A1: for x st x in X/\Y holds x in  X and
              A2: X c= X/\Y and
              A3: ex x st x in X & not x in Y;
```

Fig. 5. An example Mizar assumption where two predicates (equality and inclusion) are unfolded in one skeleton step.

IV. Simplification of Mizar proofs by cut introduction

The Mizar proof style, inspired by Jaśkowski [23], provides various natural deduction steps (called *Skeleton steps* in Mizar). The steps generally modify the current part of a given *thesis* that still remains to be proven. Thesis is the same as the current goal at the beginning of every proof, but further it becomes *implicit*, as is done in informal proofs. Indeed, mathematicians do not often indicate what has been done or what is left in the middle of proofs. It means that Mizar authors must know the current thesis and predict how it will be changed by a particular skeleton step to finish a given proof. Mizar proof is finished if the thesis is reduced to `verum` (*true*). Then the Mizar Reasoner tries to adapt the skeleton steps proposed by authors even if this requires unfolding the definitions of several predicates.

A. Unfolded Predicates

An example of an *implicit* unfolded definition is presented in Fig. 1 in line 11. The generalization step (keyword let) can be used if the current thesis is a universally quantified formula, but the current thesis after line 10 is a formula (`X c= Y\{x}`) that becomes a universally quantified formula if we unfold a definition of set inclusion. Such an approach gives a lot of freedom for authors, but is very hard to control by existing external tools. It is important to note that the Mizar semantic representation has been enriched by Urban with a list of definitions that the Mizar Reasoner actually needs to unfold in every step. Such information is sufficient to cross-verify a given thesis modification done by Reasoner (for more detail see, [24]). However this information is not sufficient to determine reasoning pattern, since it does not determine positions of unfolded predicates, or even their number. An example of an assumption accepted by Reasoner supported by two definitions, namely the definition of equality as two inclusions and the definition of inclusion is presented in Fig. 5. The equality `X/\Y = X` is introduced as two inclusions, where the first one `X/\Y c= X` has been unfolded, and further the indirect proof is started where the inclusion in the indirect assumption not `X c= Y` has been unfolded. Determination of a reasoning pattern for such an assumption is a severe problem. However, combining the syntactic and semantic information we can automatically

eliminate the more advanced skeleton steps, generating the corresponding reasoning patterns while introducing fewest changes in the reasoning.

B. Procedure Overview

Note that we can transform modified thesis by a given reasoning step back to the thesis before this step or to an equivalent formula, if we take into account the meaning of the definitions unfolded there (for more details see [24]). It means that we can reconstruct an equivalent of the thesis by analyzing the skeleton steps from the end of a given proof, if there are only *simple* (i.e., without definitional expansions) kinds of steps, such as generalizations, assumptions, conclusions (or shorter let–assume–thus) that correspond directly to the Isar fix–assume–show. Moreover, in such cases we can indicate a list of natural deduction rules that precisely correspond to the related created thesis and proof. In our representation we only use implication introduction and the following four rules (expressed in the Isabelle syntax):

lemma impMI: (A1 ⟹ A2 ⟶ C) ⟹
  A1 ∧ A2 ⟶ C
lemma conjMI: C2 ⟹ C1 ⟹ C1 ∧ C2
lemma ballI: (⋀x. x be D ⟹ P(x)) ⟹
  inhabited(D) ⟹ ∀x:D. P(x)
lemma bexI: P(x) ⟹ x be D ⟹
  inhabited(D) ⟹ ∃x:D. P(x)

where impMI connects uncurry and impI; conjMI is a modification of conjI; ballI, bexI are bounded quantifier introduction and elimination rules which apart from the condition ensure that the given Mizar types are inhabited. These correctly correspond to the Mizar foundations (see [17]). Note that conjMI corresponds to the Mizar conclusion where a given proposition is a conjunct of the current thesis and impI separates a list of conjunctions in an assumption to give them independent labels as follows:

have A ∧ B ∧ C
proof(rule conjMI,rule conjMI)
  show A ⟨proof⟩
  show B ⟨proof⟩
  show C ⟨proof⟩
qed
have A ∧ B ∧ C ⟶ D
proof(rule impMI,rule impMI,rule impI)
  assume a: A **and** b: B **and** c: C
  show D ⟨proof⟩
qed

As shown in Fig. 5, a reconstructed thesis can not be easily matched to a given thesis in a proof, if some definitions have been unfolded in a let–assume–thus step. Therefore, in our approach we introduce a cut in the reasoning at every place where such steps occur. Let us fix such a let–assume–thus step. We encapsulate a part of deduction beginning from the step using the created list of rules as a sub-deduction that proves the reconstructed thesis (the correctness condition of such cut introduction have been developed in [25]). Then we replace this step by a conclusion where the original thesis is given as the proposition and we refer to the sub-deduction and unfolded definitions and as the justification. An example of such a cut introduced to our representation is presented in Fig. 6.

```
have ∀ y : ⟨ object HIDDENM1 ⟩ .
    y in HIDDENR3 X →
     y in HIDDENR3 Y \ XBOOLE_0K4 { TARSKI1 x }
proof(rule ballI,rule impI)
 fix y being ⟨ object HIDDENM1 ⟩
 assume A4: y in HIDDENR3 X
 hence y <> HIDDENR2 x using A3;
 hence A5: ¬ y in HIDDENR3 { TARSKI1 x }
    using tarski_def_1;
 have X c= TARSKIR1 Y using A1;
 hence y in HIDDENR3 Y using A4;
 thus y in HIDDENR3 Y \ XBOOLE_0K4 { TARSKI1 x }
    using xboole_0_def_5, A5;
qed
thus X c= TARSKIR1 Y \ XBOOLE_0K4 { TARSKI1 x }
   using tarski_def_3;
```

Fig. 6. An example of a cut introduction related to the skeleton step located in line 11 in Fig. 1.

C. take *steps*

The Mizar take is a kind of skeleton step that is a challenge for other declarative proof languages, including expressing the proofs in Isabelle/Mizar, as such steps cannot be omitted. take indicates terms suitable for instantiating an existentially quantified thesis. Such terms can be constructed using any available constants in Mizar. For comparison, there is a limitation for kinds of constants in the Isabelle/Isar language i.e., constants introduced inside obtain steps have to be available before a deduction where we use them to construct such suitable term. Unfortunately, the obtain step is the only equivalent of the Mizar consider (for more detail see [21]) and most of take steps are using consider constants. A cut introduction is one and only one solution that we introduce in our representation. For example, we can introduce the following cut

```
show ∃ t : object HIDDENM1 .
 t in HIDDENR3 Y ∧
   X c= TARSKIR1 Y \ XBOOLE_0K4 { TARSKI1 t }
proof(rule bexI[of _ x],rule conjMI)
 show x in HIDDENR3 Y using A2;
 have ∀ y : ⟨ object HIDDENM1 ⟩ .
    y in HIDDENR3 X →
    y in HIDDENR3 Y \ XBOOLE_0K4 { TARSKI1 x }
 proof(rule ballI,rule impI)...
 thus X c= TARSKIR1 Y \ XBOOLE_0K4 { TARSKI1 x }
    using tarski_def_3;
qed
```

in relation to the step take x; presented in Fig. 1.

Note that to introduce such cut we have to extract a given term and also its type, but the type can be *implicit* even in the semantic representation. Generally, we can extract this type comparing the thesis before and after a given take step, but not in most cases where some definitions have been unfolded. For them we use the following solution. Let us regard t as such take step and

denote by `f` the first skeleton step after `t` that is not a take step where we can not extract a type. First, we encapsulate a part of deduction beginning from `f` as a sub-deduction that proves a thesis that corresponds to `f`. Then we formulate a conclusion with the original thesis of `t` and as a justification we refer to the sub-deduction, unfolded definitions in `t`, but also in all skeleton steps between `t` and `f`; and a list of bexl rules substituted by terms given in `t` and skeleton steps between `t` and `f`. In the case of the take step presented in Fig. 1. proposed solution should introduce the following cut:

```
show ∃ t : object HIDDENM1 .
 t in HIDDENR3 Y ∧
   X c= TARSKIR1 Y \ XBOOLE_0K4 { TARSKIK1 t }
proof-
 have x in HIDDENR3 Y ∧
    X c= TARSKIR1 Y \ XBOOLE_0K4 { TARSKIK1 x }
 proof(rule conjMI)...
 thus ?thesis using unfolded definitions bexI[of _ x];
qed
```

D. The rest of skeleton steps

The remaining Mizar skeleton steps not explained so far are: given, hereby and also per cases steps that play a similar role as skeleton steps, but do not modify a considered thesis.

The given step is an abbreviation for an assume step that as a valid proposition introduces an existential statement and a consider step that creates a fresh constant and provides access to the instantiated existential statement with the constant. We replace every such step via assume, consider/obtain.

To describe the hereby step, we must first introduce the now concept, since hereby is a simply abbreviation of thus now. In the Mizar language now opens a sub-deduction where the proved statement is not written explicitly but is reconstructed from the sub-deduction. The sub-deduction can be conducted with the support of all kinds of skeleton steps with only one restriction that each take steps have to be formulated as "take *new constant=term*" to indicate all terms that should be replaced by a variable bounded by the appropriate existential quantifier. Weaker equivalent of the now blocks are present in some proof languages, for example the Isabelle/Isar `{...}` concept supports only deduction via fix–assume where the last `have` step is chosen as the conclusion. Therefore we replace the now blocks by *normal* steps with reconstructed statements in our export.

The per cases step is a kind of step that generally reflects the idea of the informal proof by cases, where a thesis has to be justified under each of logically complementary alternatives, but not necessarily with identical skeleton steps. We encapsulate the sub-deduction in each case as a justification of the corresponding implication "*case assumption* implies *reconstructed thesis*".

V. Our representation as a next stage to cross-verify MML in Isabelle

In this section we describe possibilities of our representation in relation to the needs of the Isabelle logical framework and in particular Isar reasoning patterns. In our previous work [18], we defined a unique and faithful equivalent of the Mizar dependent type system and higher-order concepts as an Isabelle object logic. This equivalent has been tested so far only on a manually reformalized part of the MML. However, the experience gained during manual re-formalization showed directions in which we can bring the Isar language closer to the Mizar one as well as unattainable goals, e.g., the take step. Our manual attempts to generate Isar reasoning patterns also showed that with enough effort, we can indicate corresponding lists of rules even for very intricate deductions. However, such list are too sensitive to minor changes, even in *simple* deductions. Our representation of proofs is an attempt to solve these problems on the MML side.

Opportunities offered by this representation at the moment have been visualized on a re-formalization of the theorem presented in Fig. 1 created in our Isabelle environment. The re-formalization is presented in Fig. 7 and reflects all elements contained in the automatically generated visualization[5].

The example demonstrates the usefulness of the reconstructed reserve concept (Section III) that is welcome in the human-readable export. Note that the reserve command is our Isabelle/Mizar equivalent of the Mizar reserve mechanism that collects variable names with their types. Therefore, we do not need mention the types of `X`, `Y`, `x` in quantified formulas same as in the Mizar proof scripts. Additionally, we can hide the first two quantifiers in the statement of the theorem, since the mtheorem command automatically binds free reserve variables, introduces them into the sub-deduction, and adds corresponding propositions of the shape "*term* is *type*" to the *background knowledge* of the proof stored by a designated theorem list (`ty`). Then the knowledge, extended by some additional informations is added to the list of premises by the proof method mauto before auto call (more detail in our formalization).

We can also observe consequences of the introduced cuts and the simplicity of the generated reasoning pattern described in Section IV. Note that Isabelle/Isar does not accept a given sub-deduction as the justification of a particular statement, even if it accepts justifications for all the steps in a sub-deduction, since the given reasoning pattern does not precisely correspond to the statement and the sub-deduction.

VI. Conclusion

We have introduced a combination of the Mizar syntactic and semantic proof representations and presented

[5]See http://alioth.uwb.edu.pl/~pakkarol/fedcsis2018/mispdf/xboole_0.pdf

```
reserve X,Y for set
reserve x for object
mtheorem xboole_0_th_8:
 ∀ X. ∀ Y. X c< Y ⟶
   (∃ x. x in Y ∧ X c= Y\{x} )
proof -
 have ∀ X. ∀ Y. X c< Y ⟶
   (∃ t:object. t in Y ∧ X c= Y\{t} )
 proof(rule balll,rule balll,rule impl)
  fix X assume [ty]:X be set
  fix Y assume [ty]:Y be set
  assume A1: X c= Y
  then obtain x where [ty]: x be object and
   A2: x in Y and A3: ¬ x in X using xboole_0_th_6
   by mauto
  show ∃ t : object. t in Y ∧ X c= Y\{t}
  proof(rule bexI[of _ x],rule conjMI)
   show x in Y using A2 by auto
   have ∀ y : object. y in X ⟶ y in Y\{x}
   proof(rule balll,rule impl)
    fix y assume [ty]:y be object
    assume A4: y in X
    hence y <> x using A3 by auto
    hence A5: ¬ y in {x} using tarski_def_1
     by auto
    have X c= Y using A1 xboole_0_def_8
     by mauto
    hence y in Y using A4 tarski_def_3 by mauto
    thus y in  Y\{x} using xboole_0_def_5 A5
     by mauto
   qed mauto
   thus X c= Y\{x} using tarski_def_3 by mauto
  qed mauto
 qed mauto
 thus ?thesis by mauto
qed
```

Fig. 7. An example Isabelle/Mizar reasoning that exactly corresponds to the combined representation of the proof script presented in Fig. 1. Note that the highlighted part of reasoning can be removed from the script without influence for its correctness just like in the Mizar proof scripts (for more detail see our formalization)

a number of possibilities that such combined data offers. We rebuild the Mizar natural deduction style to the fix–assume–thus proof outlines present in declarative proof modes, including that of Isabelle/Isar. The transformation preserves all Mizar components using cut introduction. This eliminates all the Mizar natural deduction constructions for which adequate equivalent constructs do not exist in other systems. In particular, it reduces the distance between the Mizar and Isabelle/Isar proof styles, which we showed in an experiment in which a transformed MML proof could be directly cross-verified in Isabelle/Mizar. The original and transformed proofs for the first 50 Mizar articles are available at:

http://alioth.uwb.edu.pl/~pakkarol/fedcsis2018/

Future work could target a further reduction of the distance between Mizar and Isabelle. The MML export combining the syntactic and semantic representations, as well as the Isabelle/Mizar object logic and its packages are under active development. Many Mizar concepts are still not completely expressed in Isabelle (e.g., the Mizar reconsider construction) or they have not been sufficiently verified (e.g., the Mizar structures [22] which are used directly or indirectly in the latter 74% of the MML). We believe that under a suitable proof translation ATPs are strong enough to accept all justifications accepted by the Mizar checker. However, to make the automation more efficient and practical, it might be necessary to extract additional knowledge from the semantic representation used as an initial information by the checker.

References

[1] G. Bancerek, C. Byliński, A. Grabowski, A. Korniłowicz, R. Matuszewski, A. Naumowicz, K. Pąk, and J. Urban, "Mizar: State-of-the-art and Beyond," in *Intelligent Computer Mathematics - International Conference, CICM 2015*, ser. LNCS, M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, and V. Sorge, Eds., vol. 9150. Springer, 2015. doi: 10.1007/978-3-319-20615-8_17 pp. 261–279.

[2] G. Bancerek, C. Byliński, A. Grabowski, A. Korniłowicz, R. Matuszewski, A. Naumowicz, and K. Pąk, "The role of the Mizar Mathematical Library for interactive proof development in Mizar," *J. Autom. Reasoning*, 2017. doi: 10.1007/s10817-017-9440-6. [Online]. Available: https://doi.org/10.1007/s10817-017-9440-6

[3] G. Bancerek and P. Rudnicki, "Information retrieval in MML," in *Mathematical Knowledge Management, MKM 2003*, ser. LNCS, A. Asperti, B. Buchberger, and J. H. Davenport, Eds., vol. 2594. Springer, 2003. doi: 10.1007/3-540-36469-2_10. ISBN 3-540-00568-4 pp. 119–132. [Online]. Available: https://doi.org/10.1007/3-540-36469-2_10

[4] J. Urban, "XML-izing Mizar: Making semantic processing and presentation of MML easy," in *Mathematical Knowledge Management (MKM 2005)*, ser. LNCS, M. Kohlhase, Ed., vol. 3863. Springer, 2005. ISBN 3-540-31430-X pp. 346–360.

[5] M. Iancu, M. Kohlhase, F. Rabe, and J. Urban, "The Mizar Mathematical Library in OMDoc: Translation and applications," *J. Autom. Reasoning*, vol. 50, no. 2, pp. 191–202, 2013. doi: 10.1007/s10817-012-9271-4

[6] J. Urban, "MPTP 0.2: Design, implementation, and initial experiments," *J. Autom. Reasoning*, vol. 37, no. 1–2, pp. 21–43, 2006. doi: 10.1007/s10817-006-9032-3

[7] C. Kaliszyk and J. Urban, "MizAR 40 for Mizar 40," *J. Autom. Reasoning*, vol. 55, no. 3, pp. 245–256, 2015. doi: 10.1007/s10817-015-9330-8

[8] A. Naumowicz and R. Piliszek, "Accessing the Mizar library with a weakly strict Mizar parser," in *Intelligent Computer Mathematics, CICM 2016*, ser. LNCS, M. Kohlhase, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016. doi: 10.1007/978-3-319-42547-4_6 pp. 77–82. [Online]. Available: http://doi.org/10.1007/978-3-319-42547-4_6

[9] C. Byliński and J. Alama, "New Developments in Parsing Mizar," in *Intelligent Computer Mathematics - 11th International Conference, AISC 2012, 19th Symposium, Calculemus 2012, 5th International Workshop, DML 2012, 11th International Conference, MKM 2012, Systems and Projects, Held as Part of CICM 2012*, ser. 7362, J. Jeuring, J. A. Campbell, J. Carette, G. D. Reis, P. Sojka, M. Wenzel, and V. Sorge, Eds., vol. 5170. Springer, 2012. doi: 10.1007/978-3-642-31374-5 pp. 427–431.

[10] J. Urban, "Translating Mizar for first order theorem provers," in *Mathematical Knowledge Management, Second International Conference, MKM 2003, Bertinoro, Italy, February 16-18, 2003, Proceedings*, ser. LNCS, A. Asperti, B. Buchberger, and J. H. Davenport, Eds., vol. 2594. Springer, 2003. doi: 10.1007/3-540-36469-2_16 pp. 203–215.

[11] J. Urban and G. Sutcliffe, "ATP-based cross-verification of Mizar proofs: Method, systems, and first experiments," *Math.*

in Computer Science, vol. 2, no. 2, pp. 231–251, 2008. doi: 10.1007/s11786-008-0053-7

[12] O. Kunčar, "Reconstruction of the Mizar type system in the HOL Light system," in *WDS Proceedings of Contributed Papers: Part I – Mathematics and Computer Sciences*, J. Pavlu and J. Safrankova, Eds. Matfyzpress, 2010, pp. 7–12.

[13] C. E. Brown and J. Urban, "Extracting higher-order goals from the Mizar Mathematical Library," in *Intelligent Computer Mathematics (CICM 2016)*, ser. LNCS, M. Kohlhase, M. Johansson, B. R. Miller, L. de Moura, and F. W. Tompa, Eds., vol. 9791. Springer, 2016. doi: 10.1007/978-3-319-42547-4_8 pp. 99–114.

[14] J. Urban, P. Rudnicki, and G. Sutcliffe, "ATP and presentation service for Mizar formalizations," *J. Autom. Reasoning*, vol. 50, no. 2, pp. 229–241, 2013. doi: 10.1007/s10817-012-9269-y. [Online]. Available: https://doi.org/10.1007/s10817-012-9269-y

[15] J. C. Blanchette, D. Greenaway, C. Kaliszyk, D. Kühlwein, and J. Urban, "A learning-based fact selector for Isabelle/HOL," *J. Autom. Reasoning*, vol. 57, no. 3, pp. 219–244, 2016. doi: 10.1007/s10817-016-9362-8. [Online]. Available: http://dx.doi.org/10.1007/s10817-016-9362-8

[16] M. Wenzel, L. C. Paulson, and T. Nipkow, "The Isabelle framework," in *Theorem Proving in Higher Order Logics, 21st International Conference, TPHOLs 2008*, ser. LNCS, O. A. Mohamed, C. A. Muñoz, and S. Tahar, Eds., vol. 5170. Springer, 2008. doi: 10.1007/978-3-540-71067-7_7 pp. 33–38.

[17] C. Kaliszyk, K. Pąk, and J. Urban, "Towards a Mizar environment for Isabelle: Foundations and language," in *Proc. 5th Conference on Certified Programs and Proofs (CPP 2016)*, J. Avigad and A. Chlipala, Eds. ACM, 2016. doi: 10.1145/2854065.2854070 pp. 58–65.

[18] C. Kaliszyk and K. Pąk, "Semantics of Mizar as an Isabelle object logic," *J. Autom. Reasoning* 2018. doi: doi.org/10.1007/s10817-018-9479-z. [Online]. Available: https://doi.org/10.1007/s10817-018-9479-z

[19] ——, "Progress in the independent certification of Mizar Mathematical Library in Isabelle," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2017. doi: 10.15439/2017F289 pp. 227–236.

[20] ——, "Isabelle formalization of set theoretic structures and set comprehensions," in *Mathematical Aspects of Computer and Information Sciences, MACIS 2017*, ser. LNCS, J. Blamer, T. Kutsia, and D. Simos, Eds., vol. 10693. Springer, 2017.

[21] M. Wenzel and F. Wiedijk, "A comparison of Mizar and Isar," *J. Autom. Reasoning*, vol. 29, no. 3-4, pp. 389–411, 2002. doi: 10.1023/A:1021935419355

[22] A. Grabowski, A. Korniłowicz, and A. Naumowicz, "Mizar in a nutshell," *J. Formalized Reasoning*, vol. 3, no. 2, pp. 153–245, 2010. doi: 10.6092/issn.1972-5787/1980

[23] S. Jaśkowski, "On the rules of suppositions," *Studia Logica*, vol. 1, 1934.

[24] J. Urban and G. Sutcliffe, "ATP cross-verification of the Mizar MPTP challenge problems," in *Logic for Programming, Artificial Intelligence, and Reasoning, 14th International Conference, LPAR 2007, Yerevan, Armenia, October 15–19, 2007, Proceedings*, 2007. doi: 10.1007/978-3-540-75560-9_39 pp. 546–560.

[25] K. Pąk, "Methods of lemma extraction in natural deduction proofs," *J. Autom. Reasoning*, vol. 50, no. 2, pp. 217–228, February 2013. doi: 10.1007/s10817-012-9267-0. [Online]. Available: http://dx.doi.org/10.1007/s10817-012-9267-0

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 155–158
DOI: 10.15439/2018F79
ISSN 2300-5963 ACSIS, Vol. 15

Modelling Legal Interpretation in Structured Argumentation Framework

Tomasz Zurek
Institute of Computer Science,
Maria Curie-Sklodowska University in Lublin
Ul. Akademicka 9, 20-033 Lublin, Poland
Email: zurek@kft.umcs.lublin.pl
Michał Araszkiewicz
Departament of Legal Theory, Jagiellonian University,
Bracka Str. 12, 31-005 Cracow, Poland
Email: michal.araszkiewicz@uj.edu.pl

***Abstract*—The paper discusses the problem of formal modeling of the interpretation of statutory legal norms. The authors propose a comprehensive framework that allows the representation of the interpretation process. The authors' proposal is illustrated by a real-life example.**

Introduction

LEGAL interpretation is one of the most important problems in legal theory and practice. The aim of our paper is to develop formalized descriptive model of statutory interpretation. We are interested in formal modelling of actual interpretive argumentation rather than in developing its idealized picture. The second aim of our work is to integrate our model of legal interpretation with one of the most popular formal model of argumentation (ASPIC+ [1]). The development of a fully-fledged descriptive model of legal interpretation is a complex research project, and perhaps rather a regulative idea rather than an operational goal. The realisation of such a goal requires dealing with certain problems that remain unsolved in the current state of the art. In this contribution we model an actual case involving statutory interpretation to represent different arguments developed by different agents for the sake of the realisation of goals important for these agents.

I. Interpretive agents

The authors of [2] notice that a significant role in the argumentation process is played not only by the interpretive argument itself, but also by the agent putting forward the argument, since in legal discourse every agent plays a particular role with his/her own preferences and goals. In [2] a semi-formal framework is presented which permits to model the agent's role in interpretive arguments. This framework will constitute the grounds for the model of agent in our argumentative system.

II. Argumentation framework

ASPIC$^+$ is a well-developed framework for structured argumentation representation ([1], [3] and many others). As such, it does not specify any logical language to represent arguments, but argumentation represented in this framework may be instantiated in different languages.

III. The Model for Legal Interpretation

Let $\mathscr{L}$ be any well-defined and closed under negation language. First, let us define certain postulated sets. Let S be a legal system in question, let C be a concrete or a hypothetical case in question, and let $\mathscr{L}$ be the language under consideration.

Definition 1 (Source Set): Let K be a knowledge base in argumentation system AS. Then $SRC(S,C)$ (Source Set under a legal system S in the context of a case C) consists of:

1) $ST(S)$ is the set of all explicit statutory norms under a system S. $ST(S) \subset K_p$. Each norm in $ST(S)$ is represented by a predicate $norm(\alpha,\beta,\gamma)$, where $\alpha \in \mathscr{L}$ is the name of the norm, β is a wff in $\mathscr{L}$ which represents the conditional part of the norm, and γ is a wff in $\mathscr{L}$ which represents the conclusion of the norm;
2) $Cases(S)$ is the set of wff in $\mathscr{L}$ which represents all accessible judicial opinions ruled under a system S. $Cases(S) \subset K_p$;
3) $Doctrine(S)$ is the set of wff in $\mathscr{L}$ which represents all scholarly opinions concerning legal issues arising under a legal system S. $Doctrine(S) \subset K_p$;
4) $Materials(S)$ is the set of wff in $\mathscr{L}$ which represents the remaining official materials that may be relevant for the sake of interpretation of statutory law under a system S, such as legislative opinions, soft law and the like. $Materials(S) \subset K_p$;
5) CSK is the set of wff in $\mathscr{L}$ which represents all available common sense knowledge propositions. $CSK \subset K_p$;
6) SK is the set encompassing propositions which are referred to as Scientific Knowledge. $SK \subset K_p$;
7) $Facts(C)$ is the complete set of propositions describing the facts of a case C in question. $Facts(C) \subset K_n$;
8) $IT(\mathscr{L})$ is the set of all Interpretive Terms in a language $\mathscr{L}$, that is, terms that may be used for the sake of the interpretation of any term of $ST(S)$. $IT(\mathscr{L}) \subset K_p$;

9) $R_i \subset R_d$ is the set of all argumentation schemes (represented as defeasible inference rules) used to generate interpretive arguments from the knowledge contained in sets 1)-8) above, hereafter referred to as Source Sets.

Definition 2 (Extensional relations): Extensional relation $INC \subset \mathscr{L}$ is a set of binary relations encompassing inclusion relation $\sqsubseteq$, strict inclusion relation $\sqsubset$, and equivalence relation ($\equiv$) defined on the set $\mathscr{L}$.
If $X \sqsubseteq Y$ and X and Y are two expressions in $\mathscr{L}$, then we claim that X is within the scope (semantic extension) of Y.

Definition 3 (Interpretation): $\bullet \in \mathscr{L}$ is a binary relation word denoting "legally counts as" or "is interpreted as". We introduce this relation in order to grasp the phenomenon in which, in certain cases, an expression may legally count as an instance of another expression even though it is situated outside its semantic extension. It it worth to notice that the relation "is interpreted as" should be understood as a presumptive (defeasible) one.
Relation $\bullet$ is reflexive (because $\phi \bullet \phi$), not symmetric (because if $\phi \bullet \psi$ then it is not necessary that $\psi \bullet \phi$), and transitive (because if $\phi \bullet \psi$ and $\psi \bullet \delta$ then $\phi \bullet \delta$)
Each of the sources from a Source Set can be interpreted as a kind of a knowledge base which allows to infer and examine whether a given proposition is in the scope of meaning of a certain expression.

Definition 4 (Interpretive Sentences): All complex expressions of a language $\mathscr{L}$ constructed by means of any of the elements from the set INC or by means of the relation word $\bullet$ will be referred to as Interpretive Sentences.
The legal theory points out that Interpretive Sentences are justified by means of interpretive arguments or canons. In our work interpretive canons will be represented by interpretive inference rules:

Definition 5 (Interpretive Inference rules): All inference rules whose conclusions are interpretive sentences or undercutters of other interpretive inference rules will be referred to as interpretive inference rules. The set of all Interpretive Inference Rules will be denoted as R_i ($R_i \subset R_d$)

Definition 6 (Interpetive Arguments): If A is an argument constructed by means of a knowledge base K in AS and the last inference rule in A is built on the basis of an interpretive inference rule ($TopRule(A) \in R_i$), then argument A is an interpretive argument.
Although we assumed that the sources of justifications do not have to be consistent, we assume that an interpretive argument should be internally consistent. By an internally consistent argument we understand an argument in which:
$\nexists_{\varphi,\psi}(\varphi \in Prem(A) \wedge (\psi \in Prem(A) \vee Conc(A) = \psi) \wedge \psi \in \overline{\varphi})$.

A. Authorship of interpretive arguments

The authors of [2] point out that the notion of interpreting agent plays a crucial role in a descriptive model of legal interpretation. The notion of interpreting agents enables us to attibute certain statements and arguments to a given agent, which allows for a more fine-grained representation of argumentation in real-life cases.

Definition 7 (Set of Agents): Let $IA \subseteq K_n$ be a collection of the agents' names. Each $ia \in IA$ will be the name of the agent present in a legal case c.

Definition 8 (Authorship of an argument): The relation of authorship is a subset of a Cartesian product:
$\mathscr{R} \subseteq IA \times \mathscr{A}$, i.e. a set of pairs: (ia, A), where $ia \in IA$ and $A \in \mathscr{A}$.
This relation shows who the author of a given argument is. The argument can have many authors; one agent may be the author of many arguments.
To consider the issue of argument authorship in structured argumentation framework, the *SAF* from [1] definition must be adjusted:

Definition 9: A structured authored argumentation framework (*SAAF*) is a tuple $\langle \mathscr{A}, \mathscr{C}, \preceq, \mathscr{R} \rangle$ where:

- $\mathscr{A}$ is the smallest set of all finite arguments constructed from a knowledge base in AS;
- $\preceq$ is an ordering on $\mathscr{A}$;
- $(X,Y) \in \mathscr{C}$ iff X attacks (is in conflict with) Y.
- $\mathscr{R}$ is an authorship relation on sets $IA \subset K_n$ and $\mathscr{A}$.

B. Model of Interpreting Agent

Definition 10 (Agent): Basing on the model from [2], it is assumed that an agent ia in a structured authored argumentation framework *SAAF* will be a tuple:
$(KB(ia), preferences(ia), authority(ia))$
Where:

- $KB(ia) \subseteq SRC(S,C)$ The knowledge base of an agent IA is a subset of the Source Set
- $preferences(ia) \subseteq K_n$ and $preferences(ia) = NormPref(ia) \cup SubPref(ia)$
  - $NormPref(ia) = (<_{NP(ia)}, \mathscr{L})$ $NormPref(ia)_t$ is a partial order on set wff in $\mathscr{L}$
  - $SubPref(ia) = (<_{SP(ia)}, \mathscr{L})$ $SubPref(ia)$ is a partial order on set wff in $\mathscr{L}$
- $authority(ia) \subseteq K_n$ The relation of Authority is a subset of a Cartesian product $authority(ia) \subseteq S(ia) \times IA$, where $S(ia)$ is the set of all sentences stated by an Interpretive Agent in the case c (formally: $S(ia) = \{Conc(A_n) : (Conc(A_n), ia) \in \mathscr{R}\}$, i.e. a set of pairs of statements given by the agent in a case c and agents formally bound by these sentences).

On the basis of the relations $preferences(ia)$, $authority(ia)$, and the relevant inference rules (to appear in future work), it will be possible to establish a relation of order between conflicting arguments in a structured argumentation framework (relation $\preceq$).

IV. Example

This section presents a modelling of interpretation in an actual case also discussed in [2]. However, here, in addition to presentation of the knowledge bases of the relevant agents, we also reconstruct the arguments developed and used by them. The legal issue at stake was as follows. Generally, according to the Personal Income Tax Act (PITA), the taxpayer's total

revenue is taken into account in the calculation of taxable income, unless this revenue is exempted. Pursuant to the provision of 21.1.47c of the PITA, revenues raised by a natural person from a governmental or an executive agency, where the agency is financed from the state budget, are exempted from tax. The protagonist of the case obtained a housing benefit from the Military Housing Agency and claimed that this revenue was exempted from tax. However, the tax authorities disagreed, pointing out that the legislative materials suggested that the exemption is question was intended to apply to entrepreneurs, while the protagonist of the case was not one. The assessment of the Court is that the opinions presented by the tax office are not sufficiently justified both in relation with the provisions of tax law and the factual circumstances of the case. Such a conclusion results primarily from the outcomes of the linguistic interpretation of the Act. It follows from this regulation that in order to apply the exemption in question, two conditions must be fulfilled: first, the taxpayer is to receive a specific amount from a government or executive agency; second, this agency is to receive funds for this purpose from the state budget. In the view of the Court, both conditions which determine the exemption are fulfilled in this case. It should be highlighted that in the interpretation of tax law provisions, the linguistic interpretation of the text of the Act has the primary and dominant weight. Under no circumstances is it permitted in the tax law system to apply teleological, systemic, or historical interpretation of a provision of tax law to the factual circumstances should its result (even if obtained correctly) be inconsistent with the result of the linguistic interpretation.

Basics:

First, the alphabet of the language is defined:

- Propositional atoms: $\{housing\_benefit, natural\_person, person, enterprise, revenue\_from\_agency, revenue, agency\_financed\_from\_the\_state\_budget, tax\_law, tax, d_{ling}, d_{hist}, d_{mod}, r1, r2\}$
- symbols: $\{\neg, \wedge, \vee, \supset, \sqsubseteq, \sqsubset, \equiv, \bullet, norm(,,)\}$
- Interpreting Agents: $IA = \{ia_{person}, ia_{taxOffice}, ia_{judge}\}$

Knowledge Base

The authors of [4] (section 4.0) distinguish two ways of utilization of the $ASPIC^+$ framework: domain-specific vs. general inference rules. In order to model our example, we will use the second one.

Facts of the case:
$Facts(C) = \{housing\_benefit, natural\_person, revenue\_from\_agency, agency\_financed\_from\_the\_state\_budget, tax\_law\}$
Commonsense knowledge:
$CSK(S) = \{revenue\_from\_agency \sqsubseteq revenue\}$
Applicable law:
$ST(S) = \{$
$norm(r1, housing\_benefit \wedge revenue\_from\_agency \wedge agency\_financed\_from\_the\_state\_budget, \neg tax,$
$norm(r2, revenue, tax)\}$
There are two legal rules: it follows from the first one that revenues raised from a governmental or an executive agency, where the agency is financed from the state budget, are exempted from tax, whereas the other rule states that all kinds of revenue are taxable.
Historical materials:
$Materials(S) : \{norm(r1, \alpha, \beta) \wedge natural\_person \supset \neg(natural\_person \wedge \alpha \bullet \alpha)\}$
According to historical materials, the legal rule $r1$ is not intended for natural persons, but for companies: a natural person does not fulfill the conditions of rule $r1$.
Doctrine:
$Doctrine(S) = \{(n(A) = d_{hist} \wedge tax\_law) \supset \neg A)\}$
The use of historical interpretation is forbidden in tax law.
Inference Rules:
Interpretive inference rules $R_i =$

- linguistic interpretation $d_{ling} : \alpha, \alpha \between \beta \Rightarrow \alpha \bullet \beta$, where $\between$ is one of the extensional relations: $\sqsubseteq, \sqsubset, \equiv$.
- historical interpretation $d_{hist} : \alpha, (\alpha \bullet \beta) \in Materials(S) \Rightarrow \alpha \bullet \beta$
- interpretation $d_{int} : \alpha, \alpha \bullet \beta \Rightarrow \beta$

Defeasible inference rules $R_d =$

- defeasible modus ponens $d_{mod} : \alpha, (\alpha \supset \beta) \Rightarrow \beta$
- legal rule application: $d_{legal} : \alpha, norm(r, \alpha, \beta) \Rightarrow \beta$

Where α, β are formulae in $\mathscr{L}$, $r \in \mathscr{L}$ is a legal rule name.

Interpreting Agents

The models of interpreting agents are adapted from [2]:
Since none of the agents from our case built arguments on the basis of the sets *Cases*, *CSK*, and *SK*, we assume that they are empty for all agents.
Agent: ia_{judge} $KB(ia_{judge})$:

- $ST(ia_{judge}) = \{$ $norm(r1, housing\_benefit \wedge revenue\_from\_agency \wedge agency\_financed\_from\_the\_state\_budget, \neg tax),$ $norm(r2, revenue, tax)\}$
- $Doctrine(ia_{judge}) = \{(InfRule(A) = d_{hist} \wedge tax\_law) \supset \neg A)\}$
- $Materials(ia_{judge}) = \varnothing$
- $Facts(ia_{judge}) = Facts(C)$
- $R_d(ia_{judge}) = R_d$

Preferences: In the analyzed case, the agent does not use his/her preferences
Authority: Interpretive statements made by the judge are binding on the tax office and the person:
If $\alpha \in S(ia_{judge})$ then $(\alpha, ia_{person}) \in authority(ia_{judge})$ and $(\alpha, ia_{taxOffice}) \in authority(ia_{judge})$
Agent: $ia_{taxOffice}$
$KB(ia_{taxOffice})$:

- $ST(ia_{taxOffice}) = \{$ $norm(r1, housing\_benefit \wedge revenue\_from\_agency \wedge agency\_financed\_from\_the\_state\_budget, \neg tax),$ $norm(r2, revenue, tax)\}$
- $Doctrine(ia_{taxOffice}) = \varnothing$

- $Materials(ia_{taxOffice}) : \{norm(r1,\alpha,\beta) \wedge natural\_person \supset \neg(natural\_person \wedge \alpha \bullet \alpha)\}$
- $Facts(ia_{taxOffice}) = Facts(C)$
- $R_d(ia_{taxOffice}) = R_d$

Preferences: $NormPref(ia_{taxOffice} = SubPref(ia_{taxOffice})$.
The agent prefers rules which increas the collected tax:
$NormPref(ia_{taxOffice}) = \{r1 <_{NP(ia_{taxOffice})} r2\}$
In the analyzed case, the agent does not use his/her preferences.

Authority: Interpreting statements made by the tax office are binding on the person:
If $\alpha \in S(ia_{taxOffice})$ then $(\alpha, ia_{person}) \in authority(ia_{taxOffice})$

Agent: ia_{person}
$KB(ia_{person})$:

- $ST(ia_{person}) = \{$ $norm(r1, housing\_benefit \wedge revenue\_from\_agency \wedge agency\_financed\_from\_the\_state\_budget, \neg tax)$, $norm(r2, revenue, tax)\}$
- $Doctrine(ia_{person}) = \varnothing$
- $Materials(ia_{person}) = \varnothing$
- $Facts(ia_{person}) = Facts(C)$
- $R_d(ia_{person}) = R_d$

Preferences: In the analyzed case, the agent does not use his/her preferences

Authority: Interpreting statements made by a person are not binding on anyone:
$authority(ia_{person}) = \varnothing$

Arguments:

First of all, the arguments of the agent *person* are presented:

A_1 : $natural\_person$
A_2 : $housing\_benefit$
A_3 : $revenue\_from\_agency$
A_4 : $agency\_financed\_from\_the\_state\_budget$
A_5 : $norm(r1,\alpha,\beta)$ where: $\alpha = housing\_benefit \wedge revenue\_from\_agency \wedge agency\_financed\_from\_the\_state\_budget$, $\beta = \neg tax$,
A_6 : $A_1, A_2, A_3, A_4 \sqsubseteq \alpha$
A_7 : $A_6 \Rightarrow (A_1 \wedge A_2 \wedge A_3 \wedge A_4) \bullet \alpha$ (Inference rule: d_{ling})
A_8 : $A_7 \Rightarrow \alpha$ (Inference rule: d_{int})
A_9 : $A_8, A_5 \Rightarrow \neg tax$ (Inference rule: d_{legal})

It follows from the above arguments that since the conditions of the legal rule $r1$ are fulfilled, the revenue should not be taxable.

Next, the arguments of agent $taxOffice$ are presented:

B_1 : $Materials(ia_{taxOffice}) : norm(r1,\alpha,\beta) \wedge natural\_person \supset \neg(natural\_person \wedge \alpha \bullet \alpha)$
B_2 : $B_1, A_1, A_2, A_3, A_4, A_5 \Rightarrow \neg((A_1 \wedge A_2 \wedge A_3 \wedge A_4) \bullet \alpha)$ (Inference rule: d_{hist})

The arguments of tax authorities are based on historical materials from which it can be concluded that the legal rule $r1$ does not apply to natural persons, and therefore the natural person (even if theoretically the conditions of $r1$ are fulfilled) cannot be interpreted as fulfilling the conditions of $r1$.

Arguments A_7 and B_2 are in conflict (B_2 rebuts A_7), and hence the case is decided by the judge:

The arguments of the agent *judge*:

C_1 : $Doctrine(ia_{judge}) = \{(InfRule(A) = d_{hist} \wedge tax\_law) \supset \neg A)\}$
C_2 : $InfRule(B_2) = d_{hist}$
C_3 : tax_law
C_4 : $C_1, C_2, C_3 \Rightarrow \neg B_2$ (Inference rule: d_{mod}).

Arguments A_7 i B_2 are, on the basis on definition 11, contradictory.

According to the doctrine (arg C_1), it is forbidden to use historical interpretation in tax law, and hence argument C_4 attacks (undercuts) argument B_2.

Since argument B_2 is undercuted by argument C_4 and:

- $(ia_{judge}, C_4) \in \mathscr{R}$
- $(Conc(C_4), ia_{person}) \in authority(ia_{judge})$,
- $(Conc(C_4), ia_{taxOffice}) \in authority(ia_{judge})$

then C_4 defeats B_2.

Since C_4 defeats B_2, then C_4 defends A_7.

V. Conclusions

The main aim of our work was to develop a formal descriptive model of statutory interpretation which can be integrated with one of the most popular argumentation frameworks (ASPIC+ [1]). Our proposal was illustrated by a model of a real life example of a legal case. Compared to the models presented in [5], [6], [7], our model is more comprehensive and abstract. We focused on the problem of integrating interpretation with the entire argumentation process, on the roles played by agents, disregarding the discussion of the structure of interpretive arguments.

References

[1] S. Modgil and H. Prakken, "The aspic+ framework for structured argumentation: A tutorial," *Argument and Computation*, vol. 5, no. 1, pp. 31–62, 2014.

[2] M. Araszkiewicz and T. Zurek, "Interpreting agents," in *Legal Knowledge and Information Systems - JURIX 2016: The Twenty-Nineth Annual Conference, Sophia-Antipolis, France*, 2016. doi: 10.3233/978-1-61499-726-9-13 pp. 13–22.

[3] B. van Gijzel and H. Prakken, "Relating Carneades with abstract argumentation via the ASPIC + framework for structured argumentation," *Argument & Computation*, vol. 3, no. 1, pp. 21–47, Mar. 2012. doi: 10.1080/19462166.2012.661766

[4] H. Prakken, "An abstract framework for argumentation with structured arguments," *Argument and Computation*, vol. 1, no. 2, pp. 93–124, 2011.

[5] F. Macagno, D. Walton, and G. Sartor, "Argumentation schemes for statutory interpretation," in *Argumentation. International Conference on Alternative Methods o Argumentation in Law.*, M. Araszkiewicz, M. Myska, T. Smejkalova, J. Savelka, and skop M., Eds., Brno, 2012, pp. 61–76.

[6] G. Sartor, D. Walton, F. Macagno, and A. Rotolo, "Argumentation schemes for statutory interpretation: A logical analysis," in *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, 2014. doi: 10.3233/978-1-61499-468-8-11 pp. 11–20.

[7] D. Walton, G. Sartor, and F. Macagno, "An argumentation framework for contested cases of statutory interpretation," *Artificial Intelligence and Law*, vol. 24, no. 1, pp. 51–91, 2016. doi: 10.1007/s10506-016-9179-0

8th International Workshop on Advances in Semantic Information Retrieval

THE International Workshop on Advances in Semantic Information Retrieval is organized as an event within FedCSIS. In 2018, we are running our eighth workshop. It is gaining popularity among researchers from Europe and Asia.

We are doing our best to follow the changes in the area of semantic information retrieval and making the necessary adjustments in the set of topics of interests. We shift the focus of this workshop on the most challenging problems. The ASIR'18 workshop will continue to maintain high standards of quality and organization, set in the previous years.

Characterizing the current tendency in development of semantic technologies, we point out that recent advances form a solid basis for a variety of methods and instruments used in multimedia information retrieval, knowledge representation, discovery and analysis. They influence the way and form of representing documents in the memory of computers, approaches to analyze documents, and techniques to mine and retrieve knowledge. In its turn, gathered knowledge is used to build problem domain models and support decision making. The abundance of video, voice and speech data also raises new challenging problems in developing multimedia information retrieval systems.

We believe that the ASIR'18 workshop will facilitate discussions of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics, data-driven decision making, data analytics and semantic information retrieval. We welcome all interested researchers to join this event.

Topics

The topics and areas include but not limited to:

- Data analytics.
- Data-driven decision making.
- Domain-specific semantic applications.
- Evaluation methodologies for semantic search and retrieval.
- Knowledge representation and management.
- Models for document representation.
- Natural language semantic processing.
- Ontology for semantic information retrieval.
- Ontology alignment, mapping and merging.
- Query interfaces.
- Searching and ranking.
- Semantic multimedia retrieval.
- Visualization of retrieved results.

Event Chairs

- **Klyuev, Vitaly,** University of Aizu, Japan
- **Mozgovoy, Maxim,** University of Aizu, Japan

Program Committee

- **Dobrynin, Vladimir,** Saint Petersburg State University, Russia
- **Goczyła, Krzysztof,** Gdansk University of Technology, Poland
- **Gutnova, Alina,** North Ossetian State University, Russia
- **Haralambous, Yannis,** Institut Telecom - Telecom Bretagne, France
- **Homenda, Wladyslaw,** Warsaw University of Technology, Poland
- **Janusz, Andrzej,** University of Warsaw, Poland
- **Jin, Qun,** Waseda University, Japan
- **Kotets, Alexey,** North Ossetian State University, Russia
- **Lai, Cristian,** CRS4, Italy
- **Makarenko, Maria,** North Ossetian State University, Russia
- **Minasyan, David,** North Ossetian State University, Russia
- **Nalepa, Grzegorz J.,** AGH University of Science and Technology, Poland
- **Pyshkin, Evgeny,** University of Aizu, Japan
- **Shtykh, Roman,** CyberAgent Inc., Japan
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates
- **Suárez-Figueroa, Mari Carmen,** Ontology Engineering Group, Scool of Computer Science at Universidad Politécnica de Madrid, Spain
- **Tadeusiewicz, Ryszard,** AGH University of Science and Technology, Poland
- **Vacura, Miroslav,** University of Economics, Czech Republic
- **Zadrożny, Sławomir,** Systems Research Institute of Polish Academy of Sciences, Poland
- **Ławrynowicz, Agnieszka,** Poznan University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 161–164
DOI: 10.15439/2018F260
ISSN 2300-5963 ACSIS, Vol. 15

A New Subject-based Document Retrieval from Digital Libraries Using Vector Space Model

Sayed Mahmood Bakhshayesh*, Azadeh Mohebi†, Abbas Ahmadi‡, and Amir Badamchi§
*Amirkabir University of Technology, Tehran, Iran
Email: s.ma.bakhshayesh@aut.ac.ir
†Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
Email: mohebi@irandoc.ac.ir
‡Amirkabir University of Technology, Tehran, Iran
Email: abbas.ahmadi@aut.ac.ir
§Amirkabir University of Technology, Tehran, Iran
Email: badamchi@aut.ac.ir

***Abstract*—Document retrieval from digital libraries based on user's query is highly affected by the terms appeared in the query. In many cases, there are some documents in the digital libraries that do not share exactly the same terms with the query, but they are related to the user's need. We address this challenge in this paper by introducing a new subject-based retrieval approach in which, apart from ranking documents based on the terms in the query, a new subject-based scoring scheme is defined between the query and a document. We define this score by introducing a new vector space model in which a vectorized subject-based representation is defined for each document and its keywords, and the terms in the query, as well. We have tested the new subject-based scoring scheme on a database of scientific papers obtained from Web of Science. Our Experimental results show that in 83% of times users prefer the proposed scoring scheme with respect to the classic scoring ones.**

I. Introduction

NOWADAYS a considerable amount of information is spread over billions of various documents saved in digital libraries. Although, various retrieval tools and algorithms have been developed to address accessing such information easily, in many cases these algorithms and tools are limited by the user's query. Many of the retrieval methods try to go beyond the exact terms in user's query. In other words, instead of only relying on Bag-of-Words (BoW) representation of the query, new approaches have been developed such as interactive query refinement, relevance feedback from user, word sense disambiguation, and clustering search results [1], [2], [3], [4], [5] to guid the user in his/her *journey of information retrieval*. More specifically, some of these methods are based on improving user involvement (implicitly or explicitly) in the retrieval process by receiving relevance feedback or providing interactive search tools. Some other methods rely on expanding user's query using query expansion techniques [6].

In this paper, we address these challenges by introducing a new *subject-based* document retrieval approach. Instead of applying query expansion techniques or using semantic relations between words and terms based on ontologies, we introduce a new subject-based representation for each document in the digital library, using vector space model. By the use of the proposed approach, we can measure how much a document and a given query are similar and share same subjects, even if they do not share same terms. The proposed approach is applicable in specialized, scientific digital libraries in which in addition to a set of keywords/tags usually assigned to each document, a set of predefined disciplines/subjects are also available and each document usually falls into a specific discipline, subject or category. In such specialized libraries, documents are usually indexed based on subjects and keywords assigned to them, to improve the indexing, retrieval and archiving tasks.

In the proposed method, first, a new subject-based vectorized representation for each keyword is introduced by relying on the knowledge obtained from all documents that have been already indexed in the digital library. Then, a probabilistic, vectorized subject-based representation for each document is estimated. Each element of this vector shows how much each document belongs to a specific subject. This consideration is based on the assumption that each document might belongs to more than one subject/category. This is a valid assumption that is usually considered in well-known retrieval/indexing approach such as topic modeling. Then we use these vectors in order to calculate subject-based similarity between a given query and documents.

After a brief literature review in section two, we describe our method in details in section three. Then, in section four, a series of experiments are presented to show the effectiveness of our approach, and The experimental results are analyzed. Finally, we present our conclusions in section five.

II. Literature Review

Different types of research have been done in order to improve the performance of retrieval algorithms, by considering semantic relationship between the query and documents. Tai et al. used supervised learning to improve vector space information retrieval model [7] by using matrices with 1s and 0s to show the relevance of queries and documents. Hofmann presented a statistical model based on Latent Semantic Analysis (LSA) leading to probabilistic latent Semantic Analysis

(PLSA) [8]. Maitah et al. investigated the use of an adaptive algorithm under vector space model, extended Boolean model, and language model in information retrieval [9]. Wang et al. presented a new document retrieval framework that learns a probabilistic knowledge model for improving document retrieval [10]. The model was represented by a network of association among concepts defining key domain entities and is extracted from a corpus of documents or from a domain knowledge base. Campos et al. proposed a probabilistic model based on Bayesian network for document retrieval [11] and used the network to compute posterior probabilities for the relevance of the documents. Mohebi et al. proposed a new subject-based retrieval method to retrieve all documents from a scientific digital library related to that subject. Their proposed method does not rely on user's query, rather the user specifies a specific topic or subject, and all related scientific documents related to this subject are retrieved [12]. Siddiqui proposed a hybrid IR model with two stages: first, the document collection is downsized using vector model based on a given query, second a conceptual graph based representation is used to rank the documents [13].

Sometimes retrieving the relevant text is hard because the query and the document may use different vocabularies. Mitra et al. trained a word2vec embedding model to improve the ranking of retrieved documents. In their model they map the query words into the input space and the document words into output space, and compute a relevance score by aggregating the cosine similarities across all word pairs [14].

Relevant document may be clustered together with other relevant items that may not contain query terms and could be retrieved through a clustered search [15].

Most of the methods in the literature rely completely or partially on the terms presented in the user's query. However, when a document does not contain any of the terms in the query, but is related to the query, then that document has a low chance to appear in the top retrieved documents. We address this challenge in this research by proposing a new method based on Vector Space Model. In this model a new subject-based representation for each document and the query is defined, that is independent of the query terms. Subject-based mapping of all documents in this method is a pre-processing activity that should be done once for all documents in the data-base.

III. Proposed subject-based scoring scheme

The proposed scoring scheme can be applied on a basic retrieval model such as BM25, in order to re-order the ranking of a set of retrieved documents. The proposed scheme calculates a new subject-based distance between a document and a query. This distance is a semantic-based one which calculates the relationship between a query and a given document apart from their joint terms. For this purpose, we assume that $\mathcal{D}$ is the document collection, with N documents, while every document has a set of keywords and a set of subjects associated with it. We aggregate all subjects and all keywords of all documents in set $\mathcal{S}$ and $\mathcal{K}$, respectively, i.e.:

$$\mathcal{D} = \{d_1, \ldots, d_N\}, \mathcal{S} = \{s_1, \ldots, s_M\}, \mathcal{K} = \{k_1, \ldots, k_L\}. \quad (1)$$

Our ultimate goal is to define a vector space model in order to represent each document as a subject-based vector. Consequently, the subject-based vector for each document can be compared with the subject-based vector of a given query to compute their relationship. In order to do so, we rely on the keywords for each document. In other words, we introduce a method to represent each keyword as a subject-based vector, with the size of M, to reflect how much the keyword is related to every subject. For a keyword k_l, this vector is defined as:

$$\mathbf{vk}_l = \Big(p_l(s_1),\ p_l(s_2), \ldots,\ p_l(s_M)\Big), \quad (2)$$

where $p_l(s_m)$ shows how much keyword k_l is related to subject s_m. In other words, $p_l(s_m)$ can be considered as the conditional probability that a given keyword belongs to a specific subject, defined by:

$$p_l(s_m) = P(s_m|k_l) = \frac{P(s_m, k_l)}{P(k_l)}. \quad (3)$$

We estimate this probability based on the data available in $\mathcal{D}$, as follows:

$$\hat{P}(s_m|k_l) = \frac{\sum_{d_i \in D_l} ds^i_m}{|D_l|}, \quad (4)$$

where D_l is the set of all documents with keyword k_l, and ds^i_m denotes the number of documents in D_l containing subject s_m. Finally, for a document d_i with L_i keywords, we represent d_i as a $L_i \times M$ matrix (X_i) where each row corresponds to each keyword of d_i and each column corresponds to a subject. For the sake of simplicity, we assume that $k_1, k_2, \ldots, k_{L_i}$ are keywords of d_i, then we have:

$$\mathbf{X_i} = \begin{bmatrix} \mathbf{vk}_1 \\ \vdots \\ \mathbf{vk}_{L_i} \end{bmatrix} = \begin{bmatrix} p_1(s_1) & p_1(s_2) & \ldots & p_1(s_M) \\ \vdots & & & \\ p_{L_i}(s_1) & p_{L_i}(s_2) & \ldots & p_{L_i}(s_M) \end{bmatrix}. \quad (5)$$

Now we can define a subject-based representation vector for every document, based on the matrix in 5. We call this vector $\mathbf{vs_l}$. Each component in this vector corresponds to a subject in $\mathcal{S}$, showing how much the document is related to that subject. Thus, each document d_i is mapped to a subject-based vector:

$$\mathbf{vs_i} = \frac{\sum_{l=1}^{L_i} \mathbf{vk_l}}{L_i} = \Big(\frac{\sum_{l=1}^{L_i} p_l(s_1)}{L_i}, \ldots, \frac{\sum_{l=1}^{L_i} p_l(s_M)}{L_i}\Big). \quad (6)$$

Every query Q, can also be mapped to a subject-based, M-sized, vector too. For this purpose, the query is processed first in order to extract its distinguished terms, i.e. $q_1, q_2, \ldots, q_r$. Thus, we have:

$$\mathbf{vs_q} = \frac{\sum_{l=1}^{r} \mathbf{vk_l}}{r} = \Big(\frac{\sum_{l=1}^{r} p_l(s_1)}{r}, \ldots, \frac{\sum_{l=1}^{r} p_l(s_M)}{r}\Big). \quad (7)$$

Based on the subject-based vectors for d_i and Q, a new subject-based scoring function is defined:

$$Score_{subject}(d_i, Q) = \frac{1}{\|\mathbf{vs_q} - \mathbf{vs_i}\|}. \quad (8)$$

A. Final combined scoring scheme

The proposed subject-based scoring scheme can be combined with different basic retrieval scoring schemes such as Okapi BM25 which is based on the probabilistic retrieval framework and ranks a set of documents based on the query terms appearing in each document. Given a query Q with r distinct terms, BM25 score is:

$$Score_{BM25}(d, Q) = \sum_{j=1}^{r} IDF(q_m) \frac{freq(q, d)(c+1)}{freq(q, d) + c(1 - b + b\frac{|d_j|}{avgdl})}, \quad (9)$$

where q_i is i-th term of query, $freq(q_i, d)$ is term frequency of q_i in document d, $|d|$ is the length of d in words and $avgdl$ is the average document length in the whole collection. Parameters c and b are usually chosen as $c \in [1.2, 2.0]$ and $b = 0.75$. $IDF(q_i)$ is the inverse document frequency (IDF) weight of the query term q_i and is usually calculated as:

$$IDF(q_i) = \log \frac{(N - n(q_i) + 0.5)}{(n(q_i) + 0.5)}, \quad (10)$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing term q_i. Now, we can define a combined scoring scheme based on the subject-based and BM25 scores:

$$Score_{final}(d, Q) = \alpha\ Score_{subject}(d, Q) + (1 - \alpha)\ Score_{BM25}(d, Q), \quad (11)$$

where $\alpha \in [0, 1]$ is a weighing parameter that need to be tuned. This score is applied on a set of documents retrieved based on a basic model such as BM25, in order to represent a new ranking for the retrieved documents.

IV. Experiments and Results

In order to examine the proposed method, we have considered a collection of scientific documents (articles) extracted from Web of Science (WoS) which contains all papers published from Iran in years $2013-2017$. The collection contains 98497 documents. Each document has a title, abstract, author, keywords and subjects. The subjects are assigned for each document by WoS, based on a list of predetermined categories in WoS. The collection contains 340836 keywords and 1200 different subjects. Two domain experts have classified 1200 subjects to eight main subjects including Art, Biosciences and Natural Sciences, Basic Sciences, Empirical Sciences, Humanities Sciences, Medicine and Treatment, Engineering.

In our experiment, we choose the top 100 documents for our query. Then, the top selected documents are ranked again based on $Score_{final}$. The subject-based vectors for keywords and documents in the database are calculated once. Thus, all vectors are calculated offline, and for every query presented to the system, only the corresponding vector for the query is calculated. For instance, given the following query:
"A Robust optimization for the milkrun problem under demand and travel time uncertainty",
the subject-based vector for the query is calculated based on the vectors of each term in the query, after stop word removal, as shown in Table I. $Score_{final}$ is calculated for the selected top documents based on the query $\mathbf{vs_q}$. Table II shows the $Score_{BM25}$, $Score_{subject}$, and $Score_{final}$ for the top retrieved documents, when $\alpha = 0.3$. Thus, we calculated retrieved documents changes as following:

$$Change(\%) = n_q \sum_{i=1}^{n_q} \min |R_i - i|, \quad (12)$$

while R_i is rank of i-th result in BM25 ranking and n_q is number of queries in experiment. In Fig. 1 we show how the ranking changes based on (12) in terms of α. In the proposed scoring scheme, when α is very small, the contribution of subject-based score is small, thus BM25 plays the key role in ranking the results. Alternatively, when α is large, near 1, the subject-based scoring share the most contribution in the final score. However, in a specific range, i.e. when $\alpha \in [0.25, 0.55]$, there is a competition between BM25 ranking and subject-based ranking. In this range, we see the maximum changes in the ranking between these two ranking schemes.

In order to evaluate the proposed approach on users' opinion, we have launched our model on a server and represented users the ranking obtained based on BM25 and proposed approach for a set of queries, while α changes. Then, we have asked the users to choose the best ranking. We have observed that the users prefer more the results based on the proposed scoring scheme than BM25 scoring scheme, when $\alpha = 0.3$.

TABLE I
SUBJECT-BASED VECTOR FOR QUERY: *"Robust optimization for the milkrun problem under demand and travel time uncertainty".*

Query term	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
Robust	0	0	0	0	0	0	1	0
Optimization	0	1	0	1	0	0	1	0
Milkrun	0	0	0	0	0	0	0	0
problem	0	0	0	0	0	0	0	0
demand	0	0	0	0	0	0	1	0
travel	0	0	0	0	0	0	1	0
time	0	0	0	0	0	0	0	0
uncertainty	1	1	0	0	1	1	1	0
$\overrightarrow{\mathbf{vs_q}}$	0.125	0.25	0	0.125	0.125	0.125	0.625	0

TABLE II
BM25 RANKING VERSUS PROPOSED RANKING FOR TOP 5 RETRIEVED DOCUMENTS WITH $\alpha = 0.3$

BM25 ranking	$Score_{BM25}$	$Score_{final}$	**Proposed ranking**
1	0.628	0.637	1
2	0.508	0.436	2
3	0.358	0.157	**5**
4	0.357	0.193	**3**
5	0.348	0.167	**4**

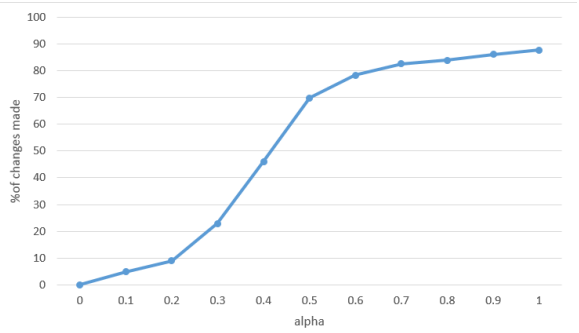


Fig. 1. How much the proposed method is able to change the ranking, when α changes. The vertical axis reflects the ranking difference between proposed method and BM25.

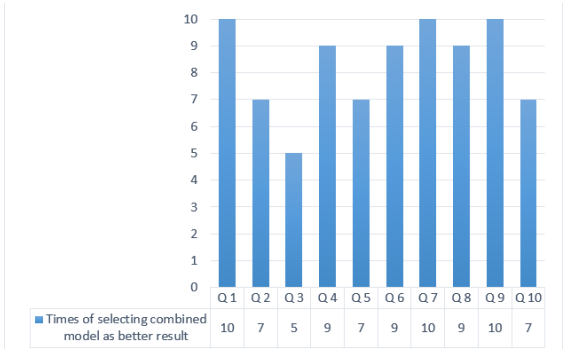


Fig. 2. How much different users prefer the proposed method versus BM25, in 100 experiments

We have also represented 10 users, 10 different queries (users and queries are independent) with both BM25 and proposed ranking scheme and ask them to choose the best results.

Based, on 100 experiments and results, we have obtained that in 83% of times, users preferred the proposed ranking scheme, as shown in Fig. 2.

V. Conclusions

This paper introduced a new vector based model for improving document retrieval, specifically when the documents are from a set of scientific databases, and each contains a set of keywords, and subjects assigned to it. A new scoring scheme is defined in which each document is represented as a vector of subjects. Based on this vector, a new subject-based scoring scheme is defined that can be combined with a basic scoring scheme such as BM25 in order to assign a new score for each document. The new scoring scheme is specifically practical when some terms in the user's query have not been appeared in the database. Thus, rather than retrieving documents based on the exact appearance of the user's term in the database, the proposed approach looks for documents related to the query conceptually, by comparing the subject-based vectorized representation. We have evaluated our proposed scoring scheme to examine how much it is able to change the results effectively, comparing with BM25. In addition we have evaluated the proposed approach based on user's satisfaction, and obtained that in 83% of times the users prefer the proposed scoring scheme than the basic frequency-based scoring scheme. For future research directions, we propose to examine other basic retrieval method rather than BM25, and combine them with the subject-based scoring scheme.

References

[1] S. Momtazi, M. Lease, and D. Klakow, "Effective term weighting for sentence retrieval," in *Research and Advanced Technology for Digital Libraries*, M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, Eds. Springer Berlin Heidelberg, 2010, pp. 482–485. [Online]. Available: https://doi.org/10.1007%2F978-3-642-15464-5_62

[2] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM, 2017, pp. 176–184. [Online]. Available: https://doi.org/10.1145%2F3130348.3130365

[3] S. Acid, L. M. De Campos, J. M. Fernández-Luna, and J. F. Huete, "An information retrieval model based on simple bayesian networks," *International Journal of Intelligent Systems*, vol. 18, no. 2, pp. 251–265, 2003. [Online]. Available: https://doi.org/10.1002%2Fint.10088

[4] J. Zhang, J. Gao, M. Zhou, and J. Wang, "Improving the effectiveness of information retrieval with clustering and fusion," *Computational Linguistics and Chinese Language Processing*, vol. 6, no. 1, pp. 109–125, 2001.

[5] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. ACM, 2006, pp. 178–185. [Online]. Available: https://doi.org/10.1145%2F1148170.1148204

[6] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1:1–1:50, 2012. [Online]. Available: https://doi.org/10.1145%2F2071389.2071390

[7] X. Tai, M. Sasaki, Y. Tanaka, and K. Kita, "Improvement of vector space information retrieval model based on supervised learning," in *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*. ACM, 2000, pp. 69–74. [Online]. Available: https://doi.org/10.1145%2F355214.355224

[8] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57. [Online]. Available: https://doi.org/10.1145%2F3130348.3130370

[9] W. Maitah, M. Al-Rababaa, and G. Kannan, "Improving the effectiveness of information retrieval system using adaptive genetic algorithm," *International Journal of Computer Science & Information Technology*, vol. 5, no. 5, p. 91, 2013. [Online]. Available: https://doi.org/10.5121%2Fijcsit.2013.5506

[10] S. Wang, S. Visweswaran, and M. Hauskrecht, "Document retrieval using a probabilistic knowledge model," in *Internation Conference on Knowledge Discovery and Information retrieval*, 2009. [Online]. Available: https://doi.org/10.5220%2F0002293400260033

[11] L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete, "A layered bayesian network model for document retrieval," in *Advances in Information Retrieval*, F. Crestani, M. Girolami, and C. J. van Rijsbergen, Eds. Springer Berlin Heidelberg, 2002, pp. 169–182. [Online]. Available: https://doi.org/10.1007%2F3-540-45886-7_12

[12] A. Mohebi, M. Sedighi, and Z. Zargaran, "Subject-based retrieval of scientific documents, case study: Retrieval of information technology scientific articles," *Library Review*, vol. 66, no. 6/7, pp. 549–569, 2017. [Online]. Available: https://doi.org/10.1108%2Flr-10-2016-0090

[13] T. Siddiqui and U. Tiwary, "A hybrid model to improve relevance in document retrieval," *Journal of Digital Information Management*, vol. 4, pp. 73 – 81, 2006 2006.

[14] E. Nalisnick, B. Mitra, N. Craswell, and R. Caruana, "Improving document ranking with dual word embeddings," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International WWW Conferences Steering Committee, 2016, pp. 83–84. [Online]. Available: https://doi.org/10.1145%2F2872518.2889361

[15] Y. Kural, S. Robertson, and S. Jones, "Clustering information retrieval search outputs," in *Proceedings of the 21st Annual BCS-IRSG Conference on Information Retrieval Research*, ser. IRSG'99. Swindon, UK: BCS Learning & Development Ltd., 1999, pp. 9–9.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 165–168
DOI: 10.15439/2018F42
ISSN 2300-5963 ACSIS, Vol. 15

Automatic intonation-based keyword extraction from academic discourse

Iurii Lezhenin*, Vadim Diachkov *, Anton Lamtev *, Artyom Zhuikov*,
Natalia Bogach*, Elena Boitsova†, Evgeny Pyshkin‡
*Institute of Computer Science and Technology Peter the Great St. Petersburg Polytechnic University
194021 St. Petersburg Polytechnicheskaya, 21 Email: bogach@kspt.icc.spbstu.ru
†Institute of Humanities Peter the Great St. Petersburg Polytechnic University
194021 St. Petersburg Polytechnicheskaya, 19 Email: el-boitsova@yandex.ru
‡Software Engineering Lab. University of Aizu
Aizu-Wakamatsu, 965-8580, Japan Email: pyshe@u-aizu.ac.jp

***Abstract*—This paper examines the perspectives of intonation processing for automatic keyword extraction. Based on a discourse intonation model from D. Brazil, automatic tone pattern recognition in speech stream is performed. It is shown that automatic classification of tone patterns can be done using simple polynomials and correlation. The original software tool *PitchKeywordExtractor (PKE)* was applied to academic discourse (on-line lectures) to extract keywords. The results were compared to the output of popular tools for speech analytics: *VoiceBase* and *IBM Watson*. All the records were processed also with Praat software and annotated by human experts. Experiments show that none of the automatic systems outperforms the others and PKE, VoiceBase and IBM Watson have the identical error rates with respect to human expert opinion. It motivates further research and supports the tendency to integrate intonation and, more generally, prosody processing in automatic keyword extraction.**

I. Introduction

AUTOMATIC keyword extraction is an important operation of textual information processing, e.g., information retrieval, summarizing, indexing, etc. Speech content occupies a large share in the overall information environment being therefore a matter for automatic keyword extraction [1]. The common practice to retrieve the keywords from speech is limited to text-based supervised and unsupervised methods applied to automatic speech recognition (ASR) output. Meanwhile, speech has its inherent feature, namely, speech prosody, that can be processed automatically to leverage keyword extraction.

Prosody processing for keyword extraction has not been thoroughly studied so far. Nevetheless, during past decades, it was repeatedly highlighted that the involvement of prosody knowledge into speech processing frameworks contributes to their performance. Even though there exists a significant diversity in phonetic and phonological approaches to prosody modelling, it is widely acknowledged, that speech prosodic markers are stable. They can be directly measured and reliably classified by means of machine learning [2], [3], [4].

Speech prosody encompasses all suprasegmental speech phenomena, but the present research is focused on only one aspect of prosody, i.e., intonation, in terms of pitch or fundamental frequency F_0. This paper addresses speech intonation in context of automatic keyword extraction in English academic discourse and contributes to the approach presented in [5] towards better understanding of applicability of computational prosodic modelling for keyword extraction and possible benefits for existing speech keyword extraction techniques.

The rest of the paper is organized as follows: Section I establishes the research background; Section II describes automatic tone pattern recognition using polynomials; Section III outlines word-to-frame mapping; Section IV presents the results of polynomial model (p-model) accuracy evaluation and cross-validation of *PitchKeywordExtractor* [5] along with two popular speech processing tools, *VoiceBase* and *Watson*; Section V summarizes the paper.

Research background for this work originates from three areas: automatic keyword extraction techniques, integration of prosody knowledge into speech processing and automatic tone pattern recognition:

A. Automatic keyword extraction techniques

Automatic keyword extraction has been a subject of extensive and detailed research in the past. An extreme demand for fast, cost-effective and accurate keyword extraction algorithms is motivated by a growing amount of digital text information. Text mining, automatic data collection indexing, extractive and abstractive text summarization, keyword-based information retrieval as well as other related tasks and applications strongly rely upon the sets of keywords (e.g., [6]).

Detailed surveys of the state-of-the-art keyword extraction techniques can be found in [7], [8], [9]. A comparative analysis of automatic keyword extraction algorithms along with text summarization challenges was presented in [10]. Existing techniques can be classified by approach as supervised and unsupervised, the latter including simple statistic, linguistics, graph-based and hybrid. Supervised techniques require annotated training data, while unsupervised operate without preliminary annotation or labelling (e.g., [7]). A comprehensive study of performance for supervised ensemble methods and base learning algorithms (Naive Bayes, support vector machines, etc.) can be found in [11].

Unsupervised methods were shown to be not less powerful than supervised ones; e.g., unsupervised morphology learning was found to produce similar results compared to a rule-based system [12]. In [9] automatic keyword extraction was performed very effectively with unsupervised graph-based keyword ranking. Keyword extraction from conversations using particle swarm optimization was shown to produce highly accurate query results [13].

B. Prosodic models in speech processing

Computational prosodic modeling integrated into speech processing workflow is a promising yet challenging area. Prosodic models have been reported to be helpful for various speech processing areas [14], e.g., automatic speech understanding (ASU), speech synthesis (TTS, text-to-speech) [15], [3], discourse tagging and segmentation [4] and automatic speaker verification [16]. It was shown that the combination of word and prosodic knowledge yielded the best results, with significant improvements over either knowledge source taken separately.

Prosodic models were found to increase speech recognition accuracy, having not been optimized for word recognition [4]. An impressive result in speech segmentation, where the prosodic model alone performed better than the language model alone [4], makes it reasonable to investigate the segmentation ability of prosodic models for keyword location within ASR output.

One of the key concepts of any prosodic model is a tone unit. In [17] the tone unit is defined as the realization of the information unit, which is extremely valuable in the context of keyword search. Both units are generated in the flow of discourse, referring to the phonological and grammatical levels respectively.

Prosodic models which motivated this research were Discourse Intonation model from D. Brazil et al. (communicative approach) [18] and Systemic Functional Linguistics of M. Hallidey et al. (grammatical approach) [17]. Both models operate with a set of tonal patterns, e.g., in Brazil model these are: *falling*, *rising*, *rising-falling*, *falling-rising* and *level*, each having a specific communicative payload. These tone patterns are connected to the categories of "given/new information" [17] or deemed to be "referring/proclaiming" [18], [19], [20], [21] This explicit relationship between intonation and meaning is exploited to search for keywords in speech.

C. Automatic tone pattern recognition

Location and classifying of tone units can be performed automatically. In [2] a 4-point model to approximate tone patterns is proposed and examined in contrast with other approximation models for tones (e.g., Bezier curves). [2] also presents a detailed study on 4-point model cascaded with several supervised classifiers and was shown to perform the best with a rule-based classifier. In [5] a continuous polynomial tone model (p-model) for Brazil tones was proposed. Functions inside p-model are used not to approximate pitch contours, but as ideal tone pattern sets to calculate correlations. Both models will be evaluated together to check p-model tone pattern recognition accuracy (see Experiment 1 in Section IV).

II. Automatic tone pattern recognition using polynomial model

Automatic tone pattern recognition implemented in PitchKeywordExtractor [5] is applied to locate a pitch pattern within a part of a record to retrieve a frame with a significant tone move. The task is to make a decision what pattern type is the closest to a frame of a record containing n readings of fundamental frequency (pitch) $F_0[k], 0 \leq k \leq n$ taken at the sample rate of f_d. $w_{min} \leq w \leq w_{max}$ is frame length range; $0 \leq l \leq n - w$ is frame shift from the first frame element. Thus, each frame contains lw-windowed signal $F_0[l : l + w]$ and one can easily see that these frames are of different length. To cope with it, a polynomial model can be easily scaled and shifted. Due to the pitch detection algorithm if for k-th sample $F_0[k]$ cannot be measured, it is defined as $F_0[k] = -1$. Median filtering is applied to smooth single pitch discontinuities.

A. Polynomial model (p-model)

We define 5 model functions ϕ_k, $k = 1..5$, which correspond to 5 Brazil tones - *falling*, *rising*, *rising-falling*, *falling-rising* and *level*. These functions are the 1st and 2nd order polynomials (Fig. 1).

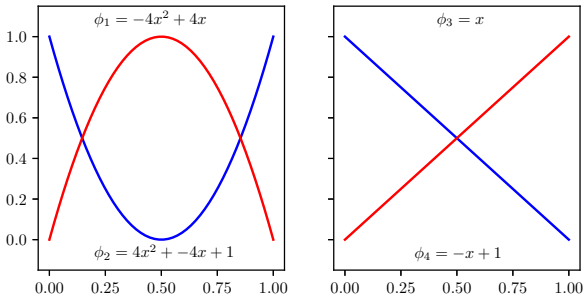


Fig. 1. p-model for Brazil tones

B. Classifier

To classify a frame by pattern type we evaluate its proximity to 4 model functions (except ϕ_5, "level"). ϕ_k, $k = 1..5$ define 5 decision regions separated by surfaces (Fig. 2).

Decision criterion to classify a frame to a region ϕ_i is

$$a_i = \frac{\sum_{k=l,F_0[k]\neq -1}^{w+l}(F_0[k] - \overline{F_0})(\phi_i((k-l)/w) - \overline{\phi_i})}{\sqrt{\sum_{k=l}^{w+l}(\phi_i((k-l)/w) - \overline{\phi_i})^2}},$$

and

$$\frac{a_i}{\sqrt{\sum_{k=l,F_0[k]\neq -1}^{w+l}(F_0[k] - \overline{F_0})}} = r_i \leq 1,$$

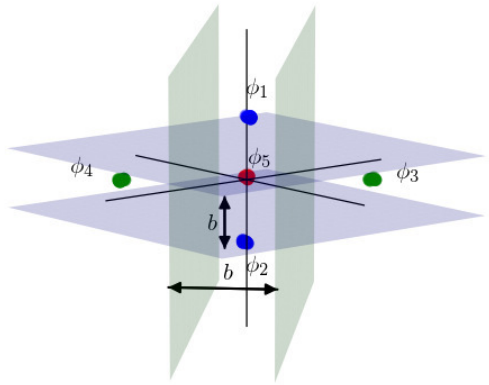


Fig. 2. Classifier decision regions $\phi_1 - \phi_4$. Parameter b sets the boundaries for *level* region ϕ_5.

where $\overline{F_0}$ is frame mean, $\overline{\phi_i}$ is model mean, r_i is normalized correlation; a_i is scaled correlation to distinguish any significant tone move from almost *level*. Decision criterion is scale and shift-invariant, thus, a_i are time and timbre-independent.

Decision is made as $k = argmax_i(a_i)$. If $max(a_k, b) \leq b$, where b is adjustable significance threshold and sets the *level* region, frame is classified as *level*, ϕ_5.

Thus, classifier outputs a pair (k, r). Frame overlaps are resolved [5] and the final frame set is successively transmitted to word-to-frame mapping.

III. WORD-TO-FRAME MAPPING

Frames where a tone move was detected during automatic tone pattern recognition and ASR output file are mapped to each other to locate a word within a frame. The goal of word-to-frame mapping is to find a word that was pronounced during a given interval defined by the frame boundaries; this word is deemed to be a keyword. Partial coincidence between segments and word timestamps is allowed and can be set as a parameter p.

Mapping in [5] could extract single words only. i. e. if one frame contained several matches with ASR output words, they were processed independently: e.g. for "computer science" the output list included both words one after another:"computer", "science". But very often a tone move refers to word collocation. Mapping is modified in order to extract keyphrases and word collocations as a single keywordlist entry. The following condition is checked sequentially for every frame:

If

$$\frac{L_{frame}}{t_2^{This_{word}} - t_1^{This_{word}}} \geq (1+p), \qquad (1)$$

then $[This_{word} \quad Next_{word}]$ is added to keyword list. This way keyphrases, e.g. of type "noun+noun" are constructed: "computer science", "artificial intelligence", "graduate student", etc.).

If

$$\frac{L_{frame}}{t_2^{This_{word}} - t_1^{Next_{word}}} \geq (1+p),$$

then $[This_{word}, Next_{word}, Next\_next_{word}]$ is added to the list to produce constructions like "noun+preposition+noun" or "particle+verb+particle/attributive construction" (e.g. "place of interest", "to follow up", "to examine closely").

A further improvement of mapping may be achieved by *break* indices processing if ToBI annotated data are available.

IV. EXPERIMENTS

PitchKeywordExtractor implementation details, libraries and tools are described in [5]. New experiments are aimed at checking the applicability of proposed p-model in comparison with one of the best existing models (4-point model) and to disclose abilities of intonation-based keyword extraction to contribute to existing speech keyword extraction techniques.

Publicly available online lectures were used as samples of academic discourse to retrieve automatically pitch patterns (Experiment 1) and extract keywords (Experiment 2). Results on three speakers are shown in Table I, II:

Speaker 1 is Benjamin Elman from Harvard University's Fairbank Center for Chinese Studies The Great Reversal: The "Rise of Japan" and the "Fall of China" after 1895 as Historical Fables.

Speaker 2 is JoAnne Stubbe, MIT 5.07SC Biological Chemistry, MIT OpenCourseWare Lexicon of Biochemical Reactions: Cofactors Formed from Vitamin B12.

Speaker 3 is Patrick Winston, MIT 6.034 Artificial Intelligence, MIT OpenCourseWare, Introduction and Scope

All the records were processed with Praat software and annotated by human experts. *Expert* row in Table II is the absolute value of agreement between two human experts about tone patterns and keyword set for each *Speaker*.

A. Experiment 1. Pattern recognition

In Experiment 1 samples of academic speech were processed to check pattern recognition ability of p-model. p-model and 4-point model [2] are evaluated together to check p-model applicability for tone pattern recognition (Table I). Both models reveal almost identical recognition recall, calculated as

$$R = \frac{T_2}{T_1} 100\%,$$

where T_1 is a number of tones in total (tone units pointed out by *Expert*), T_2 is a number of tones found automatically.

TABLE I
PATTERN RECOGNITION WITH P-MODEL AND 4-POINT MODEL

Speaker	Tones in total	$A_{p-model}$	$A_{4-point}$
Speaker 1	50	52%	49%
Speaker 2	22	36%	36%
Speaker 3	40	22%	25%

B. Experiment 2. Intonation-based keyword extraction vs. other algorithms

Cross-validation of PitchKeywordExtractor (PKE) algorithm [5] vs. *Expert* and two popular speech processing tools, *VoiceBase* and *Watson* was performed. All the sets of

TABLE II
INTONATION-BASED KEYWORD EXTRACTION VS. HUMAN EXPERTS, VOICEBASE AND WATSON

Experiment	Speaker 1	Speaker 2	Speaker 3
E (human experts)	51	53	26
PKE	54	40	24
W(Watson)	58	51	18
VB(VoiceBase)	26	50	33
$PKE \cap E$	15	12	9
$W \cap E$	16	30	10
$VB \cap E$	10	11	4
$PKE \cap VB$	8	5	1
$PKE \cap W$	9	13	4
$VB \cap W$	16	9	3

keywords were compared and their intersections were counted. Numbers in cells show absolute value of keywords found. The observations that can be done based on Table II:

1) None of the systems outperforms the others
2) All the keyword sets found by the systems of automatic extraction (*PKE*, *W* and *VB*) have nearly the same intersection with *Expert*
3) All the keyword sets found by the systems of automatic extraction (*PKE*, *W* and *VB*) have nearly the same intersections with each other
4) There exist "core" keywords, extracted by either of the systems

V. CONCLUSION

Keywords are informative milestones of speech, therefore, they are frequently marked by prosodical emphasis; that is why specific discernible prosodic characteristics (tone moves) can mark keyword presence. Prosodic features in the form of F0 estimates allow computation of pitch contours along the utterances or single words, or over the length of windows positioned in a location of interest (e.g., around a word boundary). The algorithm is based on tone and information unit boundaries juxtaposition.

The goal of this paper is to provide evidence that automatic keyword extraction systems can benefit from intonation analysis. A software tool *PitchKeywordExtractor* was evaluated along with popular tools for speech analytics and revealed the identical ability to locate the keywords. A moderate percentage in intersections of human and automatic keyword sets, pointed out either by intonation-based and other algorithms, motivates further research towards the elaboration of a hybrid approach to automatic keyword extraction.

REFERENCES

[1] Polykarpos Meladianos, Antoine J-P Tixier, Giannis Nikolentzos, and Michalis Vazirgiannis, "Real-time keyword extraction from conversations," *EACL 2017*, p. 462, 2017.

[2] David O. Johnson and Okim Kang, "Automatic prosodic tone choice classification with brazil's intonation model," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 95–109, Mar 2016.

[3] Anton Batliner and Bernd Möbius, *Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground?*, pp. 21–44, Springer Netherlands, Dordrecht, 2005.

[4] Elizabeth Shrieberg and Andreas Stolcke, "Prosody modeling for automatic speech recognition and understanding," 2002.

[5] Yurij Lezhenin, Artyom Zhuikov, Natalia Bogach, Elena Boitsova, and Evgeny Pyshkin, "Pitchkeywordextractor: Prosody-based automatic keyword extraction for speech content," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, Prague, Czech Republic, September 3-6, 2017.*, 2017, pp. 265–269.

[6] Alan Darmasaputra Chowanda, Albert Richard Sanyoto, Derwin Suhartono, and Criscentia Jessica Setiadi, "Automatic debate text summarization in online debate forum," *Procedia Computer Science*, vol. 116, no. Supplement C, pp. 11 – 19, 2017, Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

[7] Slobodan Beliga, Ana Mestrovic, and Sanda Martincic-Ipsic, "Selectivity-based keyword extraction method.," *Int. J. Semantic Web Inf. Syst.*, vol. 12, no. 3, pp. 1–26, 2016.

[8] Slobodan Beliga, "Keyword extraction techniques," 2016.

[9] Yan Ying, Tan Qingping, Xie Qinzheng, Zeng Ping, and Li Panpan, "A graph-based approach of automatic keyphrase extraction," *Procedia Computer Science*, vol. 107, no. Supplement C, pp. 248 – 255, 2017, Advances in Information and Communication Technology: Proceedings of 7th International Congress of Information and Communication Technology (ICICT2017).

[10] Santosh Kumar Bharti and Korra Sathya Babu, "Automatic keyword extraction for text summarization: A survey," *CoRR*, vol. abs/1704.03242, 2017.

[11] Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, no. Supplement C, pp. 232 – 247, 2016.

[12] Yanzhang He, Brian Hutchinson, Peter Baumann, Mari Ostendorf, Eric Fosler-Lussier, and Janet B. Pierrehumbert, "Subword-based modeling for handling oov words inkeyword spotting," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7864–7868, 2014.

[13] D. Sowmya and J.I. Sheeba, "Keyword extraction using particle swarm optimization," *Procedia Computer Science*, vol. 85, no. Supplement C, pp. 183 – 189, 2016, International Conference on Computational Modelling and Security (CMS 2016).

[14] Janet Pierrehumbert, *Prosody, intonation, and speech technology*, p. 257–280, Studies in Natural Language Processing. Cambridge University Press, 1993.

[15] Grażyna Demenko, "Intonation processing for speech technology przetwarzanie intonacji na potrzeby technologii mowy," 2012.

[16] Mustafa Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub, "Modeling dynamic prosodic variation for speaker verification," 01 1998.

[17] M. A. K. Halliday and William S. Greaves, *Intonation in the grammar of English / by M. A. K. Halliday and William S. Greaves*, Equinox Pub London ; Oakville, CT, 2008.

[18] David Brazil et al., *Discourse intonation and language teaching.*, ERIC, 1980.

[19] Miriam P. Germani and Lucia Rivas, "Discourse intonation and systemic functional phonology," *Colombian Applied Linguistics Journal*, vol. 13, no. 2, pp. 100–113, 2011.

[20] Dorothy M Chun, *Discourse Intonation in L2 – From Theory and Research to Practice*, 01 2002.

[21] Malcolm Coulthard and David Brazil, *The place of intonation in the description of interaction*, Linguistic Agency University of Trier, 1981.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 169–172
DOI: 10.15439/2018F22
ISSN 2300-5963 ACSIS, Vol. 15

Lithuanian Author Profiling with the Deep Learning

Jurgita Kapočiūtė-Dzikienė
Vytautas Magnus University
K. Donelaičio 58, LT-44248,
Kaunas, Lithuania
Email: jurgita.kapociute-dzikiene@vdu.lt

Robertas Damaševičius
Kaunas University of Technology
K. Donelaičio 73, LT-44029,
Kaunas, Lithuania
Email: robertas.damasevicius@ktu.lt

***Abstract*—We address the Lithuanian author profiling task in two dimensions (AGE and GENDER) using two deep learning methods (i.e., Long Short-Term Memory – LSTM) and Convolutional Neural Network – CNN) applied on the top of Lithuanian neural word embeddings. We also investigate an impact of the training dataset size on the author profiling accuracy. The best results are achieved with the largest datasets, containing 5,000 instances in each class. Besides, LSTM was more effective on the smaller datasets, and CNN – on the larger ones. We compare the deep learning methods with the traditional machine learning methods (in particular, Naive Bayes Multinomial and Support Vector Machine), and frequencies of elements as the feature representation). The comparison revealed that the deep learning is not the best solution for our author profiling task.**

I. Introduction and Related Work

AUTHOR Profiling (AP) is a specific subfield of Authorship Identification that aims at revealing characteristics of authors (e.g., age, gender, psychometric traits, etc.) from their writing style: synonymy and sentence structures used, grammatical or syntax errors made, etc. Thus, the AP task is solvable due to the stylometric "fingerprint" (so-called human stylome [1]): a phenomenon of individuals to express their thoughts in the written text in the specific unique ways. The stylome is also valid for the groups of individuals sharing the same demographic or psychometric characteristics. In some cases, the stylome is even attributed to the other human biometrics, as handwriting, gait or voice, and it tends to develop over time [2], depending on the age, education, social status of a person. Due to a number of potential applications in such fields as forensics, security or e-commerce, the importance of AP is constantly growing. These tasks are tackled with the automatic methods and continuous improvements of these methods contribute to the increase of the AP accuracy.

The majority of AP tasks are solved with the traditional machine learning methods and the weight vectors of features [3], [4]. The most influential examples of this field refer to Support Vector Machines (SVMs) [5], Multi-Class Real Winnow [6], Mean Proximity Clustering [7] and Holomorphic Transforms [8]. While a range of explored feature types usually covers stylistic (e.g., average sentence length, standardized type/token ratio), lexical (e.g., bag-of-words, function words), character (e.g., document or word-level character n-grams), morphological (e.g., part-of-speech tags) levels of feature representation types. The detailed description of these techniques can be found in [9].

Since methods are usually tested under different experimental conditions (various languages, profiling dimensions or datasets) it is difficult to determine, which one is the best. It is the reason why the scientific PAN competition of shared tasks plays an important role in the AP research field.[1] The comprehensive comparative analysis on the benchmark datasets reveals potential of tested methods and the new trends.

In 2013 [10], 2014 [11] and 2015 [12] PAN competition age and gender profiling was done on the English and Spanish datasets with the traditional supervised machine learning approaches: Logistic Regression, Random Forest, SVMs, etc. In 2016 PAN competition [13] the goal was to test the robustness of methods from the cross-genre perspective and SVMs were the dominant paradigm. In 2017 [14] two more languages (i.e., Arabic and Portuguese) were added to the dataset. Despite SVMs were still chosen by many participants, deep neural networks (in particular, Windowed Recurrent Convolutional Neural Network as an extension of Recurrent Convolutional Neural Network) achieved state-of-the-art performance on the gender dimension.

In the whole area of authorship identification, authorship attribution is the most explored topic for the morphologically complex Lithuanian language (the recent research work is described in [15], [16]). Unfortunately, the deep learning methods have never been applied on the Lithuanian language in any of these tasks, including AP. The aim of this research is: 1) to test their robustness on the AGE and GENDER dimensions; 2) to compare obtained results with the results produced by the traditional machine learning methods, described in [17].

II. Deep Learning Methods

Our solving task can be formulated as the supervised machine learning, where classifiers are the deep learning methods:

- *Long Short Term Memory* (LSTM) [18]. This method is a modification of Recurrent Neural Network (RNN) having a memory unit and able to learn long-term dependencies. The memory unit with input, output and forget gates is used to remember the values over arbitrary time intervals. The output with 256 nodes in the LSTM layer is an input to the fully connected softmax layer which output is the probability distribution over classes.

[1]More information about the PAN competition is in http://pan.webis.de/.

- *Convolutional Neural Network* (CNN) [19]. The convolution is performed on the sequentially connected word vectors (the detailed description is in [20]). The feature map is produced when the filters (in particular, of 3, 4, and 5 widths) are applied on each possible window of words in the text. The max-over-pooling operation on the feature map generates a single maximum value for each filter. Values from different filters are passed to a fully connected layer which outputs the probability distribution over classes.

The LSTM and CNN methods were tested using *deeplearning4j*[2] – the open-source distributed deep learning library for the Java Virtual Machine. Original method implementations were adjusted to solve only binary classification problems, therefore necessary adjustments to multi-class classification were done by the authors of this paper. All parameters were set to their default values, except for the maximum text length: i.e., it was set to 300 tokens (i.e., words or other text elements separated by spaces or punctuation) to match the maximum possible length of the input text (described in Section III-A).

Both deep learning methods were applied on the top of Lithuanian neural word embeddings (the description is in [21]), in particular, continuous bag-of-words of 300 dimensions generated with the negative sampling as the training algorithm. Since Seimas transcripts of ∼23.9 million tokens (described in Section III-A) are also the part of word embeddings corpora, our deep learning methods are protected from the out-of-vocabulary problem in all AP tasks. Despite we analyze the spoken edited language, the vocabulary of each speaker remains untouched. Since the vocabulary itself becomes one of the strongest evidence of the authorship, word embeddings should be the proper feature type for our solving task.

III. Experimental Set-Up and Results

A. Datasets

The datasets for our AP tasks are composed of the Lithuanian parliamentary text transcripts, representing speeches and debates by the Lithuanian Seimas members produced at regular parliamentary sessions and cover the period of 7 parliamentary terms from 1990 till 2013.

All texts perfectly represent formal spoken Lithuanian language, because: 1) the language of transcripts is unedited (texts match soundtracks), 2) words are grammatically correct. Only texts of the length between 100 and 300 tokens are considered, because: 1) very short texts are less informative; 2) too long texts might have the unclear authorship, i.e., long parliamentary speeches for parliamentarians might be written by someone else.

The experiments are carried out on the datasets for these dimensions:

- *AGE* dimension was composed of 6 classes (25,439 texts, 5,395,677 tokens, 161,010 types, ∼212.10 tokens/per text) related with the age intervals: *to-29* (inclusive), *30-39*, *40-49*, *50-59*, *60-69*, and *from-70* (inclusive).[3]
- *GENDER* dimension was composed of 2 classes (10,000 texts, 2,168,664 tokens, 101,951 types, ∼216.87 tokens/per text): *male* and *female*.

Each dimension was tested with 6 balanced datasets of 100, 300, 500, 1,000, 2,000, and 5,000 texts (i.e., instances) in each class. Except for the AGE dimension: the *to-29* class contained 707 and *from-70* class contained 4,732 instances at most. All datasets were composed by randomly selecting the determined number of text documents from the whole set of texts.[4]

The experiments with AGE and GENDER dimensions were performed with relevant datasets (described in Section III-A) of different sizes, containing 100, 300, 500, 1,000, 2,000, and 5,000 instances in each class.

B. Evaluation

We have tested two deep learning methods (in particular, LSTM and CNN) with the Lithuanian neural word embeddings (described in Section II) on the dataset described in Section III-A. Rough texts (without any normalization and dimensionality reduction) were given as the input. The stratified 10-fold cross-validation was used in all our experiments. The effectiveness of methods was evaluated with the *macro-accuracy* and *macro-f-score* measures (explanation is in [22]) averaged over classes and folds.

To determine if 1) obtained results are reasonable, and 2) differences between results are statistically significant, we have 1) calculated random and majority baselines, and 2) performed McNemar [23] test with one degree of freedom, respectively. The random ($\sum(P(c_j)^2)$) and majority ($max(P(c_j))$) baselines are the same in all datasets except for the AGE dimension with 1,000, 2,000 and 5,000 instances in each class (because it's classes *to-29* and *from-70* contained 707 and 4,732 instances, respectively). For the McNemar test, we have set the significance level equal to 95%, which means that the differences are considered statistically significant, if the calculated *p-value* is lower than 0.05.

The results produced by the deep learning methods with the neural word embeddings were compared to the results of the traditional classification methods (in particular, Naive Bayes Multinomial – NBM and Support Vector Machine - SVM) with the frequencies of elements as the text document feature representation. The results for NBM and SVM were taken from [17]. NBM and SVM were tested with the different feature representation types: ultimate style markers, document-level character n-grams (with *n*=[2,7]), function words, token n-grams (with *n*=[1,3]), token lemmas (with *n*=[1,3]), part-of-speech tag n-grams (with *n*=[1,3]), and n-grams of concatenated lexical and morphological features. There is no single the

[2]The deep learning library is in https://deeplearning4j.org/.

[3]The chosen grouping is also used in the largest European data archive (http://www.gesis.org) and in the Lithuanian Data Archive for Social Science and Humanities (http://www.lidata.eu).

[4]The *AMŽIUS_PROF* and *LYTIS_PROF* datasets of the *AGE* and *GENDER* dimensions, respectively, can be downloaded from http://dangus.vdu.lt/~jkd/eng/?page_id=16.

best feature representation type: it depends on the classification method and the dataset size (for the best types see Table I). Here *lemmorf* denotes lemmas + fine-grained POS information; *lex* – tokens, *lexpos* – tokens + coarse-grained POS; *lem* – lemmas; *lempos* – lemmas + coarse-grained POS; *lexmorf* – tokens + fine-grained POS information; *chr* – characters. The number next to each tag represents *n* of their n-gram.

C. Results

The results of the deep learning on the top of neural word embeddings and traditional machine learning methods with the best feature types (presented in Table I) are summarized in Figure 1. The figures do not present the *f-score* values, demonstrating the same trend as the *accuracy* values.

Figure 1 allow us to make the following claims. All obtained results are reasonable, because exceed random and majority baselines, except for LSTM with the dataset size of 100 in the GENDER dimension.

Marginally the best accuracies of 0.316 and 0.609 with the AGE and GENDER, respectively, were achieved with the CNN method and the largest datasets of 5,000 instances in each class. Besides, CNN method achieves higher profiling accuracy compared to LSTM on the larger datasets for all dimensions (with 1,000-5,000 for AGE and GENDER). Whereas, CNN is often outperformed by LSTM on the smaller datasets: with 100-500 for the AGE dimension; with 300-500 for GENDER.

According to the McNemar test, the differences in accuracies between tested LSTM and CNN methods are significant with $p < 0.05$ for the AGE and GENDER dimensions with 1,000, 2,000 and 5,000 instances in each class. For 100, 300, and 500 instance datasets for the AGE dimension are not statistically significant with $p = 0.32, 0.22, 0.19$, respectively. The p values for 100, 300, and 500 instance datasets for GENDER are 0.37, 0.99, and 0.38, respectively.

The comparison of LSTM or CNN + neural word embeddings with NBM or SVM + element frequencies as the feature representation type revealed that the deep learning methods are not the best choice for our AP tasks. The neural methods are significantly outperformed by the traditional machine learning methods: i.e., except for GENDER with 100 or 300 datasets.

The accuracies improve by increasing a number of instances in each class. In this research the purpose was to equalize the experimental conditions (in terms of dataset sizes) and to compare the effectiveness of deep learning methods with traditional machine learning methods. However, the deep learning results improve with the increase of the dataset size, whereas, e.g., NBM seems already have reached its limits (i.e., the peak on AGE and GENDER, are with 500 instance datasets, respectively). Maybe it is possible to find the breaking point where the deep learning methods reach or even bypass the effectiveness of traditional methods. Thus, the deep learning experiments with the larger datasets could be possible accuracy improvement direction for the future research.

Despite the experiments are performed with the grammatically correct texts and in-the-vocabulary words, for NBM and SVM lexical features (bag-of-words) are not always the best representation. The deep learning methods are applied on the top of neural word embeddings, however, in the future research would be useful to test the other types of embeddings (e.g., based on characters or lemmas). Moreover, the parameter (i.e., the numbers of layers, filters, or neurons in each hidden layer) tuning of LSTM and CNN could result in the higher AP accuracy, therefore this important step is also on the list of our future plans.

IV. Conclusions and Future Work

The main contribution of this research – the Lithuanian author profiling experiments with the AGE and GENDER dimensions, performed using the deep learning methods (applied on the top of neural word embeddings) that have never been tested for this task on the Lithuanian language. During this research the impact of the dataset size (with 100, 300, 500, 1,000, 2,000, 5,000 instances in each class) was also investigated. Moreover, the achieved results were compared with the traditional machine learning methods with element frequencies as the feature representation type.

The experiments on the grammatically correct texts of the Lithuanian parliamentary transcripts revealed the superiority of the Convolutional Neural Network over the Long Short-Term Memory method with the larger datasets on both profiling dimensions.

The highest accuracies of 0.316 and 0.609 on the AGE and GENDER, respectively, do not exceed the accuracies achieved by the traditional machine learning methods. Summarizing, deep learning methods are not the best choice for our profiling tasks with AGE and GENDER. Despite that in the future research we are planning to continue exploring the deep learning methods (by increasing training set sizes, tuning parameters, selecting different types of word embeddings) for the author profiling.

References

[1] H. van Halteren, R. H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. "New machine learning methods demonstrate the existence of a human stylome". *Quantitative Linguistics*, vol. 12(1), 2005, pp. 65–77.

[2] P. Juola. "Future trends in authorship attribution". *Advances in Digital Forensics III – IFIP International Conference on Digital Forensics*, vol. 242, 2007, pp. 119–132.

[3] H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño, and A. Gelbukh. "Automatic authorship detection using textual patterns extracted from integrated syntactic graphs". *Sensors*, vol. 16(9), 2016, pp. 1374, https://doi.org/10.3390/s16091374.

[4] V. Ong, A. D. S. Rahmanto, Williem, D. Suhartono, A. E. Nugroho, E. W. Andangsari, and M. N. Suprayogi. "Personality prediction based on Twitter information in Bahasa Indonesia". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, https://doi.org/10.15439/2017F359.

[5] Sh. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker. "Lexical predictors of personality type". *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[6] J. Schler, M. Koppel, Sh. Argamon, and J. W. Pennebaker. "Effects of age and gender on blogging". *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, AAAI, 2006, 199–205.

[7] R. Aljumily. "Hierarchical and non-hierarchical linear and non-linear clustering methods to "Shakespeare authorship question"". *Social Sciences*, MDPI AG, vol. 4(3), 2015, pp. 758–799, https://doi.org/10.3390/socsci4030758.

TABLE I
FEATURE REPRESENTATION TYPES WITH DIFFERENT CLASSIFICATION METHODS AND DATASET SIZES (I.E., A NUMBER OF INSTANCES IN EACH CLASS).

Dataset size	AGE		GENDER	
	NBM	SVM	NBM	SVM
100	lemmorf-2	lex-2	lexmorf-1	chr-7
300	lex-2	lemmorf-2	lempos-1	lempos-3
500	lexpos-1	lempos-1	lem-1	lexmorf-2
1,000	lempos-1	lem-3	lem-1	lem-3
2,000	lemmorf-1	lem-3	lem-1	lem-1
5,000	lem-1	lem-3	lem-1	lem-3

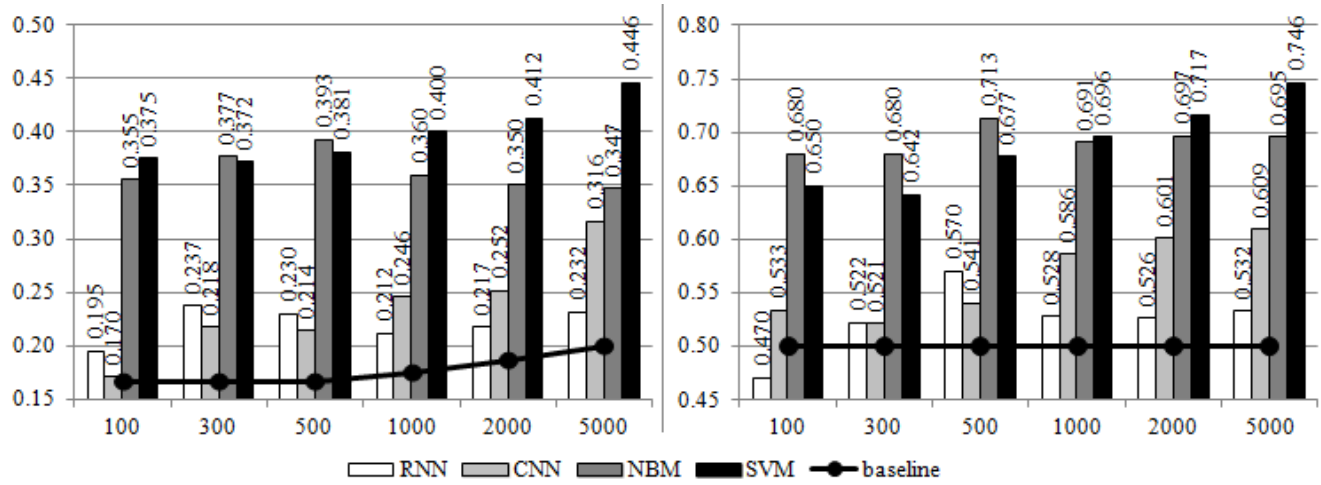


Fig. 1. AGE (left chart) and GENDER (right chart) profiling results: accuracies with different methods and dataset sizes. The *baseline* label denotes the higher value of the random and majority baselines.

[8] Ch. Napoli, E. Tramontana, G. Lo Sciuto, M. Woźniak, R. Damaševičius, and G. Borowik. "Authorship semantical identification using holomorphic Chebyshev projectors". *2015 Asia-Pacific Conference on Computer Aided System Engineering*, IEEE, 2015, https://doi.org/10.1109/APCASE.2015.48.

[9] E. Stamatatos. "A survey of modern authorship attribution methods". *Journal of the Association for Information Science and Technology*, John Wiley & Sons, Inc. vol. 60(3), 2009, pp. 538–556, https://doi.org/10.1002/asi.21001.

[10] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. "Overview of the author profiling task at PAN 2013". *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, 2013.

[11] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. "Overview of the 2nd author profiling task at PAN 2014". *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, 2014.

[12] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. "Overview of the 3rd author profiling task at PAN 2015". *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*, 2015.

[13] P. Rangel, M. Francisco, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, Martin, and B. Stein. "Overview of the 4th author profiling task at PAN 2016: Cross-Genre Evaluations". *Working Notes Papers of the CLEF 2016 Evaluation Labs*, 2016.

[14] P. Rangel, M. Francisco, P. Rosso, M. Potthast, and B. Stein. "Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter". *Working Notes Papers of the CLEF 2017 Evaluation Labs*, 2017.

[15] J. Kapočiūtė-Dzikienė, A. Venčkauskas, and R. Damaševičius. "Comparison of authorship attribution approaches applied on the Lithuanian language". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, pp. 347–351, https://doi.org/10.15439/2017F110.

[16] A. Venčkauskas, A. Karpavičius, R. Damaševičius, R. Marcinkevičius, and J. Kapočiūtė-Dzikienė. "Open class authorship attribution of Lithuanian Internet comments using one-class classifier". *Federated Conference on Computer Science and Information Systems, FedCSIS 2017. In the 2nd International Workshop on Language Technologies and Applications (LTA'17)*, 2017, pp. 373–382, https://doi.org/10.15439/2017F461.

[17] J. Kapočiūtė-Dzikienė, L. Šarkutė, and A. Utka. "Author profiling of Lithuanian parliamentary speeches: exploring the influence of features and dataset sizes". *Human Language Technologies – The Baltic Perspective: Proceedings of the 6th International Conference Baltic HLT*, IOS press, 2014, pp. 99–106, https://doi.org/10.3233/978-1-61499-442-8-99.

[18] S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural Computation*, vol. 9(8), 1997, pp. 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE*, 1998, pp. 2278–2324, https://doi.org/10.1109/5.726791.

[20] Y. Kim. "Convolutional neural networks for sentence classification". *Empirical Methods in Natural Language Processing*, EMNLP, 2014, pp. 1746–1751, https://doi.org/10.3115/v1/D14-1181.

[21] J. Kapočiūtė-Dzikienė and R. Damaševičius. "Intrinsic evaluation of Lithuanian word embeddings using WordNet". *CSOC 2018: 7th computer science on-line conference*, 2018, pp. 394-404, https://doi.org/10.1007/978-3-319-91189-2_39.

[22] M. Sokolova and G. Lapalme. "A systematic analysis of performance measures for classification tasks". *Information Processing and Management*, vol. 45(4), 2009, pp. 427–437, https://doi.org/10.1016/j.ipm.2009.03.002.

[23] Q. McNemar. "Note on the sampling error of the difference between correlated proportions or percentages". *Psychometrika*, vol. 12(2), 1947, pp. 153–157, http://doi.org/10.1007/BF02295996.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 173–177
DOI: 10.15439/2018F103
ISSN 2300-5963 ACSIS, Vol. 15

Named Property Graphs

Dominik Tomaszuk
Institute of Informatics, University of Bialystok
ul. Ciołkowskiego 1M, 15-245 Białystok, Poland
Email: d.tomaszuk@uwb.edu.pl

Łukasz Szeremeta
Institute of Informatics, University of Bialystok
ul. Ciołkowskiego 1M, 15-245 Białystok, Poland
Email: l.szeremeta@uwb.edu.pl

***Abstract*—The amount of information that is stored and processed by computer systems is constantly increasing. The relational model is still popular. Unfortunately, despite its simplicity, it has many disadvantages, which more often exclude it from large-scale applications. The property graph model seems to be a good alternative for describing real world data with its relationships. Therefore, property graph based databases become more and more popular every day. In this paper we introduce Named Property Graph model that allows to group graphs into separate units and describe information about them. We also present Cypher$_n$ query language that supports our proposal, mapping algorithms, use cases with the chemical data, and SDFEater that is our tool for processing data. Presented solutions are fully backward compatible with existing databases.**

I. Introduction and Motivation

THE amount of information that is stored and processed by computer systems is constantly increasing. The way they are stored becomes more and more important. The relational model is still popular. Unfortunately, despite its simplicity, it has many disadvantages, which more often exclude it from large-scale applications. Today, with much more attention is also looking at alternatives, e.g. graph and document databases [1], [2], [3], [4]. The graph data model is often very well suited for describing real relationships. It can be successfully used for example, on social networks. Presenting of the relationships between users, as well as the relationship of their posts, seems much more natural in this model. Users are represented by vertices, and the edges describe relationships between them. Using property graphs [5], we can additionally add some information about each person, and even that, from when they are friends.

Importantly, the property graph data model can be simply mapped back to other data models [5], [6], [7], [8]. This means that the presentation of data, for example in tabular and semi-structured form, is also possible.

The property graphs model [5], unlike the Resource Description Framework (RDF) [9], does not support named graphs [10]. In this article, we present our solution that allows to group property graphs, and give them a name and properties. This solution allows to describe graphs (context, provenance information, graph hashes and graph signatures or other metadata).

A similar approach was presented in [11], [12]. The authors propose graph collections that are logical partitions of a graph called logical graphs. These graphs are subsets of shared sets of vertices and edges. Unfortunately, logical graphs may have common vertices and edges, so it in not possible to unambiguously hash and sign these graphs.

Another approach was presented in [13] and [14]. In the first paper, Levene et al. present hypergraph which is a generalization of graphs where the concept of edge is extended to hyperedge, which relates an arbitrary set of nodes. In the second paper, the authors propose hypernodes that are directed graphs whose nodes can themselves be graphs, allowing nesting of graphs. Unfortunately, these both solutions do not support the property graph model, which is used most widely in databases.

In [15] the authors present the GOOD data model with directed labeled graph and object in database as nodes. In this model, nodes can be printable or not. In addition, two edge types are distinguished – functional and non-functional. Functional edges allow to define functional relationships between objects. GOOD data model supports edge and node labeling. Another data model was presented in [16]. The LDM data model is based on labeled directed multigraph which means graph that can have one and more edge between pair of nodes. All edges have one specific type: Basic, Product, Power or Union. LDM supports only node labels, edge labels are not supported. Unfortunately, GOOD and LDM data models do not support properties.

The GRAD data model presented in [17] is based on property graphs and extends property graphs by specific semantics. Ghrab et al. define different types of nodes, eg. entity, attribute, and literal node. Authors also introduced four types of entity edges such as association, generalization, aggregation, and composition edges. GRAD supports hypernodes which are represented as subgraphs. Unfortunately, authors do not provide any algorithms for mapping from their proposal into existing databases and query languages. Furthermore, hypernodes in GRAD do not support properties that can be used in storing metadata about subgraphs.

The paper is constructed as follows. In Section II we formalize Named Property Graph data model and propose how to map our approach into Property Graphs. Section III is devoted to a use case that presents our proposal in a molecular entities scenario. In Section IV we present tested datasets and our experiments. The paper ends with conclusions.

II. Named Property Graph

This section describes the Named Property Graph (NPG) model and shows how to map our proposal to property graphs

(Subsection II-A). Then, we discuss the possible use of our proposal. In Subsection II-B we present Cypher$_n$ that is a query language for NPGs. Then, we show a mapping algorithm that transforms our proposal into openCypher [18].

Our following proposal allows to group graphs into separate units and describe information about them. The ability to express metainformation about graphs can be required for:

- access control – ability to add additional metadata for precise access control,
- information usage control – ability to add additional metadata about authorship, license, and policy to graph in order to limit information usage,
- data syndication – ability to store and update original information,
- graph singing – allows to apply good practices, where all singing data is kept in a different graph,
- aggregation or encapsulation of graph elements – provide a wider view of the graph as it enables a higher level design and analysis,
- expressing propositional attitudes – such as trust and temporal metrics.

Potential drawbacks of Named Property Graph model is data redundancy. However, there are mechanisms for removing redundant vertices [19], [20].

A. Named Property Graph Model

According to [5], we provide a formal definition below.

Definition 1 (Property Graph): A *Property Graph* is a tuple $PG = \langle V, E, S, P, h_e, t_e, l_v, l_e, p_v, p_e \rangle$, where:

1) V is a non-empty set of vertices,
2) E is a multiset of edges, which are elements of $V \times V$,
3) S is a non-empty set of character strings,
4) P is a Cartesian product $S \times S$, where each member has a form $p = \langle k, v \rangle$,
5) $h_e : E \to V$ is a function that yields the source of each edge (head),
6) $t_e : E \to V$ is a function that yields the target of each edge (tail),
7) $l_v : V \to S$ is a function mapping each vertex to a label,
8) $l_e : E \to S$ is a function mapping each edge to a label,
9) $p_v : V \to 2^P$ is a function that assigns vertices to their multiple properties, and
10) $p_e : E \to 2^P$ is a function that assigns edges to their multiple properties.

We propose a general and simple variation on PG model, called Named Property Graphs. A named property graph is a property graph which is assigned a name (label), and properties. A name should be in the form of a string, and properties should be in the form of a set of key-value.

Definition 2 (Named Property Graph): A *Named Property Graph* is a tuple $NPG = \langle PG, N, l_n, p_n \rangle$, where:

1) PG is a Property Graph (see Definition 1),
2) N is a non-empty set of named nodes,
3) $l_n : N \to S$ is a function mapping each named node to a label, and
4) $p_n : N \to 2^P$ is a function that assigns named nodes to their multiple properties.

Algorithm 1: Mapping Named Property Graph into Property Graph

input : Named Property Graph NPG
output: Property Graph PG

```
n ← getName(NPG) ;
foreach g ∈ NPG do
    v ← getVertex(g) ;
    PG ← addVertex(v) ;
    PG ← addEdge(n, "related", v) ;
return PG;
```

Named nodes, unlike vertices, cannot connect to each other using edges. The set of all PG and NPG graphs is the Named Property Graph Database that allows to group graphs into separate units and describe information about them.

Definition 3 (Named Property Graph Database): A *Named Property Graph database* consists of a (possible empty) set of Named Property Graphs (with distinct labels) and a set of Property Graphs.

In order for our solution to work on current databases, we show the transformation of our NPG into PG in Algorithm 1. The algorithm adds one additional vertex, with the same label as the name of the PG graph, and leaves the properties assigned to it. The next step is to assign an edge to the label related to each of the vertices.

B. Cypher$_n$ Query Language

Named Property Graphs need a query language. We propose Cypher$_n$ that supports our proposal. It is a simple extension of openCypher [18]. OpenCypher a high-level declarative graph query language with an ongoing standardization work. We add `FROM` clause which specify name of property graph. Listing 1 presents Cypher$_n$.

```
MATCH (n)
FROM (m)
RETURN n, m
```

Listing 1. Cypher$_n$ example

The extension of openCypher grammar[1] that defines our query language consists of one production, which adds a FROM clause. A fragment of grammar is presented in Listing 2 in EBNF. The key fragment of grammar is shown in Fig. 1 in the form of a railroad diagram.

```
From = ((F,R,O,M), SP, NodePattern);
```

Listing 2. Extension of openCypher grammar

In order to ensure interoperability with solutions that already exist, we present Algorithm 2 that transforms our Cypher$_n$ into openCypher. The algorithm gets the name of the graph from the `FROM` clause and modifies the `MATCH` clauses so that the

[1]https://www.opencypher.org/resources

Algorithm 2: Mapping Cypher$_n$ into openCypher

input : Cypher$_n$ query CN
output: openCypher query C

$f \leftarrow$ getFormClause(CN) ;
if $f \notin \emptyset$ **then**
  $P \leftarrow$ getMatchPattern(CN) ;
  foreach $p \in P$ **do**
    $C \leftarrow$ addRelationship(p, f) ;
  $C \leftarrow$ cloneWhereAndReturnClause(CN) ;
  return C;
else
  return CR ◁ `From clause is optional` ;
return PG;

relation selects all the paths connecting the special vertex with the metadata with other vertices. The algorithm allows some graphs to be unnamed, thanks to which we retain backward compatibility with existing solutions. Listing 3 shows how openCypher query is generated from a Cypher$_n$ query given in Listing 1.

```
MATCH (n)<-[:RELATED]-(m)
RETURN n, m
```

Listing 3. OpenCypher after transformation

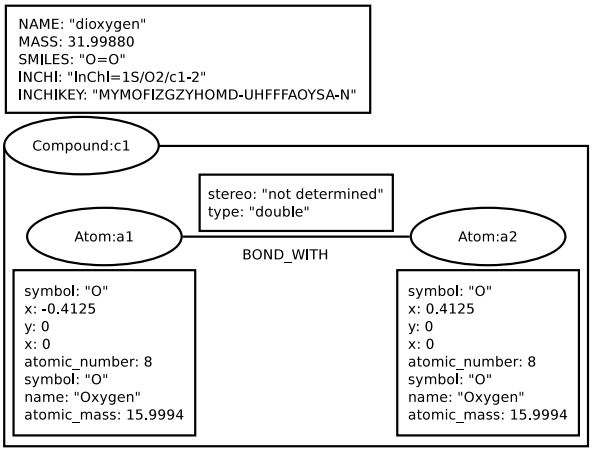


Fig. 2. Example of a named property graph

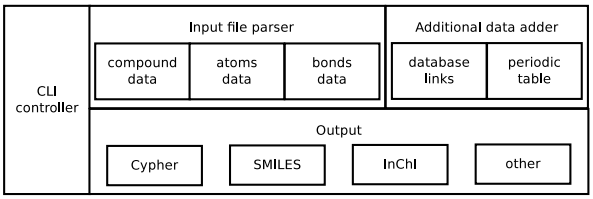


Fig. 3. Parser architecture

III. Use Case: Molecular Entity Representation

Our approach may have many practical applications. One of them is the molecular entity representation. In the standard property graphs, it is possible to describe individual atoms and their relationships. Using our solution, it is also possible to describe the entire molecule as shown in Fig. 2. In this particular case, we have compound "dioxygen" with the properties describing it. All atoms and chemical bonds are also additionally described with properties.

We have developed chemical data parser called SDFEater, available on GitHub[2] under MIT license. Our cross-platform parser is written in Java and works from the command line. It reads molecules, atoms, and bonds data from the file, and then places it in the appropriate program structures. Moreover, the program can add additional atoms data from periodic table, and tries to match the hyperlinks to the database identifiers placed in the input file. The parser accepts Structure-data file (SDF), which is part of Chemical Table file (CT File) [21] family. CT File is the collection of text formats describing chemical data. Among them there is Molfile that contains information about atoms and bonds which is stored in the tabular form. SDF, in addition to Molfile, may also contain additional information about the whole molecule such as description, Simplified Molecular Input Line Entry Specification (SMILES), International Chemical Identifier (InChI), mass, and others in the key-value form. Parser supports four output formats. One of them is the output to the openCypher query language. This allows to easily import data into the Neo4J[3] graph database.

The parser architecture is shown in Fig. 3. SDFEater has 4 main modules – CLI controller, input file parser, additional data adder, and output. The first one is responsible for operating of the command line, writing help, and running of the appropriate parts of the code depending on the selected options. In the second, the data from the source file is parsed and saved to the appropriate program structures. In the next module, additional data not included in the input file is added. Depending on the options chosen by the user, the program tries to replace the IDs listed in the compound properties with full database URL. It can also add additional information about atoms by searching the built-in data from the periodic table of chemical elements. Finally, there is the output module, where the program prints data in the appropriate format depending on the option chosen by the user.

[2]https://github.com/lszeremeta/SDFEater

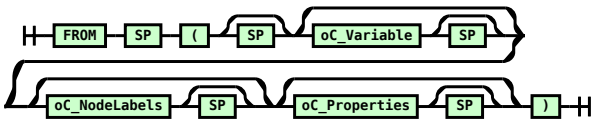


Fig. 1. Cypher$_n$ railroad diagram

[3]https://neo4j.com/

TABLE I
TEST QUERIES

Name	Query	Description
$\mathbf{Q}_1$	MATCH (c:Compound) RETURN c.CASNumber	Selects CAS number from Compound data
$\mathbf{Q}_2$	MATCH ()−[r:BOND_WITH]−>() RETURN r	Displays all information about atoms bonds
$\mathbf{Q}_3$	MATCH (a1:Atom)−[r:BOND_WITH]−>(a2:Atom) WHERE r.type = 'double' RETURN a2	Chooses second atom with "double" bond type

IV. EXPERIMENTS AND EVALUATION

In this section we evaluate the creating, loading, and querying openCypher based on our Cypher$_n$, presented in Section III. We performed Cypher$_n$ generation tests, importing data to the Neo4j graph database, as well as query tests.

The loading and query tests were performed using the cypher-shell command-line tool[4].

All experiments were executed on a laptop with quad-core AMD A6-6310 APU with AMD Radeon R4 Graphics @ 2.4 GHz (4 cores, 4 threads), single channel 2x4GB RAM (clock speed: 800 MHz, available: 6.77 GB) and 5400 RPM HDD with reading speed rated at about 95 MB/sec[5]. The system used was Ubuntu 16.04.4 LTS with Oracle Java 1.8.0_171 and Neo4j 3.3.5 graph database (community server edition).

We prepared two SDF datasets based on ChEBI [22] (subset of ChEBI complete 3-star dataset[6]) and DrugBank [23] (subset of DrugBank open structures[7]).

In the first step, we measured the performance of generating Cypher$_n$ based on the prepared SDF subsets. We used our SDFEater which is publicly available on GitHub under MIT license. The generated data in openCypher has been published on Figshare [24] under Creative Commons (CC BY 4.0) license. In total, we created 8 datasets in openCypher (2 SDF subsets and 4 variants) marked as $\mathrm{DB}_{card}^{parm}$ and $\mathrm{CB}_{card}^{parm}$, where $parm$ represents one of four variants and $card$ is the number of compounds. For example, CB_{4000}^{r} means `r` variant of ChEBI subset with 4000 compounds. Similarly, DB_{7000}^{p} is `p` variant of DrugBank subset with 7000 compounds. We distinguish the following variants of sets:

- **r** – standard SDF to openCypher,
- **p** – standard SDF with added additional periodic table data to atoms,
- **u** – standard SDF with `-u` parser option enabled (try to generate URLs to other databases instead of IDs),
- **up** – standard SDF with added additional periodic table data to atoms and `-u` parser option enabled,

The **r** variant contains only data present in SDF files. In other cases, we add extra data that are not present in the SDF datasets.

Table II shows openCypher generation times for all discussed openCypher datasets. In the case of DrugBank, a huge increase in execution time is noticeable when additional data is added to atoms. In the case of the ChEBI subset, this is visible for CB_{4000}^{u} and CB_{4000}^{up}. We also provide SDF to PubChem JSON (pcj) [25] conversion times using OpenBabel 2.4.1[8]. Comparing execution times, we can see that our parser is much faster in processing DrugBank and slower in generating ChEBI Cypher$_n$ dataset. In the case of the DB_{7000}^{r}, the Cypher$_n$ is generated almost 11.5 times faster than PubChem JSON.

In the next step, we tested importing data to Neo4j graph database. Mapping Named Property Graph into Property Graph was based on Algorithm 1. Table III shows that for both sets, the importing time grows significantly only in the case of the variant with additional data from the periodic table.

In the last step, we prepare 3 openCypher queries based on Algorithm 2. The prepared queries are representative and checks all features. These queries are presented in Table I.

Table IV shows the execution times for discussed queries. The execution times do not change significantly even in the case of different openCypher datasets variants. The exception to this are subsets with added additional periodic table data (DB_{7000}^{p}, CB_{4000}^{p}, DB_{7000}^{up}, and CB_{4000}^{up}). This is noticeable only for `Q3`.

TABLE II
GENERATION TIMES

$\mathbf{DB}_{7000}^{r}$	$\mathbf{DB}_{7000}^{u}$	$\mathbf{DB}_{7000}^{p}$	$\mathbf{DB}_{7000}^{up}$	$\mathbf{DB}_{7000}$ (pcj)
18.898 s	18.374 s	167.633 s	168.163 s	215.54 s
$\mathbf{CB}_{4000}^{r}$	$\mathbf{CB}_{4000}^{u}$	$\mathbf{CB}_{4000}^{p}$	$\mathbf{CB}_{4000}^{up}$	$\mathbf{CB}_{4000}$ (pcj)
891.308 s	4123.369 s	965.62 s	4218.856 s	111.67 s

TABLE III
IMPORTING TO NEO4J

$\mathbf{DB}_{7000}^{r}$	$\mathbf{DB}_{7000}^{u}$	$\mathbf{DB}_{7000}^{p}$	$\mathbf{DB}_{7000}^{up}$
1951.68 s	1860.018 s	4313.933 s	4665.185 s
$\mathbf{CB}_{4000}^{r}$	$\mathbf{CB}_{4000}^{u}$	$\mathbf{CB}_{4000}^{p}$	$\mathbf{CB}_{4000}^{up}$
1308.953 s	1547.778 s	2622.072 s	2834.983 s

TABLE IV
QUERIES EXECUTION TIMES

	$\mathbf{DB}_{7000}^{r}$	$\mathbf{DB}_{7000}^{u}$	$\mathbf{DB}_{7000}^{p}$	$\mathbf{DB}_{7000}^{up}$
Q_1	3.586 s	3.616 s	3.659 s	3.636 s
Q_2	7.478 s	7.314 s	7.654 s	7.951 s
Q_3	6.188 s	6.208 s	8.894 s	8.909 s
	$\mathbf{CB}_{4000}^{r}$	$\mathbf{CB}_{4000}^{u}$	$\mathbf{CB}_{4000}^{p}$	$\mathbf{CB}_{4000}^{up}$
Q_1	3.336 s	3.336 s	3.354 s	3.319 s
Q_2	6.37 s	6.581 s	6.29 s	6.426 s
Q_3	5.223 s	5.402 s	6.339 s	6.387 s

[4]https://neo4j.com/docs/operations-manual/current/tools/cypher-shell/
[5]tested using `hdparm -t`
[6]https://www.ebi.ac.uk/chebi/downloadsForward.do
[7]https://www.drugbank.ca/releases/latest#open-data
[8]http://openbabel.org/wiki/Main_Page

V. Conclusions

The property graph model is increasingly used in databases. We present a Named Property Graph model which allows to group graphs into separate units and describe information about them. Named Property Graphs provide a high-value and incremental change to the property graph model. We also introduce Cypher$_n$ query language that supports our proposal. In the paper we also present mapping algorithms, use cases with the chemical data, and SDFEater that is our tool for processing data. Our proposal can be easily applied to existing databases. The results of the experiments show the good potential of the presented solutions.

The future work will focus on providing support for temporal, uncertainty, and trust metrics. Another challenge is to find a relationship between our solution and hypernodes. We would also like to focus on developing methods for transforming our solution into other query languages.

Acknowledgments

This publication has received financial support from the Polish Ministry of Science and Higher Education under subsidy for maintaining the research potential of the Faculty of Mathematics and Informatics, University of Bialystok.

References

[1] J. Webber, "A programmatic introduction to Neo4J," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, ser. SPLASH '12. New York, NY, USA: ACM, 2012. doi: 10.1145/2384716.2384777. ISBN 978-1-4503-1563-0 pp. 217–218. [Online]. Available: http://dx.doi.org/10.1145/2384716.2384777

[2] C. Tesoriero, *Getting started with OrientDB*. Packt Publishing Ltd, 2013. ISBN 978-1782169956

[3] L. Dohmen, "Algorithms for large networks in the NoSQL database ArangoDB," Bachelor's Thesis, RWTH Aachen University, Aachen, 2012.

[4] K. Chodorow, *MongoDB: The definitive guide: powerful and scalable data storage*. O'Reilly Media, Inc., 2013. ISBN 978-1449344689

[5] D. Tomaszuk, "RDF data in property graph model," in *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*. Springer, 2016. doi: 10.1007/978-3-319-49157-8_9 pp. 104–115. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-49157-8_9

[6] R. De Virgilio, A. Maccioni, and R. Torlone, "Converting relational to graph databases," in *First International Workshop on Graph Data Management Experiences and Systems*, ser. GRADES '13. ACM, 2013. doi: 10.1145/2484425.2484426. ISBN 978-1-4503-2188-4 pp. 1:1–1:6. [Online]. Available: http://dx.doi.org/10.1145/2484425.2484426

[7] R. De Virgilio, A. Maccioni, and R. Torlone, "R2G: A tool for migrating relations to graphs," in *Proceeding of the 17th International Conference on Extending Database Technology (EDBT 2014)*, 2014, pp. 640–643.

[8] S. Lee, B. H. Park, S. H. Lim, and M. Shankar, "Table2Graph: A scalable graph construction from relational tables using Map-Reduce," in *2015 IEEE First International Conference on Big Data Computing Service and Applications*, March 2015. doi: 10.1109/BigDataService.2015.52 pp. 294–301. [Online]. Available: http://dx.doi.org/10.1109/BigDataService.2015.52

[9] G. Schreiber and Y. Raimond, "RDF 1.1 Primer," W3C, W3C Note, 2014. [Online]. Available: http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

[10] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named graphs, provenance and trust," in *Proceedings of the 14th International Conference on World Wide Web*. ACM, 2005. doi: 10.1145/1060745.1060835 pp. 613–622. [Online]. Available: http://dx.doi.org/10.1145/1060745.1060835

[11] M. Junghanns, A. Petermann, N. Teichmann, K. Gómez, and E. Rahm, "Analyzing extended property graphs with Apache Flink," in *Proceedings of the 1st ACM SIGMOD Workshop on Network Data Analytics*, ser. NDA '16. New York, NY, USA: ACM, 2016. doi: 10.1145/2980523.2980527. ISBN 978-1-4503-4513-2 pp. 3:1–3:8. [Online]. Available: http://dx.doi.org/10.1145/2980523.2980527

[12] M. Junghanns, P. André, and R. Erhard, "Distributed grouping of property graphs with GRADOOP," in *Datenbanksysteme für Business, Technologie und Web (BTW 2017)*, B. Mitschang, N. Daniela, L. Frank, S. Harald, H. Melanie, T. Jens, H. Theo, K. Oliver, and W. Matthias, Eds. Gesellschaft für Informatik, Bonn, 2017, pp. 103–122.

[13] M. Levene and A. Poulovassilis, "An object-oriented data model formalised through hypergraphs," *Data & Knowledge Engineering*, vol. 6, no. 3, pp. 205–224, 1991. doi: 10.1016/0169-023X(91)90005-I. [Online]. Available: http://dx.doi.org/10.1016/0169-023X(91)90005-I

[14] M. Levene and A. Poulovassilis, "The hypernode model and its associated query language," in *Information Technology, 1990. 'Next Decade in Information Technology', Proceedings of the 5th Jerusalem Conference on (Cat. No.90TH0326-9)*, Oct 1990. doi: 10.1109/JCIT.1990.128324 pp. 520–530. [Online]. Available: http://dx.doi.org/10.1109/JCIT.1990.128324

[15] M. Gyssens, J. Paredaens, and D. V. Gucht, "A graph-oriented object model for database end-user interfaces," *ACM SIGMOD Record*, vol. 19, no. 2, pp. 24–33, 1990. doi: 10.1145/93605.93616. [Online]. Available: http://dx.doi.org/10.1145/93605.93616

[16] G. M. Kuper and M. Y. Vardi, "The logical data model," *ACM Transactions on Database Systems (TODS)*, vol. 18, no. 3, pp. 379–413, 1993. doi: 10.1145/155271.155274. [Online]. Available: http://dx.doi.org/10.1145/155271.155274

[17] A. Ghrab, O. Romero, S. Skhiri, A. Vaisman, and E. Zimányi, "Grad: On graph database modeling," *arXiv preprint arXiv:1602.00503*, 2016.

[18] J. Marton, G. Szárnyas, and D. Varró, "Formalising openCypher graph queries in relational algebra," in *Advances in Databases and Information Systems*, M. Kirikova, K. Nørvåg, and G. A. Papadopoulos, Eds. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-66917-5_13. ISBN 978-3-319-66917-5 pp. 182–196. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-66917-5_13

[19] R. Zhou and E. A. Hansen, "Parallel Structured Duplicate Detection," in *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'07. AAAI Press, 2007. ISBN 978-1-57735-323-2 pp. 1217–1223.

[20] D. Tomaszuk and K. Pak, "Reducing vertices in property graphs," *PLOS ONE*, vol. 13, no. 2, pp. 1–25, 02 2018. doi: 10.1371/journal.pone.0191917. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0191917

[21] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited," *Journal of Chemical Information and Computer Sciences*, vol. 32, no. 3, pp. 244–255, 1992. doi: 10.1021/ci00007a012. [Online]. Available: http://dx.doi.org/10.1021/ci00007a012

[22] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1214–D1219, 2016. doi: 10.1093/nar/gkv1031. [Online]. Available: http://dx.doi.org/10.1093/nar/gkv1031

[23] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 2018. doi: 10.1093/nar/gkx1037. [Online]. Available: http://dx.doi.org/10.1093/nar/gkx1037

[24] Ł. Szeremeta and D. Tomaszuk, "SDFParser example Cypher outputs," 5 2018. doi: 10.6084/m9.figshare.6249962.v1. [Online]. Available: http://dx.doi.org/10.6084/m9.figshare.6249962.v1

[25] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, and J. Zhang, "PubChem BioAssay: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D955–D963, 2016. doi: 10.1093/nar/gkw1118. [Online]. Available: http://dx.doi.org/10.1093/nar/gkw1118

1st International Workshop on AI Methods in Data Mining Challenges

THE scope of DMGATE is to establish a forum for data scientists working at the edge of academic research and commercial applications, keen on developing and using AI-based methods in data mining projects and challenges. DMGATE continues the tradition of data mining competitions organized at FedCSIS.

This year's competition starts on April 3 and lasts until May 7. The task is to predict win-rates of Hearthstone decks used by players with various skill and experience. Details are available here.

As usual, competition winners will be granted with financial awards and/or free conference registration. Moreover, we will invite a small group of authors of the most interesting competition solutions to submit papers that will be reviewed on special fast-track basis.

Starting from this year, we also welcome regular paper submissions describing new approaches for analyzing data sets published online during previous competitions. These papers will be peer-reviewed by top specialists in the area of data science and – if accepted – they will be presented at the DMGATE session together with the papers prepared by this year's competition winners.

We believe that it is highly important for the data science community to keep up the discussion on both current and past data mining competitions, in order to jointly develop a firm knowledge base on how to apply AI-based data mining methods in real life.

Event Chairs

- **Janusz, Andrzej,** University of Warsaw, Poland
- **Ślęzak, Dominik,** University of Warsaw, Poland

Program Committee

- **Carrizosa, Emilio,** Universidad de Sevilla, Spain
- **Kayakutlu, Gulgun,** Istanbul Technical University, Turkey
- **Raghavan, Vijay,** University of Louisiana at Lafayette, United States
- **Stefanowski, Jerzy,** Poznan University of Technology, Poland
- **Weber, Richard,** University of Chile, Chile
- **Woźniak, Michał,** Wroclaw University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 181–184
DOI: 10.15439/2018F364
ISSN 2300-5963 ACSIS, Vol. 15

Regression networks for robust win-rates predictions of AI gaming bots

Ling Cen
EBTIC, Khalifa University, UAE
cen.ling@kustar.ac.ae

Andrzej Ruta
ING Bank Slaski, Katowice, Poland
andrzej.ruta@ingbank.pl

Dymitr Ruta
EBTIC, Khalifa University, UAE
dymitr.ruta@kustar.ac.ae

Quang Hieu Vu
Zalora, Singapore
quanghieu.vu@zalora.com

***Abstract*—Designing a robust and adaptable Artificial Intelligence (AI) opponent in a computer game would ensure the game continues to challenge, immerse and excite the players at any stage. The outcomes of card based games such as "Heartstone: Heroes of Warcraft", aside the player skills, heavily depend on the initial composition of player card decks. To evaluate this impact we have developed a new robust regression network in a context of the AAIA Data Mining Competition 2018, which tries to predict the average win-rates of the specific combinations of bot-player and card decks. Our network is composed of 2 levels: the entry level with an array of finely optimized state of the art regression models including Extreme Learning Machines (ELM), Extreme Gradient Boosted decision tree (XGBOOST), and Least Absolute Shrinkage and Selection Operator (LASSO) regression trained via supervised learning on the labeled training dataset; and just a single ELM at the 2^{nd} level installed to learn to correct the predictions from the 1^{st} level. The final solution received the root of the mean squared error (RMSE) of just 5.65% and scored the 2^{nd} place in AAIA'2018 competition. This paper also presents two other runner-up models receiving RMSE of 5.7% and 5.86%, scoring the 4^{th} and the 6^{th} place respectively.**

I. Introduction

Computer games, or more precisely computer-controlled games where players interact with objects displayed on computer screens, provide entertainment [1] and challenge players' physical and mental abilities. Beside entertainment, playing computer games has been found to combat stress, promote health and keep brain fit and active [2]. In recent years, fast development and penetration of Internet, multi-medial graphic devices, emergence of virtual reality, on-line open games led to the rapid growth of gaming popularity and combined with improved affordability, accessibility, ease and customization of gameplay, opponents choices, have driven the game industry to the enormous success and a bright future ahead [2].

To keep players interested and enthralled, computer games usually offer various stages and complexity levels to suit people from beginners to masters, and keeping them equally entertained for as long as possible. The fun of computer games is magnified when players play against their friends or other opponents from all over the world in on-line games since human opponents guarantee fresh, distinctive and engaging challenge [2]. With the recent advancement in Machine Learning (ML) and the Internet of Things (IoT), Artificial Intelligence (AI) has attracted increasing attention and heavily penetrated many industries including gaming industry. In many computer games, designing a robust and adaptable AI opponent would ensure the games continues to challenge, immerse and excite the players at any stage, which is one of the most important aspects of success.

In the card based games such as Heartstone: Heros of Warcraft, aside the player skills, the outcomes heavily depend on the initial composition of card decks. To evaluate this impact, 2018 Advances in Artificial Intelligence and Applications (AAIA) Data Mining Competition was proposed and focused on the prediction of win-rates of 4 AI bot players, playing the Heartstone game among each other with different initial decks of cards and hero characters. The objective of the competition was to use these data to build the prediction model capable of accurately estimating win-rates of the same 4 AI bots but playing with one of the 200 new test card decks, gameplay of which and their results were not available to the contestants.

This paper presents a new robust shallow regression network to predict the average win-rates of the specific combinations of bot-player and card decks in a response to the context of AAIA Data Mining Competition 2018. Our network is composed of two levels. The first level is built with an array of individually trained regression models that have proven to be effective for sparse binary regression problems, including Extreme Learning Machine (ELM), Extreme Gradient Boosted Decision Tree (XGBOOST) and the Least Absolute Shrinkage and Selection Operator (LASSO) regression models, while the second level contains only a single ELM that learns to correct the predictions from the preceding level. The final solution submitted as a competitive entry in the AAIA'2018 Data Mining Competition received the RMSE of 5.65% and scored the 2^{nd} place, marginally trailing the winning solution.

The remainder of the paper is organized as follows. AAIA Data Mining Competition 2018 is introduced in Section II. The feature extraction method and regression network for predicting the average win-rates of the specific combinations of bot-player and card decks are presented in Sections III and IV, respectively. The experimental results obtained through model evaluation are summarized in Section V, followed with a discussion in VI and the concluding remarks provided in Section VII.

II. Competition Description

The AAIA Data Mining Competition 2018 is related to the turn-based card game of "Heartstone: Heros of Warcraft". In this game, two players choose their heroes with a unique power and compose a deck of thirty cards that represent various spells, weapons, and minions, and can be summoned in order to attack the opponent with the goal of reducing the opponent's health to zero and win the game. The outcomes of the game, aside the player skills, heavily depend on the initial composition of player card decks. To evaluate this impact, the competitors were expected to predict win-rates of four AI bot players, automatically playing many games against each other with different initial decks of cards and hero characters.

The training data provided by the competition contained a collection of JSON files describing in detail more than 300k games played by all pairs from the set of 4 different bots, each starting with one of 400 unique Hearthstone card decks. The data included the initial composition of card decks, heroes selected, the results of each game, and detailed turn-by-turn gameplay states and related statistics. The objective of the competition was to utilize these datasets to build the prediction model capable of accurately predicting win-rates of the 4 AI bots assigned to any previously unseen composition of card decks and related class of hero character. To evaluate the competitive models the win rates of all 4 bots were tested in combinations with specific 200 new test decks, however this time provided without any gameplay nor game results details to the contestants to properly simulate realistic predictive power of competing win-rates prediction models.

The solutions were evaluated using the root of the mean squared error (RMSE) measure. The preliminary score of each submitted solution was evaluated externally on a fixed 10% subset of the full test records and published on the competition leaderboard. The final evaluation on the complete testing set was performed after the completion, i.e. when the competitors submitted their final solutions with no further changes allowed.

III. Feature Engineering

Estimation of average win-rates of the specific combination of bot-player and card decks can be solved via regression analysis that is a methodology for estimating the relationships between a dependent variable (response) and one or multiple independent variables (predictors). The dependent variable here was the win-rate expressed as a continuous real number from the $[0, 1]$ interval.

From the outset it has been decided, that since no gameplay details, beyond the initial deck, was available in the test stage, the training data need to be trimmed consistently down to the same content. It included the id of the player-bot and the initial Heartstone deck composition, i.e. the id of one of the 9 distinct hero characters and the cardinalities (0,1, or 2) of other cards from the pool of over 300 available card types. All above were cascaded to form a feature vector as shown in Fig. 1.

The initial modeling tasks involved generating features from the available data and after a brief experimentation with simple

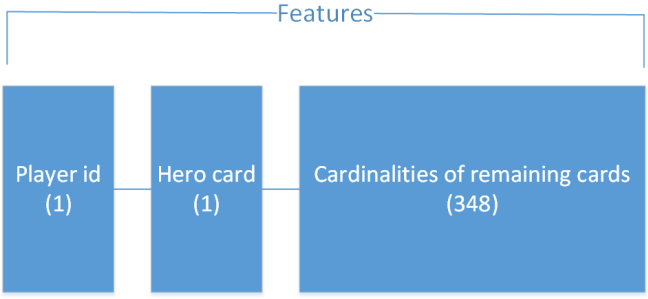


Figure 1. Feature representation.

linear regression models, the highest predictive power associated with the win-rate predictions appeared to come from numerical encoding of raw categorical features. The player id took the values of [1,2,3,4] representing the 4 bot-players, and the hero card took the values of [1,2,...,9] representing the 9 hero characters. The remaining card features took the values of [0,1,2] depending on the cardinality of specific card types in the decks. For each data record associated with a single game, this formed a sparse 348-dimensional vector describing the cardinalities of card types appearing both in the training and test sets. The final feature set included $1 + 1 + 348 = 350$ features as shown in Fig. 1.

Initial feature selection experiments did not result in any improvement of the cross-validated performance measure, although in-sample (training-set) RMSE was reduced significantly after selection of around 100 greedily found card features. To prevent model overfitting, it was decided to include all 350 features in the model building phase. With these features, a robust regression network has been developed for predicting win-rates of four AI bots playing the "Heartstone: Heros of Warcraft" game against each other with different initial decks of cards and hero characters, which will be elaborated further in the following section.

IV. Regression Networks

Artificial neural networks (ANNs) have been successfully applied in various fields due to their ability to approximate complex nonlinear mappings directly from input samples as well as model natural and artificial phenomena that are difficult to express using classical parametric techniques. Gradient-based learning algorithms are commonly used to train neural networks and tune the parameters iteratively, which, however, requires long training time.

To improve learning efficiency of neural networks, Huang and his colleagues proposed extreme learning machines (ELMs) that are feed-forward neural networks with a single or multiple layers of hidden nodes. Instead of tuning the parameters of hidden nodes, the ELMs randomly choose hidden nodes and analytically determine the output weights of the network [6]. In Comparison to many state-of-the-art computational intelligence methods, such as the conventional back-propagation (BP) algorithm and Support Vector Machines

(SVM), ELMs have the advantage of much faster learning rate, ease of implementation, the least human intervention, and better generalization performance in terms of lower training error and smaller norm of weights. It has been reported by Huang et al. based on their experimental results that ELMs are able to achieve better generalization performance and learn thousands of times faster than traditional learning algorithms for feed-forward neural networks [6].

In order to extend the generalization performance of the ELM, a novel shallow regression network composed of 2 stages has been developed. In the first stage an array of finely optimized state-of-the-art regression models are trained directly on the input data to predict the desired regression outputs. The models shortlisted for this stage based on best preliminary ad-hoc evaluation included beside kernelized ELMs, XGBOOST, LASSO, SVM, Gaussian process (GP) and simple Multi-Layer Perceptron (MLP) models.

The outputs of all base models, i.e. the proposed regression outputs are passed on to the second and final stage of the shallow network in which just a single or multiple regression are trained again, however this time their inputs are multiple propositions of the predicted outputs, hence their role is just to learn to optimally correct multiple predictions to minimize final regression error. The decision to limit such corrective layers to just a single 2^{nd} layer follows from extensive experimentations which confirmed that adding more corrective layers does not improve the performance but only contributes to the network complexity.

We have dedicated a lot of experimentation to the selection of the best subset of primary regressors as well as the final stage corrective models. We have, however consistently received ELM to be the single most effective 2^{nd} stage corrective regressor, while also in the primary first layer ELM appeared to dominate in terms of performance but showed the best overall results if combined in the first layer with XGBOOST and LASSO regression models only.

A structure of the best performing network with 9 base kernelized ELMs, 1 LASSO and 1 XGBOOST models in the primary layer and a single ELM in the final layer is shown in Fig. 2.

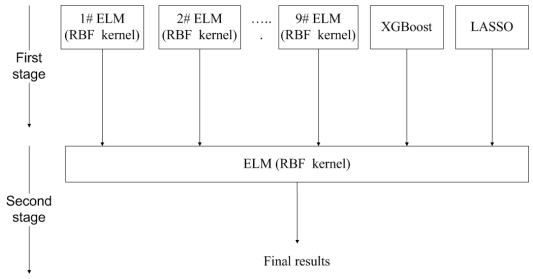


Figure 2. A sample structure of the learning model.

Multiple ELM models with radial-basis-kernels of increasing width parameter (gamma) from 20 to 60 dominated the first layer of the network. The RBF kernel is defined as [7]

$$K_{\text{RBF}}(x, x') = exp(-\frac{||x - x'||^2}{2\sigma^2}), \quad (1)$$

where $\gamma = 2\sigma^2$.

As mentioned above, these 9 ELM models in the optimized network setup have been complemented with just a single XGBOOST and LASSO models, therefore for completeness few details on only the added models are provided below.

- A decision tree builds a regression model in the form of a tree structure, which breaks down a dataset into multiple smaller subsets and incrementally builds a tree with decision nodes and leaf nodes for the purpose of classification or regression. XGBOOST, based on Extreme Gradient Boosting model [3], is an implementation of the gradient boosted decision trees algorithm with a goal of pushing the limit of compute resources for boosted tree algorithms [4]. In recent years, XGBOOST, due to its advantages of fast processing speed and high prediction accuracy, has been employed by many winning teams of a number of machine learning competitions, e.g. [5].
- LASSO regression is a shrinkage and variable selection technique aimed at enhancing the prediction accuracy and interpretability of the linear regression model it produces [8], [9], [10]. It attempts to find a subset of predictors that minimize the prediction error of the response variable, which is achieved by imposing a constraint on model parameters to make regression coefficients for some predictor variables shrink down to 0. Given the feature vectors encoding cardinalities of cards are very sparse, LASSO is employed as another base regression model in the first stage of our network. It attenuates and effectively excludes certain variables from the model, while the variables with non-zero coefficients are considered as strongly associated with the target variable.

Among other primary models that deserve some attention despite not being selected to the final network was a Multi-Layer Perceptron (MLP) with variable number of neurons. Among a wide range of configurations trialled we found a network with 50 input neurons, one hidden layer of size 20, and a single linear-activation output neuron to be the best performing model of this kind. Rectified Linear Unit (ReLU) activation [11] was set for all input- and hidden-layer neurons. It should be noted however, that we managed to maximize the generalization performance of this network only after introduction of recently popular regularization techniques: batch normalization and dropout [12] after the first two dense layers. We decided to use this particular model as a benchmark model for our regression network, yet did not include it in the network itself.

Each of the base regression models in the first stage was individually trained over the whole training set. The second stage was built on top of the first stage with a goal of learning to correct its predictions. Experimentations concluded very decisively that just a single ELM with optimized hyper-parameters is best at learning to correct the primary regressors'

outputs and hence to further improve the generalization ability of the whole network. As a result, the entire regression network became a hybrid model with a decision level fusion in the top layer realized using the ELMs. It was very important, however, for the robustness of the emerging 2-level regression network to train the second layer on the cross-validated outputs of the first layer such that the second layer regression used only out-of-sample rather than in-sample prediction outputs.

V. Experiment Results

As already partly explained in the previous section, many experimental trials were performed to determine the best composition of the first and the second stages of the regression network as well as optimize all the individual and joint hyper-parameters. All the experiments were based upon both k-fold cross-validation over the training dataset and the external feedback in a form of performance scores published in the web-based KnowledgePit platform and calculated for only 10% of the test examples. Eventually, the best structure of the network consists of 9 kernelized ELMs, an XGBOOST, and a LASSO regression models in the first level that are connected to another ELM model in the 2^{nd} level, is shown schematically in Fig. 2.

The parameters of the individual regression models were optimized over the k-fold cross-validated training set using Bayesian or grid optimization. The optimal network setup included 9 ELM models with radial-basis-function kernels of width [20,25,30,35,40,45,50,55,60], XGBOOST model with learning rate 0.01, re-sampling rate 0.2, maximum tree depth 2 and 100000 iterations, and the LASSO regression model with 100 lambdas and up to 100 non-zero weights. The ELM in the second stage used RBF kernel with a small width $\gamma < 1$.

The final solution that we submitted to the competition received the RMSE of 5.0% based on the preliminary evaluation on the 10% of all test examples, and the final score of 5.65% on the whole test set. The best RMSE scores on the preliminary leaderboard evaluation achieved individually using each base regression model were 5.88% for ELM with 40-wide RBF kernel, 6.64% for XGBOOST, and 6.87% for LASSO. For comparison, our benchmark single-stage MLP regression model achieved RMSE of 5.69% on the same 10% subset of the test set and 5.86% on the whole test set (6^{th} best score), showing robustness to over-fitting yet still remaining slightly behind the proposed two-level regression network.

The above figures prove that the introduction of the shallow hierarchy with just a single regressor in the 2^{nd} level was an adequate choice leading to a noticeable performance improvement compared to the base models.

VI. Discussion

It is found that better individual performers of base models may not lead to better combined output. Indeed the removal of GP and SVM regressors, although individually top in-sample performers, surprisingly led to improved performance of the whole network.

To further improve the network performance we have introduce specific regularization filter applied on the final test outputs in order to enforce similar global (higher order) statistics observed in the training set. The filter included 3 constraints: shift towards the desired mean, stretching or compressing the variance around the desired mean and forcing the shift of the differences among bot-player individual win-rates towards the same relative differences observed in the training set.

Deeper structures with multiple concatenated ELMs in the 2^{nd} level have also been tested to no statistically significant improvement in the generalization ability of the network compared to the architecture shown in Fig. 2. If 2 ELMs were concatenated in the 2^{nd} stage, with different kernel widths, the resulting preliminary test RMSE was in a range of $[5.05, 5.1]$. Similarly, a network with a 4-ELMs chain in the second level received the same RMSE of 5.1%. These observations indicate that further attempts to correct regression errors bring no additional value to the design instead just modeling propagated noise and bringing re-optimization overhead.

VII. Conclusions

The regression network presented in this paper has been developed and submitted as a competitive entry to the AAIA Data Mining Competition 2018, concerned with the prediction of win-rates of four AI bot players, playing the game "Heartstone: Heros of Warcraft" among each other with different initial decks of cards and hero characters. The proposed regression was hierarchically designed to combine the advantages of Extreme Learning Machine and few other complementary state-of-the-art regression models in the first level and improve the final performance through supervised decision fusion and error correction in the second level. Our solution received the final RMSE of 5.65% and scored the 2^{nd} place in AAIA'2018 Data Mining Competition.

References

[1] ScienceDaily, https://www.sciencedaily.com/.
[2] Five reasons why online games have become so popular, https://www.belfasttelegraph.co.uk/woman/life/five-reasons-why-online-games-have-become-so-popular-28656803.html
[3] J.H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189-1232, 2001.
[4] XGBoost, https://github.com/dmlc/xgboost/.
[5] XGBoost:Machine Learning Challenge Winning Solutions, https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions, Retrieved 2016-08-01.
[6] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489-501, 2006.
[7] J.P. Vert, K. Tsuda, and B. Schölkopf "A primer on kernel methods," *Kernel Methods in Computational Biology*, 2004.
[8] R. Tibshirani, "Regression Shrinkage and Selection via the lasso," *Journal of the Royal Statistical Society, Series B (methodological)*, Wiley, vol .58, no. 1, pp. 267–88.
[9] L. Breiman, "Better Subset Regression Using the Nonnegative Garrote," *Technometrics. Taylor and Francis*, vol. 37, no. 4, pp. 373–384, 1995.
[10] R. Tibshirani, "The lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, vol. 16, pp. 385–395.
[11] V. Nair and G. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *27th International Conference on Machine Learning*, pp. 807-814, 2010.
[12] N. Srivastava, G. Hinton, A. Krishevsky, I. Sutskever, and R. Salakhutdinov, "A Simple Way to Prevent Neural Networks from Overfitting", *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 185–188
DOI: 10.15439/2018F365
ISSN 2300-5963 ACSIS, Vol. 15

A Neural Network Approach to Hearthstone Win Rate Prediction

Jan Jakubik
Wrocław University of Science and Technology
Faculty of Computer Science and Management, Department of Computational Intelligence
Wrocław, Poland
Email: jan.jakubik@pwr.edu.pl

***Abstract*—This paper describes a solution to the AAIA'18 data mining challenge, which concerns prediction of win rates for decks in Hearthstone collectible card game. A neural network model assigning win rate to decks is learned based on maximisation of log probability of observed match results. A representation of deck contents is based on a second network, which performs the role of a dual-task encoder. Two tasks learned by the encoding networks are encoding decks in such a way that the full deck can be reconstructed, and encoding individual cards so that their specific properties can be decoded. Shared representation for these tasks allows the knowledge of individual cards to be taken into account.**

I. Introduction

COMPETETIVE multiplayer video games have been a thriving market in recent years, posing new challenges in areas of artificial intelligence and data analysis. In online game communities, the search for optimal tactics has lead to the development of "metagaming" - an entire layer of strategy related to the knowledge of certain playing styles and changes in their popularity on a community level.

The collectible card game (CCG) landscape, which includes popular and highly profitable games such as Hearthstone and Magic: The Gathering, naturally leads to the development of a particular type of metagame. In these games, players compete using decks which consist of a limited number of cards selected from a much larger pool of cards possible to collect. In a realistic scenario, the flow of knowledge between players leads to rapid development of popular deck types, with players often directly copying known well-performing decks. Any practical approach to applying artificial intelligence methods in such an environment has to consider not only the typical issues of moment-to-moment gameplay but also the metagame information.

In this paper, we explore the problem of predicting win rates for Hearthstone decks within a particular metagame environment. The proposed approach was developed as a submission to the AAIA'18 data mining challenge organised by Silver Bullet Labs and Knowledge Pit as part of the FedCSIS 2018 conference. The paper is arranged as follows: Section II describes the competition and available challenge data, Section III explains the use of external data not provided by organisers. Neural networks which serve as core components of the proposed approach are described in Sections IV and V. Section VI describes how ensembling was used to improve the results and Section VII summarises the conclusions.

II. Competition Description

The competition posed the task of predicting deck win rates for a set of 200 decks based on the record of 300000 games between another set 400 decks. The decks were played by four distinct AI agents, with the AI choice influencing win rates significantly. The goal of the prediction model was to compute win rates for all possible AI-deck pairs in the test dataset. This equates to 800 test samples. The training dataset included:

- name and the number of copies for each card present in the deck
- basic description of games including AIs playing, decks being played and the winning player
- detailed description of games - a recorded data of all turns, including actions taken by respective players

The proposed solution uses the basic descriptions of games while utilising an external dataset to represent cards present in the test, but not training set.

During the contest, it was possible to upload solutions and receive an evaluation of RMSE on an evaluation subset of 10% test samples (i.e. 80 AI-deck pairs randomly sampled from all possible 800). This influenced the chosen approach, as "over-tuning" parameters to increase fitness on the evaluation subset of the test set was possible. In fact, the order of top 4 results on the competition leaderboard was reversed in the final results, suggesting multiple submissions including the one described here were over-tuned to some extent. The solution described in this paper was in 2nd place on the competition leaderboard when the submissions closed but placed 3rd in the final evaluation. Possible causes are discussed in the Conclusions section.

III. External Data Use

The proposed solution employs a set of data from the hearthstonejson database [1]. This database contains information on all cards present in the training and test decks. All numerical properties such as life and attack of minions, weapon durability etc. are accounted for. Keywords such as Battlecry, Taunt, Adapt etc. are recognized as binary variables (whether the card has a keyword or does not). Full card text

(a) Minion (b) Weapon (c) Spell (d) Quest (e) Hero

Fig. 1: Types of cards available in Hearthstone CCG [2]

for every hearthstone card present in the challenge datasets is also available. The dataset also contains certain conditions that need to be met to play the cards (such as "there needs to be a valid minion target").

IV. Encoding Network

To build an encoding of hearthstone decks, first, an encoding of a card is created. The goal of this encoding is to represent cards present in the test, but not training data. There are 330 unique cards in all training decks, and 18 unique cards in test decks that do not appear in training set. These cards include both spells and minions.

Types of cards available in Hearthstone CCG are shown in Fig. 1. Note that minions (a) and weapons (b) have informative numerical properties of attack, health, and durability. However, even when these properties are accounted for, card text can still have a significant effect on the gameplay. In the case of the presented minion, Deathwing, its battlecry ability drastically alters the game state by destroying all minions on the game board. In case of spells (c), the only available information is the card text. Finally, quest (d) and hero (e) cards can alter the overarching game strategy of the entire deck by replacing the player's hero or offering a powerful reward for fulfilling the quest condition. For these cards, even card text does not offer a sufficient explanation. However, no quests or heroes that are not in the training data appear in the test dataset.

Taking this knowledge into account, we build the representation of a card as a concatenation of two vectors. First contains numeric properties, mechanics and other data available from hearthstonejson.com. Each numeric property is encoded as a continuous variable, each keyword is encoded as a binary variable, and all conditions required to play a card are encoded as binary variables. We use all properties available in hearthstonejson descriptions, as long they actually occur in training and test dataset.

The second vector is a word occurrence vector based on the card text. Card text is cleaned by removing punctuation, after which we build a dictionary of all strings that occur in the dataset (separated by whitespace characters) and count their occurrences in each card's text. Word occurrence serves as a simplified way to contain information regarding card function. Terms such as "destroy" or "heal" describe the interactions of a card to some extent, and can be relevant to the AI's ability to efficiently use the card. Without actually simulating the game logic, this is an easy way to represent effects such as the Deathwing battlecry mentioned above.

The representation of decks is then built by a neural network trained on all 600 training and test decks. The encoding network's loss function is defined as a sum of two terms, representing two distinct tasks. First is a standard autoencoder [3], i.e., the loss is based on the network's ability to reconstruct exact input vectors from a lower-dimensional encoding in the hidden layer. Inputs used in optimising this objective are decks from both training and test sets, represented as simple card occurrence vectors - each dimension in the input space represents the count of a particular unique card in the deck.

The second task is to learn an encoding which makes it possible to decode the properties of each card. For this purpose, we use the matrix of card properties C, in which each row represents hearthstonejson information of a single card. The assumption here is that encoding a single card in the same space as full decks can be decoded as card's specific properties. Given shared encoding Enc, autoencoder decoding Dec_1 and card property decoding Dec_2, the combined loss for both tasks can be calculated as:

$$\|Dec_1(Enc(X)) - X\|_F^2 + \|Dec_2(Enc(I)) - C\|_F^2 \quad (1)$$

Where X is the matrix of deck vectors, C is the matrix of card property vectors and I is an identity matrix of a size corresponding to the number of cards. Combining both tasks ensures the network encodes decks in a way that retains full information, but also encodes similar cards in a similar way.

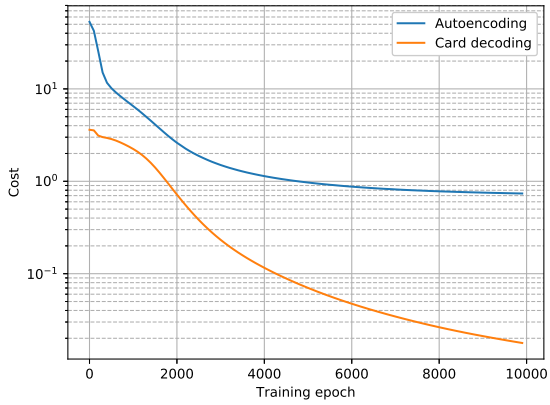


Fig. 2: Training curves of two objectives in the encoding network

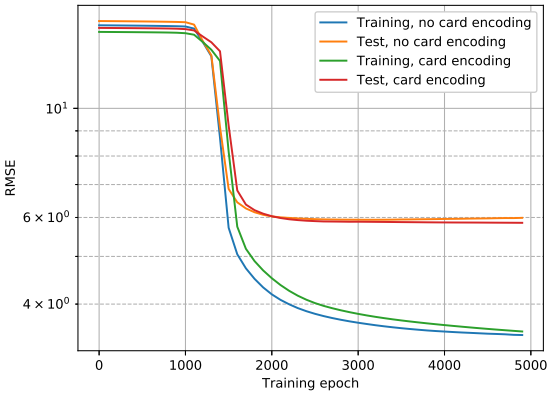


Fig. 3: Training curves of the prediction network, with and without using the card decoding objective in the encoding network

The latter is relevant for the cards that do not appear in training data.

Each of the functions: Enc, Dec_1, Dec_2, is learned by a single neural network layer, resulting in a network with one hidden layer and two separate output layers. The encoding layer uses ReLU nonlinearity [4], and decoders are linear layers. The dimensionality of encoding was set to 200 after preliminary tests. The number of training epochs for encoding network was set to 10000, and the network was trained with a gradient-based method Adadelta [5] with $\rho = 0.9$. Using a Theano [6] implementation, training this network takes approximately 3.5 seconds per 1000 epochs on a Nvidia GTX970 GPU.

Training curves of the encoding network are presented in Fig.1. While autoencoding objective reaches a visible plateau, card decoding objective could still be trained beyond 10000 epochs. However, we found this did not improve the prediction network results.

V. Prediction Network

Features encoded by the encoding network are used as and input to the prediction network. Prediction network is a standard feedforward neural network [7] with three hidden layers, respectively 300, 200 and 100 neurons. ReLU nonlinearity is used for activation in hidden layers. The network is optimized with Adadelta, using $\rho = 0.9$. To avoid errors (loss function can return NaN values if the output is outside of $(0, 1)$ interval), final layer activation was implemented as:

$$\sigma(x) = 1 - ReLU(1 - ReLU(x)) \tag{2}$$

However, in practice, outputs do not exceed 1.0 or 0.0 during optimisation if the hyperparameters are tuned for the task, i.e., the bias of final layer is initialised to 0.5 and other parameters to very small values. Therefore, the final layer effectively works as linear.

A basic loss function over all outputs, where o_i denotes the prediction for i-th deck that approximates a known win rate y_i, is defined as:

$$\sum_i -d_i(y_i \log(o_i) + (1 - y_i) \log(1 - o_i)) \tag{3}$$

where d_i represents the number of observations based on which the y_i win rate was calculated. In other words, the solution maximises log probability of the observed sequence of games, assuming all games by any particular deck can be modelled as a win-loss boolean random variable. Note that this completely ignores the two-sided nature of games and the essential property of decks having varying win rate against specific enemies. However, attempts to estimate matchup-specific win rates resulted in worse performance, possibly caused by to overfitting due to an insufficient number of games for each specific matchup.

Additionally, decks from the test set were employed in training of the network to counteract the positive win rate bias which appeared when using training data. When the number of games per deck is not distributed uniformly, it is possible for the average win rate to be over or under 0.5, thus creating an unwanted bias in the model. In the training set for the competition, the average deck has a win rate of 0.517, leading models trained on the data to overestimate win rates of test decks. Using prediction for test games, this bias can be removed. Assuming u_j is the output for j-th test set game, the full loss is defined as:

$$\sum_i -d_i(y_i \log(o_i) + (1-y_i) \log(1-o_i)) + (\lambda - \sum_j \frac{u_j}{200})^2 \tag{4}$$

The second term leads the average predicted win rate over the test set to be close to λ. This also works as regularisation for training. The λ value was chosen experimentally to maximize performance on leaderboard evaluation. We tested a range of values from 0.48 to 0.51, with 0.005 step size, and set $\lambda = 0.49$.

In Fig. 3, the training curve for the prediction network is shown. We compare the prediction network's performance

TABLE I: Results of the competition - top 5 submissions

Team	RMSE
hieuvq	5.57339852
amy	5.6482014
jj	**5.66759451**
dymitrruta	5.696228
amorgun	5.8473786

using two different encodings of training data. First is using the dual-task encoder described in Section IV, while the second one is encoded by a standard autoencoder, with no card decoding objective (i.e. the loss function is equal to the first term in Eq. 1). Training-test split for these experiments was the same, using 300 decks for training and 100 for tests. It is noticeable that the prediction network starts on a plateau and requires more than 1000 epochs to escape it, then rapidly improves the results. Past 2000 epochs little improvement is seen in test results although the minimum on the training set is not yet reached. Because of this, we set the early stopping point at 5000 epochs. Training time for this network was approximately 8 seconds per 1000 epochs on a GTX970 GPU.

Moreover, the improvement from using card decoding objective can also be seen in Fig.3. Prediction network performs better on training set but worse on test set when the card encoding objective is ignored. This indicates worse generalization without using the card decoding objective.

VI. Ensembling

The best performing single network achieved RMSE of approximately 5.0 on the 10% of the test data used to calculate leaderboard results. This result was further improved by ensembling, averaging results over multiple deep network models. Since throughout the competition we uploaded multiple results, models for the final ensemble were chosen from these according to their leaderboard evaluation results. The parameters given in sections IV and V describe the best single-network model. Other models in the final ensemble were variants of the described one with minor alterations, previously tested during parameter tuning: one with a larger number of training epochs (100000 for the encoder, 10000 for prediction network), one with added l2 regularization term in loss function (0.01 weight), and one with card matrix ignoring properties of cards other than word occurrence. These four best single-network models were averaged to obtain the final submission, resulting in approximately 0.2 RMSE improvement on the evaluation subset of test data.

VII. Conclusions

The final result placed the proposed solution as third in the competition. Results for other top submissions can be seen in Table I. It is worth noting that during the competition, two top solutions on the evalutation leaderboard reached RMSE below 5.0.

It can be argued that the chosen approach to representing Hearthstone decks was not sufficient to represent all intricacies of cards present in the test, but not training data. However, the change between results on the evaluation subset and final leaderboard suggests another explanation of the results, namely, that the described approach (along with some other top submissions) was over-tuned for the evaluation subset of test data.

While identifying the exact cause of this over-tuning is not possible without extensive tests on full data, the most likely explanation lies in the chosen approach to reducing positive win rate bias. As mentioned in Section V, the bias reduction is achieved by explicitly forcing the average win rate over test data to be close to an experimentally chosen value. The value 0.49 was set to maximise the performance according to the leaderboard. This means an implicit assumption was made that the 10% evaluation subset provides an accurate estimate of mean win rate for the entire test set.

Additionally, the choice of models to build an ensemble was based on the evaluation leaderboard, further contributing to the exact fitting of the model to evaluation subset of test data. A more refined ensemble building strategy would likely improve the results.

Regarding future work, the possibilities of the proposed approach are somewhat limited. The chosen model predicts win rates given a specific opponent distribution while using an unrealistic assumption that the test decks themselves are not part of the metagame. In a practical setting, due to fluctuating opponent distribution, a model that computes win rate for a matchup between two specific decks should be more applicable. We attempted to build such a model during the competition, however, since its training was significantly more computationally expensive than the prediction of win rates, the described approach was chosen instead. Nevertheless, predicting matchup-specific win rates remains a possible goal for further development that would require designing a new prediction network. The encoding network, on the other hand, can be potentially re-used without changes for any task that requires deck representation.

Acknowledgements

We would like to thank Silver Bullet Labs and Knowledge Pit for providing the simulation data and a platform for the competition.

References

[1] https://hearthstonejson.com/docs/cards.html
[2] https://hearthpwn.com
[3] Coates, Adam, Andrew Ng, and Honglak Lee. "An analysis of single-layer networks in unsupervised feature learning." Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011
[4] R Hahnloser, R. Sarpeshkar, M A Mahowald, R. J. Douglas, H.S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit" Nature, 405, pp. 947-951, 2000
[5] Zeiler, Matthew D. "ADADELTA: an Adaptive Learning Rate Method." Computing Research Repository, 2012
[6] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions", arXiv:1605.02688 (2016).
[7] Deng, Li. "A tutorial survey of architectures, algorithms, and applications for deep learning." APSIPA Transactions on Signal and Information Processing 3, 2014

Preprints of Position Papers of the Federated Conference on Computer Science and Information Systems pp. 189–192
DOI: 10.15439/2018F386
ISSN 2300-5963 ACSIS, Vol. 15

Toward an Intelligent HS Deck Advisor: Lessons Learned from AAIA'18 Data Mining Competition

Andrzej Janusz*†, Tomasz Tajmajer*‡, Maciej Świechowski‡,
Łukasz Grad†, Jacek Puczniewski†‡ and Dominik Ślęzak*
*Institute of Informatics, University of Warsaw, Poland
†eSensei, Poland
‡Silver Bullet Labs, Poland
Contact Email: janusza@mimuw.edu.pl

***Abstract*—We summarize AAIA'18 Data Mining Competition organized at the Knowledge Pit platform. We explain the competition's scope and outline its results. We also review several approaches to the problem of representing Hearthstone decks in a vector space. We divide such approaches into categories based on a type of the data about individual cards that they use. Finally, we outline experiments aiming to evaluate usefulness of various deck representations for the task of win-rates prediction.**

***Keywords*—Data Mining Contest; Data Representation Learning; Hearthstone: Heroes of Warcraft; Win-rates Prediction**

I. Introduction

AAIA'18 Data Mining Challenge: *Predicting Win-rates of Hearthstone Decks* was the fifth contest organized in association with the FedCSIS conference series. The topic was a follow-up of the previous edition of the challenge, related to a popular collectible card game *Hearthstone: Heroes of Warcraft* [1]. This time, participants were asked to predict win-rates of various Hearthstone (further abbreviated as HS) decks played by AI bots, based on games played with similar decks.

HS is a good framework for carrying out AI research. One kind of research is development of autonomous game-playing agents. The game is popular (more than 70M active players), highly competitive (one of the biggest eSport games) and yet it has combinatorial-game-like structure. Some notable bots reported in the literature are MetaStone (https://github.com/demilich1/metastone), Hearthranger (http://www.hearthranger.com), HearthBot [2] and Silverfish [3]. The second type of research revolves around analysis of the game in order to, e.g., help the players build better decks [4], [5]. A common need in all such investigations refers to an appropriate data representation that can be considered with respect to cards, decks or players. As one could see in the entries in the '17 installment of the competition, a good card representation is the backbone of ML-based playing agents. This aspect is even more crucial when it comes to win-rates prediction.

The competition outlined in this paper refers to both bot and non-bot research. On the one hand, we employ our AI algorithms [6] to generate massive bot vs. bot game logs data set. Our bots play using different decks and simulate different levels of real players. On the other hand, the competition task is related to another thread of our investigations, i.e., designing an advisory platform that helps players compose better decks [7]. Indeed, the top competition solutions, especially those taking into account the importance of the aforementioned card representations, can lead us toward new insights with respect to what decides about the win-rates of particular decks.

The rest of the paper is organized as follows: In Section II, we summarize the competition. In Section III, we discuss several approaches to constructing hybrid vector representations of HS decks and compare empirically usefulness of the obtained representations for predicting the win-rates. In Section IV, we draw some directions for future research.

II. AAIA'18 Data Mining Challenge

The competition (https://knowledgepit.fedcsis.org/contest/view.php?id=123) took place on April 3 – May 7, 2018, under the auspices of 13^{th} International Symposium on Advances in Artificial Intelligence and Applications (https://fedcsis.org/2018/aaia) which is a part of the FedCSIS conference series. The purpose of this challenge was to discover reliable methods for predicting win-rates of HS decks. The task was to construct a prediction model that can learn win chances of new decks, based on the history of match-ups between AI bots playing with similar decks. To give participants freedom of choosing a representation of the data, apart from a preprocessed data set in a tabular format, there were provided JSON files with detailed descriptions of each game. We were interested whether the data regarding the way in which cards are played during the game can be useful in the proposed task.

The training data set contained logs from 299680 games played between four bots which used 400 decks. Another 200 decks – combined with the same bots as in the training set – were used as a test set. The win-rates of the bot-deck pairs from the test set were computed based on 300000 simulated play-outs. In those games, one of the bots used a deck from the training set, and the other one – from the test set. The decks were created by randomly mutating a set of 13 deck archetypes that at the time of the competition were commonly used in ladder matches by human players (12 top-rated archetypes and one group of decks consisting of only basic cards).

To generate the games, we defined four HS bots that differed in: a) available time limit for performing a move and b) available knowledge about the opponent hand (*full_info* or *limited_info*). Eventually, the following configurations were used: A1 – limited_info & 1 second per move, A2 – limited_info & 2 seconds per move, B1 – full_info & 1 second per move, B2 – full_info & 2 seconds per move.

TABLE I: Final RMSE results and number of submissions from top-ranked teams. The last row shows the result obtained by a baseline solution – a fully-connected neural network with two hidden layers, trained on the bag-of-cards representations of decks.

team name	rank	number of submissions	final result
hieuvq	1	195	5.5734
amy	2	149	5.6482
jj	3	225	5.6676
dymitrruta	4	258	5.6962
amorgun	5	6	5.8474
...	...	...	...
baseline	26	–	8.8645

A. Evaluation of results and participation in the challenge

Submissions from participants were managed by Knowledge Pit [8]. Each submission had to be properly formatted, containing predictions of win-rates for every deck-bot pair from the test set. Each of the teams could submit multiple solutions. As a quality criterion for submissions, we selected the RMSE measure. The submitted solutions were evaluated on-line and the preliminary results were published on the competition leaderboard. The preliminary scores were computed on a subset of the test set, fixed for all participants. The size of this subset corresponded to randomly chosen 10% of the test decks. The final evaluation was conducted after competition's completion using the remaining part of the data.

Apart from submitting their predictions, each team was obligated by competition rules to provide a brief report describing the approach used. The description had to cover utilized learning models, as well as the steps of data preprocessing and feature extraction. Only teams which sent a valid report qualified for the final evaluation. In this way, we were able to collect a vast amount of information regarding various representations of HS decks and the state-of-the-art approaches to this type of prediction problems. After completion of the challenge, the final results were published on-line. The scores obtained by top-ranked teams are presented in Table I.

B. Summary of the competition results

Our contest attracted 204 teams from 28 countries. The countries with the highest number of registrations in the challenge were Poland (119), Russia (28), United Kingdom (9), United States (8) and India (5). Among the participating teams, 82 submitted at least one solution file which was ranked at the public leaderboard. Over a half of those teams decided to disclose their approach by uploading short reports.

The top solutions were obtained by ensembles of regression models. The winners combined linear regression with deep neural networks. The second team blended a tree-based boosting model (XGBoost) with Extreme Learning Machines and LASSO regression. Other models that performed well were SVR (SVM ϵ-regression) and Gaussian processes.

To represent the decks as vectors, many of the top-ranked teams encoded them as bags-of-cards, i.e., vectors of a size equal to the number of distinct cards in the data, where each element indicates how many cards of a corresponding type are present in a deck. The winners augmented such a representation using aggregated card properties, e.g., the total health of minions in a deck, the number of spells, the number of minions with a taunt ability, etc. A few teams incorporated into their representations advanced knowledge about HS, e.g., indicators of cards' relative strength defined by experts. Still, none of the top 20 teams used information from game logs to augment their representations of decks.

III. Representations of Hearthstone Decks

Since one of our objectives for this competition was to find out whether information regarding the way cards are played during HS duel can be useful for predicting win-rates of decks, we further investigated this problem in a series of experiments. We created various card representations using three different data sources. Based on those representations, we built vector embeddings of decks and compared their usefulness by measuring performance of several prediction models. Let us discuss the obtained results.

A. Bag-of-cards and its transformations

The most common representation of HS decks is a bag-of-cards – by an analogy to a bag-of-words representation of textual documents. A deck is regarded as a set of card IDs. Its vector representation has a length equal to the number of available cards. The i-th vector's position expresses the amount of copies of the i-th card in the given deck.

Such a simple representation turned out to be very effective for predicting win-rates. It was utilized by many of competition entrants, including all the top three teams. However, in nearly all cases, it was augmented by additional information extracted from the cards, e.g., a distribution of card mana costs. The augmentation was usually done by aggregating properties of cards included in the deck. For this purpose, participants often used external knowledge bases, such as the one provided by HearthstoneJSON API (https://hearthstonejson.com/).

The dimensionality of bag-of-cards representation can be reduced using some text mining techniques, such as SVD. A deck representation in the space of latent concepts can be used by itself. It can be also combined with the others to express combinations of cards often appearing in the same deck.

B. Aggregation of cards represented in a vector space

Representation of a deck can also be created by aggregating representations of individual cards. Information about the cards can be acquired from various sources, such as:

- a database with card properties and textual descriptions (e.g.: HearthstoneJSON, Wiki)
- a database of players' decks (e.g.: HearthPWN.com)
- logs from games between human players or AI bots (e.g.: the data used in our competition)

Specific algorithms for creating vector embeddings of HS decks based on the data from the first two of the above sources are described in [7]. Game logs can be utilized to generate embeddings of cards, e.g., using a word2vec model [9] in which card IDs correspond to terms and their use sequences extracted from game logs are treated as documents.

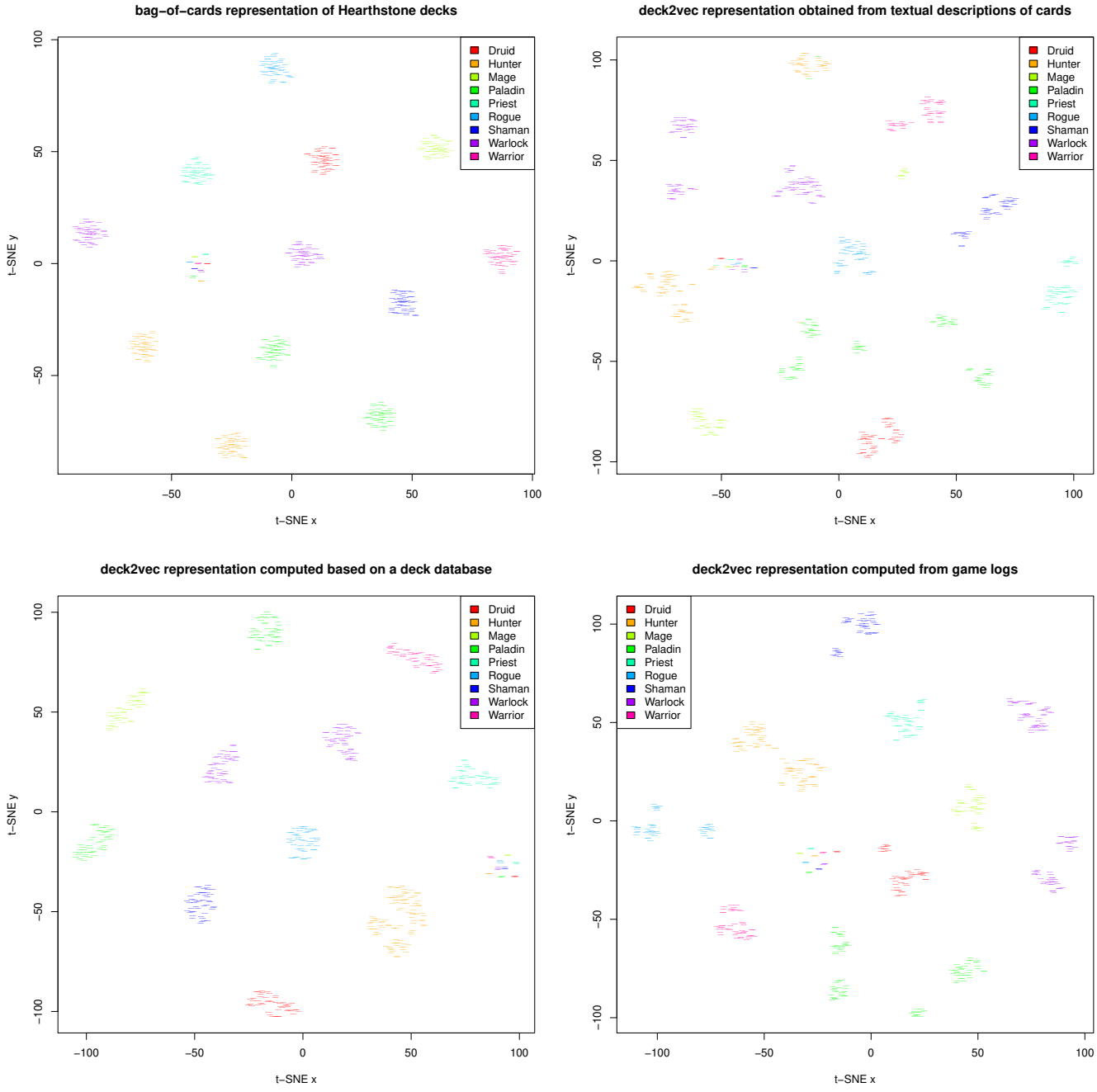


Fig. 1: A t-SNE-based visualization of four deck representations. Top-left: bag-of-cards. Top-right: aggregated embeddings derived from textual descriptions. Bottom-left: representation based on an on-line deck database. Bottom-right: representation based on the competition game logs.

The derived representations may cover different aspects of card similarity. For instance, the word2vec embeddings of cards computed from their descriptions can capture information about their basic properties. A representation derived from compositions of decks created by players may be better at expressing the aforementioned interchangeability [7]. Finally, a representation created from game logs may convey more information about the way cards are used during a game.

The aggregation of cards can be performed in many ways too. The simplest approach is to take a mean of the card vectors. To prevent losing too much information, such a deck representation can be extended by, e.g., min, max and standard deviation of card vectors (computed dimension-wise).

To visualize differences between the above approaches, we used them to represent the decks from our competition. Figure 1 shows these representations embedded into a two-

dimensional space using the t-SNE algorithm [10]. Only the bag-of-cards method allows to identify all deck archetypes which we used to generate the data (12 visible groups of decks for nine hero classes and one mixed group with decks containing only basic cards). For the representation obtained by analyzing a deck database, two groups of decks for the Hunter hero class were merged together. This is quite a good indicator given the fact that these archetypes were *FaceHunter* and *AggroBeast*, which share significant portion of cards and have a similar game plan. For the representation derived from textual descriptions, some archetypes were split into separate groups. Such a division is also visible when using representations extracted from game logs. Therein, however, apart from the group of decks composed of basic cards, there are no clusters with mixed decks from different hero classes.

C. Predictive power – experimental evaluation

We performed a series of experiments to evaluate the impact of the deck representation methods on a predictive performance of various regression models. We trained four models, namely the aforementioned Gaussian Process Regression (GPR) [11], SVR [12], Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN) on the competition data and compared their results obtained for four representations of decks. Table II shows the final scores.

The best score was achieved by combination of the bag-of-cards and GPR. On the other hand, the same representation combined with KNN regression was the worst. This fact highlights a need of adjusting data representation for a given prediction model. The second best representation was the one derived from a deck data set. Its score was even slightly better than for the bag-of-cards used together with MLP and KNN. The most disappointing were results of the representation based on game logs. It suggests that more advanced methods of learning deck representations from logs need to be developed to fully utilize this source of information.

Nevertheless, motivated by a diversity of the deck clustering results, we decided to check if the learned representations can contribute some additional knowledge to prediction models. We concatenated each of the vector representations with the bag-of-cards and evaluated their performance in a combination with GPR. These results are included in Table II. Surprisingly, for every representation we obtained considerably better results than for the plain bag-of-cards. Moreover, when we concatenated all four representations, we achieved the best score (RMSE 5.465) among all entries in our challenge (see Table I). This confirms the benefit of using diverse sources of the data for constructing prediction models.

TABLE II: RMSE scores obtained for four prediction models and discussed deck representation methods, where BC stands for bag-of-cards and DD, TD and GL denote representations derived from a deck data set, textual card descriptions and game logs, respectively. The last row shows results for selected concatenated vector representations.

representation:	BC	DD	TD	GL
GPR model	5.812	7.059	7.768	9.274
SVR model	6.185	7.107	7.435	7.873
MLP model	7.035	6.823	8.573	8.018
KNN model	10.158	9.152	9.348	8.960
representation:	BC+DD+TD+GL	BC+DD	BC+TD	BC+GL
GPR model	**5.465**	5.503	5.733	5.629

IV. Conclusions

We summarized AAIA'18 Data Mining Competition organized at the Knowledge Pit platform, whose topic was predicting win-rates of Hearthstone decks. The outcomes of the competition show that various machine learning models are capable to accurately assess the quality of new decks, based on the data regarding performance of similar decks.

Our own comparison of deck representations created using various sources of information about cards revealed that the simplest approach, i.e., the bag-of-cards, can be successfully employed. Moreover, our experiments show that it is possible to considerably improve performance of prediction models by training them on combined representations from different sources. We believe that such a hybrid approach will move us one step further in our ultimate goal of designing an advisory platform for helping players in composing their decks.

Acknowledgments

This research was co-funded by Smart Growth Operational Programme 2014-2020, financed by European Regional Development Fund under GameINN projects POIR.01.02.00-00-0150/16 and POIR.01.02.00-00-0184/17, operated by National Centre for Research and Development in Poland.

References

[1] A. Janusz, T. Tajmajer, and M. Świechowski, "Helping AI to Play Hearthstone: AAIA'17 Data Mining Challenge," in *Proceedings of FedCSIS 2017*, 2017, pp. 121–125.

[2] A. R. da Silva and L. F. W. Goes, "HearthBot: An Autonomous Agent based on Fuzzy ART Adaptive Neural Networks for the Digital Collectible Card Game Hearthstone," *IEEE Transactions on Computational Intelligence and AI in Games*, pp. 170–181, 2017.

[3] S. Zhang and M. Buro, "Improving Hearthstone AI by Learning High-level Rollout Policies and Bucketing Chance Node Events," in *Proceedings of IEEE CIG 2017*, 2017, pp. 309–316.

[4] P. García-Sánchez, A. Tonda, G. Squillero, A. Mora, and J. J. Merelo, "Evolutionary Deckbuilding in Hearthstone," in *Proceedings of IEEE CIG 2016*, 2016, pp. 1–8.

[5] A. Stiegler, C. Messerschmidt, J. Maucher, and K. Dahal, "Hearthstone Deck-construction with a Utility System," in *Proceedings of SKIMA 2016*, 2016, pp. 21–28.

[6] M. Świechowski, T. Tajmajer, and A. Janusz, "Improving Hearthstone AI by Combining MCTS and Supervised Learning Algorithms," in *Proceedings of IEEE CIG 2018*, 2018, In print.

[7] A. Janusz and D. Ślęzak, "Investigating Similarity between Hearthstone Cards: Text Embeddings and Interchangeability Approaches," in *Proceedings of IEEE SMC 2018*, 2018, In print.

[8] A. Janusz, D. Ślęzak, S. Stawicki, and M. Rosiak, "Knowledge Pit – A Data Challenge Platform," in *Proceedings of CS&P 2015*, 2015, pp. 191–195.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Proceedings of NIPS 2013*, 2013, pp. 3111–3119.

[10] G. Hinton and S. Roweis, "Stochastic Neighbor Embedding," *Advances in Neural Information Processing Systems*, vol. 15, pp. 833–840, 2003.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[12] A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 193–196
DOI: 10.15439/2018F362
ISSN 2300-5963 ACSIS, Vol. 15

Predicting winrate of Hearthstone decks using their archetypes

Jan Betley zweryfikujfirme.pl
Email: jan.betley@zweryfikujfirme.pl
Anna Sztyber Warsaw University of Technology
Email: sztyber.anna@mchtr.pw.edu.pl
Adam Witkowski University of Warsaw
Email: adam.witkowski@mimuw.edu.pl

***Abstract*—This paper describes our solution for the AAIA'18 Data Mining Challenge: Predicting Win-rates of Hearthstone Decks. Train and test decks were clustered by DBSCAN algorithm with precomputed distance matrix dependent on the number of common cards. We observed that each cluster can be represented by an archetype deck - one of popular decks used by human players. For each deck we created features describing cards quality and types. Additionally we used differences of these features with respect to archetype decks. Finally we used XGBoost to build a model predicting outcome of a game played between two decks.**

I. Introduction

Hearthstone is the most popular online collectible card video game[1]. Despite simple rules, game requires high strategic skills, of two separate types:

- Ability to create a high quality deck (set of cards)
- Ability to use given deck as well as possible

Article describes our solution to AAIA'18 Data Mining Challenge, where teams were asked to predict winrates of different decks, played by different bots. Full data about deck composition was easily available, while bots strategies could be only guessed from training games history. Because of that, of the two factors (deck variability and player variability), corresponding to the two skills mentioned, we decided to focus on the first, using bot type only as a control variable dividing training/testing data into 16 subsets (4 bots playing against each other).

In Hearthstone, every deck has a strategy - a way to beat the opponent. The three main strategy types are aggro, control and combo. An aggro deck tries to kill the opponent as fast as possible. A control deck tries to destroy minions played by the opponent and finish the game with strong, high-cost minons. A combo deck uses a special combination of cards that work very well together to gain great advantage or even kill the opponent in one turn.

[1]While describing Hearthstone mechanics is clearly out of the scope of this work, some basic rules are enough to make it understandable. There are two players, each of them starts with deck of 30 cards and a 'hero' card. They play alternating rounds, in each round player may play any number of cards, limited by their costs and the resource called 'mana'. Cards have different types, most important being spells and minions. Played cards more or less directly contribute to dealing damage to the other player, and the only goal is to be the first player to deal 30 damage.

Although there is virtually infinite set of possible hearthstone decks, only few of them are good enough to be played on at least semi-professional level. One of the most popular webpages with hearthstone statistics, https://hsreplay.net/decks/, currently defines 48 deck "archetypes" (exact number varies in time), such as "Aggro Hunter" or "Cube Warlock". Archetypes are based on existence of certain key cards, cleverly matched to other key cards. Those connections build deck strength, and with some of those cards missing deck would become unplayable.

Archetypes usually define the only one correct strategy. Without going too much into detail, strategy is about maximizing played card value by playing them in the right moment. E.g. strategy for Aggro Hunter is to deal as much damage as possible as fast as possible, while Cube Warlock defends until he has enough mana to play very strong cards cheaply. Each strategy has a counter strategy that might be more or less available to the opponent, so most decks - even top quality ones - do much better against some certain decks, and much worse against other.

II. Problem statement

The goal of the competition was to predict winrates (percentage of games won) of 200 test decks played by bots[2]. There were 4 different bots (denoted by $A1, A2, B1, B2$). A priori nothing was known about the bots, in particular it was not known what algorithms were used by the bots (a bot could just use a set of simple heuristics to play the game, or it can be a deep neural network, like AlphaGo) The training set consisted of results of 299680 games played between 400 training decks. For each game, we had a tuple (bot1, deck1, bot2, deck2) and the result of the game. This tuple denoted that bot1 played deck1 against bot2 using deck2.

The winrates that we had to predict were calculated based on a large number of games played between test decks and the training decks. The score was calculated as RMSE between the actual winrate and the predicted winrate for each bot, deck pair (so there were 800 numbers to predict).

[2]bot is a computer program that plays a game, here Hearthstone

III. Bot winrates

While we knew nothing about specific bot strategies[3], they were certainly different. Overall winrates for each bot are presented in Table I.

TABLE I
Bot winrates - overall, as first/second player, vs others

bot	overall	1st	2nd	vs A1	vs A2	vs B1	vs B2
A1	0.45	0.51	0.39	0.50	0.44	0.38	0.40
A2	0.52	0.57	0.47	0.56	0.50	0.45	0.43
B1	0.54	0.58	0.50	0.62	0.55	0.50	0.46
B2	0.56	0.61	0.51	0.60	0.57	0.54	0.50

It is obvious that every reasonable predicting model must include information about bots playing, and - when predicting single game result - also about starting deck/bot. Having stated that, later in this article we won't be explicitly referring to bots and starting positions - they are present in every model, but our solution is based entirely on differences between decks.

IV. Deck archetypes via clustering

A natural question is: are the given decks just random collections of available cards, or are they created from some archetypes? We used clustering to answer this question. We defined a distance between decks as

$$d = \frac{30 - n_c}{30} \tag{1}$$

where n_c is the number of common cards (counting with repetitions) in both decks.

Then we used DBSCAN [1] algorithm from scikit-learn library [2] with this metric. The parameters of the algorithm were eps = 0.4 and min_samples = 5. The algorithm found 11 clusters:

- 2 different clusters for Paladin and Warlock heroes;
- 1 cluster for each other hero

There were also 25 decks that did not have any cluster assigned. Each cluster had 47 or 48 decks, except for the Hunter cluster which had 96 decks.

This result looked promising — if the decks were random, because of the big number of possible cards, there would be no meaningful clusters.

A. Deck archetypes

For each cluster we calculated the most frequent cards used in the decks from this cluster. In each cluster there were from 5 to 10 cards that appeared in almost every deck of the cluster. For example, in one of the Paladin clusters cards Vilefin Inquisitor, Murloc Tidecaller, Bluegill Warrior, Grimscale Chum, Murloc Warleader and Rockpool Hunter that are all murloc minions were among top 10 most frequent cards (see Table II). This fact combined with the knowledge that there exists a popular Murloc Paladin deck allows to easily classify decks from this cluster as being of this archetype.

Based on the most frequent cards and domain knowledge (one of the authors used to play a lot of Hearthstone) we determined the archetype of the decks in every cluster. The archetypes were: dead man's hand warrior, inner fire priest, jade shaman, aggro hunter, jade druid, zoo warlock, cube warlock, tempo rogue, secret mage, murloc paladin, and dude paladin.

[3]Competition data included full training games courses, so those strategies could be somehow extracted, but we did not use them

TABLE II
Top 10 most frequent cards in the Murloc Paladin cluster.

card name	frequency
Vilefin Inquisitor	97.92 %
Murloc Tidecaller	95.83 %
Righteous Protector	95.83 %
Bluegill Warrior	93.75 %
Corridor Creeper	93.75 %
Grimscale Chum	93.75 %
Call to Arms	91.67 %
Murloc Warleader	91.67 %
Rockpool Hunter	91.67 %
Unidentified Maul	91.67 %

B. Model decks

We had decks clustered and we knew their archetypes. We were interested in how much the given decks differ from decks of those archetypes played by professional human players. Ideally, we would prefer to compare provided decks with 'optimal' decklists but there is no such thing as a "perfect" deck — optimal decklist depends on the decks played by the opponents. For example, there is a card Golakka Crawler that is very effective against decks that use pirates. If the decks with pirates are popular, then the decks with the Golakka Crawler will be more successful than those without it. On the other hand, if no one uses pirates, then the card is useless. [4]

For each cluster (except of 'other') we chose one model deck from the website Tempostorm (https://tempostorm.com/hearthstone/meta-snapshot/standard/2018-01-08) that creates reports about popular/strong decks. One problem with this approach was that the used decklists change every week and we did not know the date from which we should take the decklists. We used the report from the beginning of January 2018, based on the following observations:

1) the decks contained cards from the Kobolds & Catacombs expansion, released in December 2017;
2) the decks did not have any cards from The Witchwood expansion, released in April 2018; and
3) high percentage of the decks contained cards Corridor Creeper (46%) and Patches the Pirate (28%) which were changed in February 2018 and lost a lot of popularity as an effect [5]

In Table IV (column average distance) we give the average distance of decks from each cluster to the model decks. The distance function is the same as the one used in the clustering.

Note the huge discrepancies in the average distances between clusters. This can be simply a matter of the prepared decks: maybe the Shaman decks were generated differently

[4]For human players it is therefore very important to know "the meta" — that is, which decks are strong and which are popular.

[5]changing a card to a weaker version is called a nerf in the Hearthstone terminology.

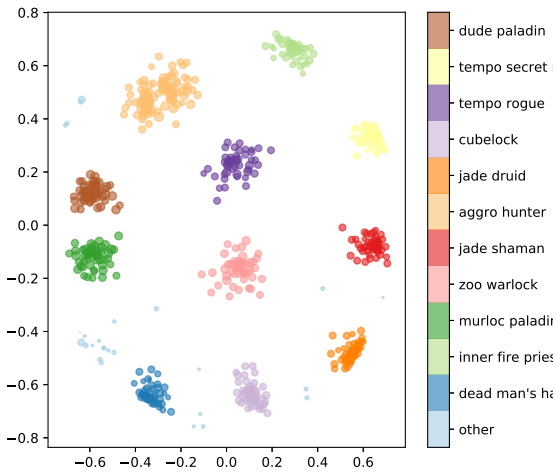


Fig. 1. 2-D decks embedding based on distance matrix using t-SNE ($perplexity = 2$, $angle = 0.9$)

than the Aggro Paladin ones. Of course this can be also a matter of poor choice of the model decks for some archetypes. We did not try to find other model decks.

To better visualize the clusters, we created a 2-D embedding using the t-SNE algorithm [4] with precomputed distance matrix. The clusters are visualised in Figure 1, Decks are coloured according to clusters found by DBSCAN. The size of dots is proportional to win-rates (with 50% substituted for test decks). In the figure we can see that clusters are clearly visible with except of the 'other' cluster. The only one unexpected fact is the division of the Hunter cluster into two groups. The number of decks in "aggro hunter" cluster is roughly two times greater than the number of decks in other clusters. Therefore it is possible that this cluster should be further divided into two clusters. However, neither DBSCAN nor multidimensional scaling (MDS) visualization confirm this division.

C. Simple cluster-based model

With established clusters, we tried the simplest possible model: for each deck D, predict the average winrate of the training decks from D's cluster as its winrate. For example, if in the training set there are 3 "murloc paladin" decks with winrates $65\%, 56\%$ and 55% respectively, each murloc paladin deck from test set gets winrate $\frac{(55+56+65)}{3} = 58.66\%$. Averaged winrates per clusters are shown in Table IV (column winrates). We can observe significant differences between clusters and poor performance of decks classified as "other". This model ignored any differences between decks in the same cluster, so it could not achieve a good score.

Another approach we tried was to use the metric given by Equation (1) directly, take the 10 training decks closest to the test deck and predict the average winrate of those 10 decks, but this gave much worse results. We decided to use more standard predictive models, improving them with features based on the clusters and model decks.

V. Predicting the winrates

A. Basic deck features

For each deck we generated a number of features that tried to capture the 'goodness' of the deck. Those included e.g.

- Average card cost
- Number of cards with cost of 0/1/2/3 or more
- Number of free/common/rare/epic/legendary cards[6]
- Number of neutral/single-hero cards[7]
- Number of minions/spells/weapons/other cards[8]
- Average overall card winrates[9]
- Number of special cards such as murlocs, beasts, minons with divine shield, taunt minions, etc.

We also gathered data from few webpages with Hearthstone statistics:

- "Card value" - overall card value when playing in arena mode[10] [http://www.heartharena.com/tierlist]
- "Card played winrate" - chance of winning the game, under condition that given card was played [https://hsreplay.net/cards]
- 2821 most popular decks [https://hsreplay.net/decks]

First and second datasets were averaged into deck "mean card value" and "mean card winrate" features. Most popular decks were used to estimate "how well cards in deck are connected". For each card pair we calculated:

- How often they appear in external decks
- How often they appear together in external decks
- "Card pair connection strength" as quotient of the above values

Deck feature "card connection strength" was calculated as a mean of connection strength between all card pairs in deck.

B. Differences with the model decks

In addition to deck-only based features, we created a set of features describing how different the deck is from the model deck of the same archetype, such as:

- General distance (as described in 1),
- How many 1-mana cards were added/removed
- Difference between added/removed card's arena value

This way we wanted to approximate how big is the "real" impact of the differences between decks and their archetypes. E.g. if deck uses many cheap minions, replacing some of them with other cheap minions is a small difference, while replacing them with expensive cards would be a drastic change.

[6]Every Hearthstone card fits into one of those categories, they approximately describe card strength

[7]Neutral card may be played by any hero, in the contrast to cards that may be played only by specific hero

[8]Every Hearthstone card has its type, "minions", "spells" and "weapons" are the most popular

[9]Taken from https://hsreplay.net/cards/

[10]Arena mode is a specific hearthstone variant, played with more or less random decks

TABLE III
FEATURES IMPORTANCES

feature	f-score
deck2 winrate	1198
mean card value p2	470
mean card value p1	442
mean minion health p2	436
mean card winrate p2	409
mean minion attack p2	387
who plays first	366
mean minion attack p1	363
mean minion health p1	352
mean card winrate p1	349
diff mean minion health p2	337
diff mean minion attack p2	334
card connection strength p1	328
diff arena value p2	327
diff arena value p1	320
card connection strength p2	320
mean minion cost p1	315
diff mean minion health p1	314
diff mean minion attack p1	308
mean minion cost p2	289

C. Final model

We considered two approaches for generating final predictions:

(a) train a regression model predicting winrate of each deck,
(b) train a classification model predicting result (win, lose) of a game between the deck and a particular opponent.

We tested both approaches and we decided on (b) due to larger training set available (300K training games versus 400 train decks) and more promising preliminary results. Since the test decks were evaluated based on the games against training decks, we added the feature "opponent's winrate" which was by far the most important one (see Table III).

Classification model was trained using XGBoost library [3], which is an implementation of gradient boosting algorithm. After training we predicted results of 4 million random games between train and test decks. Final winrates were averages of these games results.

Using XGBoost model we analysed features importances given by f-score, which is a measure of how often given variable was used to split node of a decision tree. Table III shows twenty most important features with respect to f-score. Each deck feature was repeated for both players, p1 and p2 denote player1 and player2 respectively. Features with diff prefix are differences between deck and its archetype.

VI. RESULT PER CLUSTER

After preparing the final model we tested how did the model work on particular clusters. Since we did not have the ground truth, the test was done taking 100 random training decks as the validation decks and training the model on the games played with the remaining 300 decks. We then calculated for each cluster the mean absolute error for the validation decks. The results are shown in Table IV (column mae).

We expected that we will not do so well on the 'other' cluster, since for those decks we did not have the model decks. One possible explanation is that the 'other' decks were quite weak (27% average winrate) and quite different from the other decks and therefore easier to predict.

TABLE IV
RESULTS FOR EACH CLUSTER SEPARATELY: AVERAGE WINRATE PER CLUSTER, AVERAGE DISTANCE FROM EACH CLUSTER TO THE MODEL DECK AND MEAN ABSOLUTE ERROR

cluster	winrate	average distance	mae
dude paladin	0.6482	5.33	3.90
murloc paladin	0.6141	11.98	3.13
zoo warlock	0.5871	6.46	5.08
aggro hunter	0.5860	8.13	5.54
cubelock	0.5074	5.83	4.70
tempo rogue	0.5013	6.62	5.79
jade shaman	0.4680	14.51	3.53
jade druid	0.4510	9.28	3.93
dead man's hand warrior	0.4174	12.04	5.24
tempo secret mage	0.4084	8.85	3.38
inner fire priest	0.3888	11.56	5.78
other	0.2651	-	3.01

VII. CONCLUSIONS

Our idea was to explore if the decks were generated randomly or followed the pattern of human players decks. Clustering revealed existence of groups. Each group can be represented by an archetype deck selected from decks of successful human players. Our final solution was generated by XGBoost model using features describing differences between each deck and its archetype. These distance features improved model results significantly with comparison to the model using only basic features. Our model achieved the RMSE score 6.349 which gave us 10th place in the competition.

The solution could be further improved by:

- building ensemble of different models,
- exploring other clustering algorithms,
- XGBoost hyper-parameter tuning.

Final result of our model evaluated on all test decks was slightly worse than on a fraction of test decks available for early evaluation of submissions, which can be caused by overfitting of the solution to the test data available. It would probably be beneficial to leave part of training decks for validation and comparison of different solutions.

REFERENCES

[1] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, pp. 226–231. [Online]. Available: http://dl.acm.org/citation.cfm?id=3001460.3001507

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016. doi: 10.1145/2939672.2939785. ISBN 978-1-4503-4232-2 pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[4] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 197–200
DOI: 10.15439/2018F363
ISSN 2300-5963 ACSIS, Vol. 15

Predicting Win-rates of Hearthstone Decks: Models and Features that Won AAIA'2018 Data Mining Challenge

Quang Hieu Vu
Data Science Group,
Zalora, Singapore
quanghieu.vu@zalora.com

Dymitr Ruta
EBTIC, Khalifa University
Abu Dhabi, UAE
dymitr.ruta@kustar.ac.ae

Andrzej Ruta
ING Bank Slaski,
Katowice, Poland
andrzej.ruta@ingbank.pl

Ling Cen
EBTIC, Khalifa University
Abu Dhabi, UAE
cen.ling@kustar.ac.ae

***Abstract*—Success of many computer games depends on designing a robust and adaptable AI opponent that would ensure the games continue to challenge, immerse and excite the players at any stage. The outcomes of card based games like "Heartstone: Heros of Warcraft", aside the player skills heavily depend on the initial composition of player card decks. To evaluate this impact we have developed an ensemble prediction model that tries to predict the average win-rates of the specific combination of bot-player and card decks. Our ensemble model consists of three sub-models: two Logistic Regression models and one Deep Learning model. The models are trained with both provided data and additional data about the cards, their health, attack power and cost. To avoid overfitting, we employ a trick to generate predictions for all possible combinations of opponent players and decks and obtain the result as the average of all these predictions.**

I. Introduction

Success of many computer games depends on designing a robust and adaptable AI opponent that would ensure the games continue to challenge, immerse and excite the players at any stage. In the history of game development, much efforts have been dedicated to the design and implementation of such a bot player. There are some very powerful examples of AI robots already, such as the famous "Deep Blue" that has defeated the world chess human champion. "Heartstone: Heros of Warcraft" is a turn-based card games between two players who select their heroes with a unique power and construct a deck of thirty cards that represent various spells, weapons, and minions, and can be summoned in order to attack the opponent with the goal of reducing the opponents health to zero and win. The outcomes of this game, aside the player skills, heavily depend on the initial composition of player card decks.

With the purpose of designing such robust and adaptable AI opponent for Heartstone, an important task is to correctly predict the marginal winning probability of AI players, given unseen decks. This very task has been defined as the target of the AAIA'18 data mining challenge. Specifically, the objective was to construct a prediction model that can learn win chances of the specific AI bots assigned to specific new card decks, based on the historic evidence of same AI bots playing with similar decks [1] against all kinds of players and decks extracted from hundreds of thousands of automated games. In this competition setup the training stage involved observing four AI players assigned to one of 400 available decks of 30 cards, battling each other in over 300000 automated games. The deviced prediction models were expected to use this data to learn how particular cards are played by the bots and evaluate their contribution or impact on the final win-rate estimate of specific decks-players. The competitive models are evaluated on the games with the configurations of the same 4 bot-players and 200 new card decks.

To solve the challenge we followed a pragmatic sequence of steps that could be in fact considered generic best practices for any competition involving predictive model build from data: understand the data, perform exploratory data analysis, conduct feature engineering and select features for the model, construct models, evaluate them and optimize the complete pipeline to maximally improve the predictive performance.

Since our solution turned out to be the best and won the 1^{st} place in AAIA'2018 data mining competition, we provide detailed information of how we follow these steps to design the top model. In addition, we also introduce a key technique that we have employed to avoid overfitting, which we believe was instrumental in winning the challenge. In general, our paper offers the following two major contributions:

- We provide a clear demonstration of how basic steps in designing a machine learning model should be executed: from data understanding and feature engineering to model design, parameter tunning and model's improvement.
- We present a critical method to avoid the overfitting in reconstructing regression based win-rate from large number of simple classification models and experimentally verify how it results in significant gains in testing set performance.

In rest of the paper is organized as follows. In Section II, we introduce related work. In Section III, we describe the problem, available data and features that we generated from the dataset to train our models. In Section IV, we present both our single models and how to combine them to form an ensemble model. We discuss how we avoid overfitting to improve model's accuracy in Section V. Finally, we conclude the paper in Section VII.

II. Related Work

In this section, we briefly introduce the machine learning techniques used in our model: Deep Learning, Logistic Re-

gression and Ensemble model.

A. Deep Learning

Deep Learning (DL) refers to a class of machine learning tech- niques and architectures, where many layers of non-linear information processing stages in hierarchical architectures are exploited for representation learning [2]. In particular, a DL network represents a multi-layer neural network with the deeper structures compared to the shallow models like Support Vector Machines and a specific method where the data is processed at and in between layers. Even though the concept of DL was introduced long time ago, it has only recently gained enormous popularity due the lower cost of computing hardware, the increased speed of chip processing, and recent advances in DL algorithms. DL has been successfully employed for computer vision, optimization, pattern recognition, signal processing, and natural language processing [3].

B. Logistic Regression

Logistic regression, developed by David Cox in 1958 [4], is a statistical method for regression analysis to describe the relationship between one dichotomous dependent variable (outcome) and one or more independent variables (predictors or features). Binary logistic model can be used for estimating the probability of a binary response based on predictors and gain insights on the presence of which factors increase the probability of a given outcome by a specific percentage.

Logistic regression has been widely used in medicine, e.g. to assess injury mortality or severity for patients [5], [6], [7], [8], [9], or help to diagnose some diseases like diabetes and coronary heart disease based on characteristics and physiological data of patients such like age, sex, body mass index, blood test results, etc. [10]. It has also been successfully applied in various areas, e.g. predicting votes of American voters based on their characteristics like age, income, sex, race, state of residence, previous votes, etc. [11], estimating probability of failure in various processes, systems or products [12], [13], predicting customers' propensity to purchase a product or cease a subscription in marketing applications [14]. Conditional random fields, the extended, sequential version of logistic regression for labeling or parsing sequential data, have been commonly used in natural language processing, biological sequences prediction, computer vision, etc. [15], [16], [17].

C. Ensemble method

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [18]. Similar to XGBoost, ensemble method is a popular technique employed by winning teams in Kaggle's machine learning competitions [19]. In our work, we employ a simple average ensemble method from a deep learning model and a wide learning model. While the idea of leveraging deep and wide learning has already been introduced by Cheng et al. in [20], our work is very different from their work because we combine deep and wide learning in an "ensemble method" to serve the purpose of binary classification. Cheng et al. do not utilize ensemble method, instead, they jointly train wide linear models and deep neural networks to combine the benefits of memorization and generalization for recommender systems.

III. Data Description and Feature Engineering

Before performing feature engineering, it is important to understand the original data in detail and exploit all available prior knowledge. Thus, the first part of this section is reserved for data description followed with a discussion of important features used in our model.

A. Data description

Available training data represented a collection of a total of 299680 labelled games played between 2 players from a set of total 4 bots: {A1,A2,B1,B2}. In each game, both bot-players, used one of 400 card decks, each including exactly 30 cards associated with specific hero class. Throughout the games 348 distinct cards and 9 distincts hero classses were observed. In each deck, a card must belong to either the hero class which the deck relates to or the neutral class.

The testing set included in turn only the composition of 200 new testing card decks linked to the same set of up to 9 hero classes and again the same 4 bot-players {A1,A2,B1,B2}. No details of the testing games were provided and the competition task was to estimate accurate win-rates for all 800 combinations of 4 bot-players playing off 200 testing decks.

In addition to the games inital setup data of players, decks and the result, training data also included detailed turn-by-turn gameplay data but since the same was not available for the testing set it was simply ignored.

B. Important features

Given that the task was to predict the likelihood of wining a game played by a particular bot using a specific deck, it was clear that the bot-player, hero class and the cards from the linked deck were the first choices for important features. Two types of features were extracted from the cards as follows:

- Card cardinality features: represented simply the number of cards of each type present in the deck and their observed values: {0,1,2} were mostly capturing just the presence of a specific card in a deck. We have narrowed the available set of cards down to 296 that appear in both training set and testing set since it does not help the model if we train on certain cards that do not exist in the testing set and vice versa.
- Card property features: each card has a set of properties that describe the *cost*, *health*, *attack* and *armor* of the card. In addition, there are properties to specify the card type (*hero*, *minion*, *spell*, and *weapon* [1]) and card rarity (*free*, *common*, *rare*, *epic*, and *legendary*). Since these properties are good indicators of the card's strength, we also consider them as features. However, if we build such a set of features for each individual card as card cardinality features, our number of features will increase significantly and is not manageable. Thus, we chose to generate card property features from the statistics of card properties in terms of the summary of card property values and maximum values

[1]Note that we do not consider two card types: *enchantment* and *hero power* because they do not help to improve our model's performance.

from certain card properties. In total, we generated 17 card property statistics features for each deck.

Given that the card properties are not provided in the dataset, we have to collect such information from a Hearthstone API website http://hearthstoneapi.com/.

IV. Win-rates Prediction Model

In our previous work, we proved that ensemble model usually outperforms single models in binary classification [21]. Thus, we continue to choose ensemble as our final model. In this competition, we constructed the ensemble model from three different models: two logistic regression models and one deep learning model.

A. Logistic regression models

Our first two models are based on the basic logistic regression method. As discussed in the above section, we have a total of 316 features: 1 feature for the bot id, 1 feature for the deck's hero and 297 cardinality features for the cards themselves and 17 statistics features from card properties. Initially, we trained a single logistic regression model on all features. However, we soon realized that training the model separately on either players or heroes leads to better result. Thus, we decided to employ two different logistic regression models as follows:

- The first logistic regression model consists of 4 sub-models trained separately on 4 different bots, each using 300 features KBest selected from 316 available features. Note that while we chose to use 300 features for the sub-models, these features are not the same for all sub-models. In our implementation, we used KBest to select 300 features from the training data for each sub-model and the selected features are slightly different from sub-model to sub-model.
- The second logistic regression model includes 9 sub-models trained separately on 9 different heroes of the deck, each using 100 features KBest selected from 316 available features. Similar to the above 4 sub-models trained separately for each bot, the 100 feature sets of each sub-model here are also different based on the feature selection returned from KBest method executed before training the sub-models.

It is interesting to see that the number of features used by the second model is much smaller than the number of features used by the first model (100 compared to 300). It is simply because for each sub-model of the second model, since the main difference between training data are only in cards belonging to the deck's hero or neutral cards, the total number of useful features is small.

B. Deep learning model

The third model is a deep learning model. For this model, we simply constructed a network with 5 Dense layers: 200, 100, 50, 25 and 1. For the first four layers, we used the basic *relu* activation. With the last layer, since this prediction issue is a binary classification, as expected, we used *sigmoid* for it. We compiled the model with *adam* optimizer. Our model was trained on all 316 features. For the epoch and batch size, we implemented a grid search to search for optimal parameter settings among epoch 5, 10, 15 and batch size 1,000, 2,000, 3,000, 4,000, 5,000, 10,000, 15,000, 20,000 and based on the experimental results, we chose to train the model in 5 epochs using batch size 5,000.

V. Overfiting prevention and model improvement

In this section, we will introduce several tricks we used to improve the model performance as well as to prevent overfitting.

A. Model improvement

We observed that we have two players in each game. However, so far we have only used data of the first player to train the model. We know that having more training data usually improves the model's performance. Thus, we simply tried to include training data for the model from both players of the games. It means that for each game, we use features of player 1 as the first sample and features of player 2 as the second sample. This way, the training set doubled in size up to 599,360 samples and the accuracy of our models improved from 0.4 to 0.5 point.

B. Overfitting prevention

Since the task of the competition challenge is to predict the likelihood of winning a game played by a bot using a specific deck, our initial models were trained using only data of a single bot or deck's hero as described in the above section. However, it is generally expected that the prediction accuracy would be better if we also consider information/features of the opponent player (as in practice some particular player may have better result when playing with certain players and vice versa). The problem of having extra features from the opponent player, however, is that we do not have opponent information in the test set to generate predictions. It means that to leverage information of the opponents in training the model, we need to find a way address this issue.

Fortunately, since the competition challenge description says that the opponents of games in the testing set are only selected from the set of provided 400 decks used in the training set, we decided to use a brute-force strategy to generate extra information for a test sample by considering all possible combinations of opponents, which gave us 4 * 400 = 1.600 test cases. The prediction results of these 1.600 test cases are then averaged to obtain the final prediction of the test sample. It means that given a test case of a bot and a deck, we generate a total of 1.600 test cases for that pair of bot and deck again 1.600 possible cases. It also means that we increase the size of our test size 1.600 times. In summary, the following changes were made in model training and predictions generation processes:

- We double the size of the features used in our training models, considering similar features of the opponent, in addition to features of the players. Specifically, for the two logistic regression models, we respectively used a total of 600 and 200 features selected by KBest selection method. On the the hand, the total number of features used by the deep learning model is 632.

- Given that the input size of the deep learning model has been double, we added an extra Dense layer size 400, again with *relu* activation, to the existing model, making the total number of layers to 6 instead of 5 as in the previous one.
- For the predictions, given each pair of a bot and a deck in the test set, we generated predictions for all 1,600 possible games between the bot using the deck against 4 bots * 400 decks. The final prediction is the average result obtained from these games's predictions

Note that while we double the size of the features by considering features from both the player and its opponent, we keep the number of training samples unchanged at 599,360 samples. It is because we see that training data of a bot *X* playing in a game with a bot *Y* is still different from training data of the bot *Y* in a game playing with the bot *X*.

From our submissions to the public board, we could see that this technique helped to improve our scores from 0.3 to 0.5. In addition, while we were only in the 4^{th} position of the public leader board with a gap of almost 1.0 point (a big gap) compared to the team in the 1st position, since our solution did not suffer much overfitting (which we believe that due to this strategy), we jumped to the 1st position and became the winner of the competition when the final ranking list was released.

VI. Implementation and Evaluation

We implemented all models in Python. For the logistic regression models, we relied on scikit-learn library, http://scikit-learn.org/stable/index.html. On the other hand, for the deep learning model, we employed keras.io library, https://keras.io/. Our single models respectively got score of -5.8892, -5.9148 and -6.0017 and the final ensemble model, which used a simple weighted average with coefficients 1.1, 1.08 and 1.0, reached the score of -5.4017 in the public leader board, a good improvement of approximately 0.5 compared to the results of single models. Note that we simply chose the co-efficients of the ensemble model based on the performance/scores of single models in the public leader board. Actually, if we had had more time, we could have tried a stacking technique to implement the second final layer model getting results from the first prediction models. In this way, we can also optimize the co-efficients and could even lead to a better result.

VII. Conclusion and Future Work

This paper has introduced not only our winning model from the AAIA'2018 data mining challenge, but also details of a step-by-step model building process, from the early problem and data understanding through feature engineering up to the model fine-tuning and ensembling the single models for improvements. We also prsented several techniques that we used to improve the model performance avoided model overfitting with that seemed to be critical in receiving excellent final testing score and wining the competition.

A. Future work

Even though our final result is good, as discussed in Section VI, we believe that there is still room for improvement if we have a better approach to ensemble our three single models into a final one. Besides, as discussed earlier, we did not try to utilize the massive amount of data on how games are played between the two players. Actually, we believe that the detailed game information once extracted and mined properly can provide some useful information about the tactics of the bots on playing with certain decks, and hence could be also useful to improve the model's performance.

References

[1] AAIA'18 Data Mining Challenge: Predicting Win-rates of Hearthstone Decks, https://knowledgepit.fedcsis.org/contest/view.php?id=123

[2] L. Deng, "Three Classes of Deep Learning Architectures and Their Applications: A Tutorial Survey", *APSIPA Transactions on Signal and Information Processing*, 2012

[3] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 17981828, 2013.

[4] D.R. Cox, "The regression analysis of binary sequences (with discussion)," *J Roy Stat Soc B.*, vol. 20, pp. 215242, 1958.

[5] C.R. Boyd, M.A. Tolson, and W.S. Copes, "Evaluating trauma care: The TRISS method. Trauma Score and the Injury Severity Score," *The Journal of trauma*, vol. 27, no. 4, pp. 370378, 1987.

[6] M. Kologlu, D. Elker, H. Altun, and I. Sayek, "Validation of MPI and OIA II in two different groups of patients with secondary peritonitis," *Hepato-Gastroenterology*, vol. 48, no. 37, pp. 147-151, 2001.

[7] S. Biondo, E. Ramos, M. Deiros, et al. "Prognostic factors for mortality in left colonic peritonitis: a new scoring system," *J. Am. Coll. Surg.*, vol. 191, no. 6. pp. 635-642, 2000.

[8] J.C. Marshall, D.J. Cook, N.V. Christou, et al. "Multiple Organ Dysfunction Score: A reliable descriptor of a complex clinical outcome," *Crit. Care Med.*, vol. 23, pp. 16381652, 1995.

[9] J.R. Le Gall, S. Lemeshow, and F. Saulnier, "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study," *JAMA.*, vol. 270, pp. 29572963, 1993.

[10] J. Truett, J. Cornfield, W. Kannel, "A multivariate analysis of the risk of coronary heart disease in Framingham,", *Journal of chronic diseases*, vol. 20, no. 7, pp. 511524, 1967.

[11] F.E. Harrell, Regression Modeling Strategies, Springer-Verlag, ISBN 0-387-95232-2, 2001.

[12] M. Strano, B.M. Colosimo "Logistic regression analysis for experimental determination of forming limit diagrams," *International Journal of Machine Tools and Manufacture*, vol. 46, no. 6, pp. 673682, 2006.

[13] S.K. Palei, S.K. Das, "Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach," *Safety Science*, vol. 47, pp. 8896, 2009.

[14] M.J.A. Berry, "Data Mining Techniques For Marketing, Sales and Customer Support," Wiley, pp 10, 1997.

[15] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. 18th Int. Conf. on Machine Learning. Morgan Kaufmann*, pp. 282289, 2001.

[16] X. He, and R.S. Zemel, and M.A. Carreira-Perpinn, "Multiscale conditional random fields for image labeling," *IEEE Computer Society*, 2004.

[17] K.Y. Chang, T.p. Lin, L.Y. Shih, and C.K. Wang, "Analysis and Prediction of the Critical Regions of Antimicrobial Peptides Based on Conditional Random Fields," *PLoS ONE*, 2015.

[18] T. G. Dietterich, "Ensemble Methods in Machine Learning", *Proc. of the* 1^{st} *Int. Workshop on Multiple Classifier Systems*, pp. 1-15, 2000.

[19] Kaggle, https://www.kaggle.com/.

[20] H. T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong , V. Jain, X. Liu and H. Shah, "Wide & Deep Learning for Recommender Systems", *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS, pp. 1-10, 2016

[21] Q. H. Vu, D. Ruta, L. Cen. "An ensemble model with hierarchical decomposition and aggregation for highly scalable and robust classification", *Proceedings of the AAIA*, 2017

6th International Workshop on Smart Energy Networks & Multi-Agent Systems

OUR energy supply infrastructure is in the middle of a transition from a conventional star-like energy supply topology with a manageable number of well-structured power plants towards a grid topology with a myriad of different generation units that are geographically widely distributed. Additionally, the increasing integration of volatile and intermittent renewable energy resources brings massive challenges to grid operations and its composition with respect to power system commitment, dispatching and reserve requirements.

The fact that renewable energy generation units will increase their share in the overall energy production, calls for technologies to be developed in the next decades to deal with the transition of the energy supply system and the distribution of renewable energy generation units. This includes technologies to integrate, handle and intelligently manage energy storage systems, grid load peak-shaving, smart supply system components, more efficient and intelligent coupling of heating with electrical power, heat storage, intelligent load shifting and balancing, to name only a few here.

All these have in common that the future power grid has to be intelligent, where generation and consumption units communicate or even negotiate their offer or their demand of energy through an 'internet of energy'. Thus, to efficiently design and develop those distributed energy management systems is one of the key challenges to be solved to transform the energy supply system, addressing distributed coordination, as well as different forms of energy like electricity, heat, natural gas and other.

Information and communication technologies are the key enablers of such envisioned systems, where especially the agent-paradigm provides an excellent modelling approach for the distributed character of energy systems. Although significant efforts and investments have already been allocated into the development of smart grids, there are, however, still significant research challenges to be addressed before the promised efficiencies or visions can be realised. This includes distributed, collaborative, autonomous and intelligent software solutions for simulation, monitoring, control and optimization of smart energy networks and interactions between them.

TOPICS

The SEN-MAS'18 Workshop aims at providing a forum for presenting and discussing recent advances and experiences in building and using multi-agent systems for modelling, simulation and management of smart energy networks. In particular, it includes (but is not limited to) the following topics of interest:

- Experiences of Smart Grid implementations by using MAS
- Applications of Smart Grid technologies
- Distributed energy management of distributed generation and storage based on MAS
- Examples of design patterns for MAS in distributed energy management systems
- Microgrids, Islands Power Systems
- Real time control of energy networks
- Distributed planning process for energy networks by using MAS
- Self-configuring or self-healing energy systems
- Load modelling and control with MAS
- Simulations of Smart Energy Networks
- Software Tools for Smart Energy Networks
- Energy Storage
- Electrical Vehicles
- Charge scheduling for electric vehicles (and fleets) based on MAS
- Interactions and exchange between networks for electricity, gas and heat
- Stability in Energy Networks
- Distributed Optimization in Energy Networks
- Safety and security issues for MAS in Smart Grids

TUTORIAL

Beside the scientific exchange, the event will provide a practical tutorial for building so called Energy Agents. Based on the open-source framework Agent.GUI, the JADE agent platform and the experiences in the project Agent.HyGrid, the tutorial will guide you through the development process that enables you to build agents that can be installed and executed beside distributed energy systems.

EVENT CHAIRS

- **Brehm, Robert,** University of Southern Denmark, Denmark
- **Derksen, Christian,** University Duisburg-Essen, Germany

PROGRAM COMMITTEE

- **Bilal, Bilal**
- **Bremer, Joerg,** joerg.bremer@uni-oldenburg.de, Germany
- **Fortino, Giancarlo,** Università della Calabria
- **Hildmann, Hanno,** Universidad Carlos III de Madrid (UC3M), Spain

- **Karnouskos, Stamatis,** SAP, Germany
- **Klusch, Matthias,** German Research Center for Artificial Intelligence, DFKI, Germany
- **Loose, Nils**
- **Moench, Lars,** FernUniversität Hagen, Germany
- **Nieße, Astrid,** Leibniz Universität Hannover, Germany
- **Paprzycki, Marcin,** Systems Research Institute Polish Academy of Sciences, Poland
- **Redder, Mareike**
- **Sonnenschein, Michael,** Professor (retired) at the University of Oldenburg, Germany
- **Sudeikat, Jan,** Hamburg Energie GmbH, Germany
- **Vale, Zita**

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 203–213
DOI: 10.15439/2018F282
ISSN 2300-5963 ACSIS, Vol. 15

Smart Micro-scale Energy Management and Energy Distribution in Decentralized Self-Powered Networks Using Multi-Agent Systems

Stefan Bosse
University of Koblenz-Landau
Faculty Computer Science, Institute of Software Technology,
Koblenz, Germany
Email: sbosse@uni-bremen.de

***Abstract*—Energy distribution as a main part of energy management in self-powered micro-scale networks like sensor networks is a challenge with the goal to satisfy a safe and reliable operational state on system and node level. Under the assumption that nodes are arranged in mesh-like networks with links posing the capability to transfer data and energy between nodes a self-organizing Mulit-agent System based on divide-and-conquer is deployed in this work successfully to distribute energy without a system/world level model and knowledge of single nodes about the system state. Different agent behaviour were investigated and the emergence evaluated. An exploring help strategy with energy deliver child agents showed the best and efficient overall behaviour. Mobile agents were programmed in JavaScript using the JavaScript Agent Platform that can be deployed in strong heterogeneous environments.**

I. Introduction

AMONG energy supply and consumer networks on a macro-scale level there are sensor networks with self-powered sensor nodes consisting of an energy storage and energy supply. Both classes of networks require distributed, adaptive and self-organizing energy management to satisfy (1) A balance of energy supply and energy consumption, and (2) Operational stability on system level **[1]**. The energy management addressing the control of consumption and production is basically a distributed resource sharing and scheduling problem **[2]**. Sensor networks consists of nodes optionally equipped with an energy harvester collecting energy from the environment from different sources posing varying availability that cannot be controlled, in contrast to macro-scale energy sources (power engines, ..). Power management and energy harvesting are central issues in sensor networks **[3]**. Additionally, self-powered nodes can be supplied by external energy sources not directly attached to the node using other nodes to transfer energy. Nodes in a sensor network can use communication links to transfer energy, for example, optical links are capable of transferring energy using Laser or LE diodes in conjunction with photo diodes on the destination side, with a data signal modulated on an energy supply signal.

Typically, energy management is performed by a central controller on software level **[1]**, with limited fault robustness and the requirement of a well-known environment world model for energy sources, sinks, and storage. In a centralized approach, energy is only transferred on node level. With a distributed approach, energy management in a network involves the transfer of energy between node instances, too. In **[4]**, multi-agent systems (MAS) are used to perform energy management (between sinks and sources) in a renewable energy grid by solving a (global) optimization problem by the MAS. In **[1]**, a MAS performs energy distribution by a token-based approach solving an optimization problem, too. In **[5]**, MAS based on the Belief-Desire-Intention architecture (BDI) are deployed in distributed sensor networks to perform goal and knowledge orientated energy management (but without considering energy distribution). These examples pose the benefit of agent-based systems solving energy management and distribution in large-scale networks. Agents are already deployed in industrial applications and the Industrial IoT **[6]**.

We propose a smart energy management and distribution approach for a broad range of applications ranging from micro-scale sensor network architectures to large-scale energy networks with nodes supplied by 1. energy collected from a local source (energy harvesting, **[3]**), and 2. by energy collected from neighbour nodes using smart energy management (SEM) and self-organization, based on early work investigating primarily technological aspects in sensorial materials and agent-on-chip hardware architectures **[7]**. For the sake of simplicity, nodes are arranged in a n-dimensional grid with connections to their direct neighbours, i.e., in a three-dimensional network there exist up to six connections in directions North, South, East, West, Up, and Down. It is assumed that the network is irregular (missing nodes) and incomplete (missing links). Each node can store collected energy and distribute energy to neighbour nodes via communication links.

Each autonomous node provides communication, data processing, and energy management. There is a focus on single System-On-Chip (SoC) design satisfying low-power and high miniaturization requirements addressing the micro-level as well as macro-scale networks with power supplies and consumers.

Energy management is performed 1. For the control of local energy consumption, and 2. For collection and dis-

tribution of energy by using the data links to transfer energy.

Considering strong heterogeneous networks and host platforms with a loosely coupling of nodes arranged in grids the distributed data processing is a challenge. Multi-agent systems (MAS) poses a distributed computation and communication model providing autonomy, self-organization, and group behaviour. MAS are already deployed in energy management and energy distribution systems**[8][9][1]**. Most published MAS perform negotiation between energy producer and consumer under varying environmental, node and system level states. The negotiation results effect and control energy production and consumption. In contrast to public energy markets, self-powered sensor networks do not provide such a control as energy harvesting is strongly influenced by not controllable environmental conditions (e.g., sun shine duration influencing solar cell harvesting efficiency).

This work investigates and evaluate the emergence behaviour of self-organizing energy management agents that are capable to transfer (virtually carry) energy between nodes of networks and that are deployed in large-scale decentralized energy supply and consumer networks under resources and reliability constraints, e.g., self-powered sensor networks**[10]**. The desired emergence is the efficient improvement of the energy distribution in such networks to satisfy operational stability of the entire network on system level and in bounded regions, i.e., avoiding nodes with too low energy being operational. Agents perform decision making based on actual and past node energy, reward, and interaction with other agents.

The next sections introduce the underlying energy model on node level, energy management and distribution, the MAS and the agent processing platform (APP), finally used in an evaluation of a large-scale network simulation.

II. Energy Model

There is no system level model in this work considering only the node level energy. The total energy of a network node is a balance of energy loss due to computation, communication, and agent creation, and energy harvesting via energy delivery by agents and local energy harvesters (power supplies). The energy balance equation is shown in Eq. **1.** and used throughout this work.

$$\begin{aligned} E_{node}(t) \quad = \quad & e_0(t_0) - \sum k_{decay}(t) \\ & - \sum k_{comp}\tau(Ac_i) \\ & - \sum k_{create}Ag_i(AC) \\ & - \sum k_{comm}k_{link}\sigma(msg_i) \\ & + \sum l_{conv}l_{link}e_{i,deliver} + \sum l_{conv}ha_i \end{aligned} \tag{1}$$

It is assumed that the time variable of the energy E is a discrete variable with a time resolution between milliseconds and seconds. The initial energy is e_0 with a time-dependent decay k_{decay} due to losses in the energy storage of a node. The energy is reduced by agent activity Ac_i(with $\tau(Ac)$ as the execution time of an agent activity scaled by a parameter k_{comp}), agent creation $Ag(AC)$ from class AC (can be expressed by a computational time, too). Communication further requires energy and depends on the size of the message$\sigma(msg)$ scaled with the parameter k_{comm} and the link specific energy consumption given by the parameter k_{link}. Finally, energy is increased by agents delivering energy via the communication links (amount can vary, and energy conversion losses are covered by the parameter l_{conv} and the loss of the link by l_{link}) and from local power sources (ha, again covering conversion losses by the parameter l_{conv}).

Nodes are classified by their *energy deposit* and operational state:

1. A node with very low energy $E<E_{Alarm}$, with restricted node operation (only agents arriving with energy are processed, only help emergency agents are sent out).
2. A node with low energy $E<E_{Thres1}$, resulting in a basically normal node operation but with execution limits (number of agents, agent processing time, agent creation, agent migration, agent class restrictions, increased barriers of processing negotiation).
3. A node with normal energy deposit $E<E_{Thres2}$ and a normal node operation state with some resource limitations. All agents are processed and the node agent can create any type of energy agents.
4. A node with very high energy $E>E_{Alarm}$, with node operation being normal with only a few or no resource limitations. All agents are processed and the node agent will only create distribute energy agents (if behaviour was enabled).

For the sake of simplicity a three-dimensional grid network connecting spatial neighbour nodes is assumed, shown in Fig. **1**. Connected nodes can exchange data and energy. The shown example network consists of three layers (z-axis, levels) and each layer consists of 5x8 nodes. Each node has any time t an energy deposit $0>E(t)>E_{max}$, illustrated in Fig. **1** by different colors (blue: low energy, red: high energy).

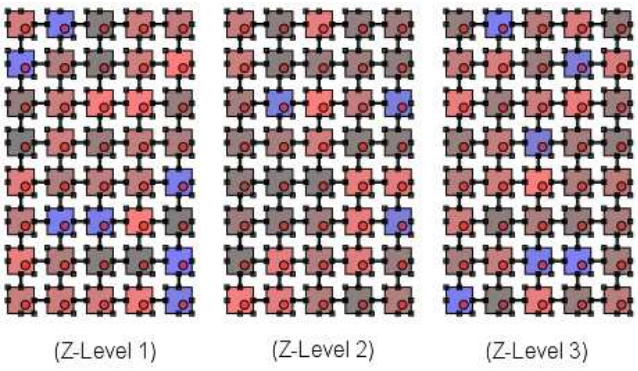


Figure 1: Example of an energy distribution in a three-dimensional network. Squares: Nodes (red color indicates enough node energy - good node - to be operational, blue color indicates critical low energy capacity - bad node), Circles: Agents, Lines: Communication and Energy links. The different z-levels are connected by up- and down links.

III. Energy Management and Distribution

In this work mobile agents perform energy management and distribution, discussed in detail in Sec. IV. The energy management relies on three main principles:

A. Adaptive Routing

Agents are responsible to find an appropriate path between a source and a distinct destination node by using adaptive routing or by using data centric routing linking an information (energy) supplier and an information (energy) sink.

B. Energy Transport

Mobile agents can carry energy tokens being able to be transferred between nodes. An energy token is requested on a node (granted or negotiated by the node agent) and can be migrated to other nodes. Each time an agent arrives on a new node it delivers its energy to the node energy deposit. If the agent has to transport the energy token to another node the agent has to collect the energy token again reduced by some technical conversion losses, shown in Fig. **2**. This approach is reasonable with respect to a technical representation of energy transfer in networks via communication links.

C. Negotiation

To avoid a high density of help and request agent populations on a specific neighbour node each help and request agent places temporary markings on the node indicating energy demand on this (good or very good) node (aka. synthetic pheromones) by other nodes. These markings are placed in the tuple space of the node and removed after a time-out automatically. If a new help or request agents arrives on this nodes and this node has a strong marking it will continue traveling (help behaviour) or dies (request behaviour). Each help and request agent negotiates energy demand with the node agent via the tuple space (by accessing active tuples using the evaluate operation and placed by the node agent using a listen operation).

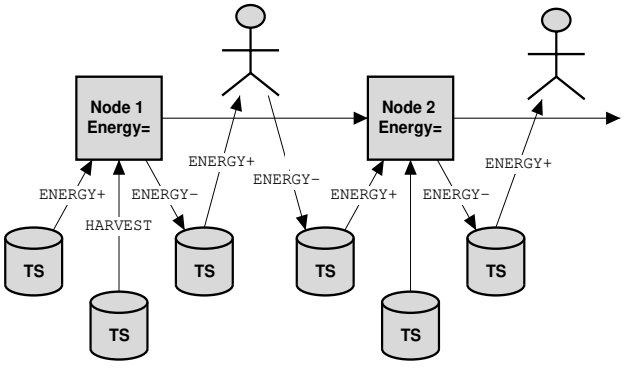


Figure 2: Energy transport by agents: Each time an agent carrying energy tokens arrives at a new node the energy is stored in the node deposit and a virtual energy tuple is stored in the tuple space. If the agent continues travelling it has to collect the energy token again.

D. Energy Management

The following parameters are used for the dynamic energy management (which can be changed at run-time) performed by the MAS:

Energy Thresholds: $E_{Alarm}<E_{Thres0}<E_{Thres1}$, $E_{Deposit}$
Energy Transfer Tokens: E_{Req},E_{Dist}
Efficiencies: k_{Conv},k_{Comm}
Exploration Radius: r_{expl}
Agent Lifetime: τ_{max}
Maximal hop-count: h_{max}

IV. MAS

The Multi-agent system (MAS) used for distributed energy management and energy distribution consists of different agent classes posing different behaviour and goals. Each network node part of the distributed energy management system provides an agent processing platform (APP). At least one stationary (non-mobile) node agent is created on each node to initialize the sensor processing and energy management. Based on the current state of the node and the power history of the node the node agent will create further energy management agents, summarized in Tab. **I**. All other energy management agents are mobile and can migrate (travel) along a path in the network crossing multiple nodes.

The node management agent can choose different energy management strategies: *Help*, *request*, and *distribute*. Depending on the operational and energy state of the node one or multiple strategies are applied to improve the energy deposit.

The different behaviour of the energy management agent is shown in the activity-transition diagram in Fig. **3**. Different parameters have an impact on each behaviour at run-time. Each agent owns a set of body variables, e.g., a delta vector indicating the position in the network, a charge and energy variable for distributed energy management. Most central parameters are the current positionΔ(relative to a source node) and the energy deposit of the current node *E* that is retrieved via the tuple space and provided by the node agent that interacts with power components via a HAL of the physical node. Energy agents are either created by the node agent (*help*, *request*, *distribute*) or by already created energy agents (*reply*, *deliver*).

Each agent can carry virtual energy as outlined in Sec. III. Since agents are mobile they can transfer energy from a node *A* to a new node *B* via the communication link (or any other power link between these nodes) just by migration. The energy transfer is handled by the APP on migration if the*charge* agent body variable holding a mobile energy token is greater than zero. Before travelling the *charge* value is transferred to the *energy* body variable of the agent to store the energy virtually. After the arrival on a new node the value of the*energy* variable is transferred back to the *charge* variable. There are now two possible cases: The agent delivers the energy finally on the current node (technically the energy was already transferred to the local energy storage), or it continues travelling and transfers the energy again to the next node via links.

Agent Class	Behaviour
`Node`	This stationary agent monitors the node state and power history. It has to initiate appropriate actions, i.e., creation of energy request, help, or distribute agents. The node agents has to asses the quality of the SEM locally and can change SEM strategies.
`Request`	*Point-to-point agent*: This mobile agent requests energy from a specific destination node, returned with a `Reply` agent. If the destination node cannot deliver energy (bad node), the request agent dies without a reply.
`Reply`	*Point-to-point agent*: Mobile reply agent created by a `Request` agent, which has reached its destination node. This agent carries energy from one node to another.
`Help`	*ROI agent*: This mobile agent explores a path starting with an initial direction and searches a good node having enough energy to satisfy the energy request from a bad node. This agent resides on the final good node (found by random walk within a region) for a couple of times and creates multiple deliver agents periodically in dependence of the energy state of the current node. If the current node is not suitable anymore, it travels to another good node.
`Deliver`	*Path agent*: This mobile agent carries energy from a good node to a bad node (response to `Help` agent). Depending on selected sub-behaviour (`HELPONWAY`), this agent can supply bad nodes first, found on the back path to the original requesting node.
`Distribute`	*ROI agent*: This mobile agent carries energy from the source node to the neighbourhood and is instantiated on a good node. It explores a path starting with an initial direction and searches a bad nodes to supply them with the energy from the agent virtual energy deposit.

Table I: SEM agents with different behaviour used to manage and distribute energy in bounded regions (ROI: region of interest) based on negotiation.

The movement of mobile agents are constrained by three parameters: The maximal hop count, the maximal mobility radius relative to the source node, and a maximal lifetime. The constrained mobility ensures a relaxation and limitation of the population of the MAS after a stimulus occurred, i.e., energy decrease or increase that can trigger the creation of energy agents.

The request and distribute strategies are the most simple ones that can be performed by bad and good nodes, respectively. The help strategy is more advanced usually performed by bad and very bad nodes. A variation of the help strategy adds help-on-way behaviour performed by the deliver agents to charge bad nodes on the way back to original requesting bad node, too.

Agents perform decision making based primarily on the node energy class (bad/good), but also on actual and past node energy recording, reward and utility feedback for charging their home node, and interaction with other agents.

V. Agent Platform

In this work agents are programmed and implemented in *JavaScript* using the *JavaScript* Agent Machine platform (*JAM*) and the *AgentJS* programming language used for the implementation of the state-based reactive agents.

In the considered use-case scenario the MAS is deployed in a large-scale strongly heterogeneous network environment, which can be additionally extended with mobile devices. This heterogeneous network requires a unified agent processing platform (APP), which can be deployed on a wide variety of host platforms, ranging from embedded devices, mobile devices, to desktop and server computers. E.g., some measuring stations are attached to buoy or installed on small islands, equipped only with low-power low-resource computers. To enable seamless integration of mobile MAS in Web and Cloud environments, agents are implemented in *JavaScript*(*JS*), executed by the *JS* Agent Machine (*JAM*), implemented entirely in *JS*, too. *JAM* can be executed on any *JavaScript* engine, including browser engines (Mozilla's *SpiderMonkey*), or from command line using*node.js*(based on *V8*) or *jxcore*(*V8* or *SpiderMonkey*), or a low-resource engine *JVM*. The last three extend the *JS* engine with an event-based (asynchronous using callback functions) IO system, providing access of the local file system and providing Internet access. But these*JS* engines have high resource requirements (memory), preventing the deployment of *JAM* on low-power and low-resources embedded devices. For this reason,*JVM* was invented. This engine is based on *jerryscript* and *iot.js* from Samsung, discussed in **[11]**. *JVM* is a Bytecode engine that compiles*JS* directly to Bytecode from a parsed AST. This Bytecode can be stored in a file and loaded at run-time. *JVM* is well suited for embedded and mobile systems, e.g., the Raspberry PI Zero equipped with an ARM processor.*JVM* has approximately 10 times lower memory requirement and start-up time compared with *nodes.js*.

JAM consists of a set of modules, with the Agent Input Output System (*AIOS*) module as the central agent API and execution level. The deployment of agents in the Internet requires an additional Distributed Organization Layer (*DOS* with capability-based security). *JAM* is available as an embeddable library (*JAMLIB*). The entire *JAM* and *DOS* application requires about 600kB of compacted text code (500kB Bytecode), and the *JAMLIB* requires about 400kB (300kB Bytecode), which is small compared to other application programs and commonly used Java-based platforms like *JADE*/*AgentSpeak*. *JVM+JAMLIB* requires only 3 MB total RAM memory on start-up.

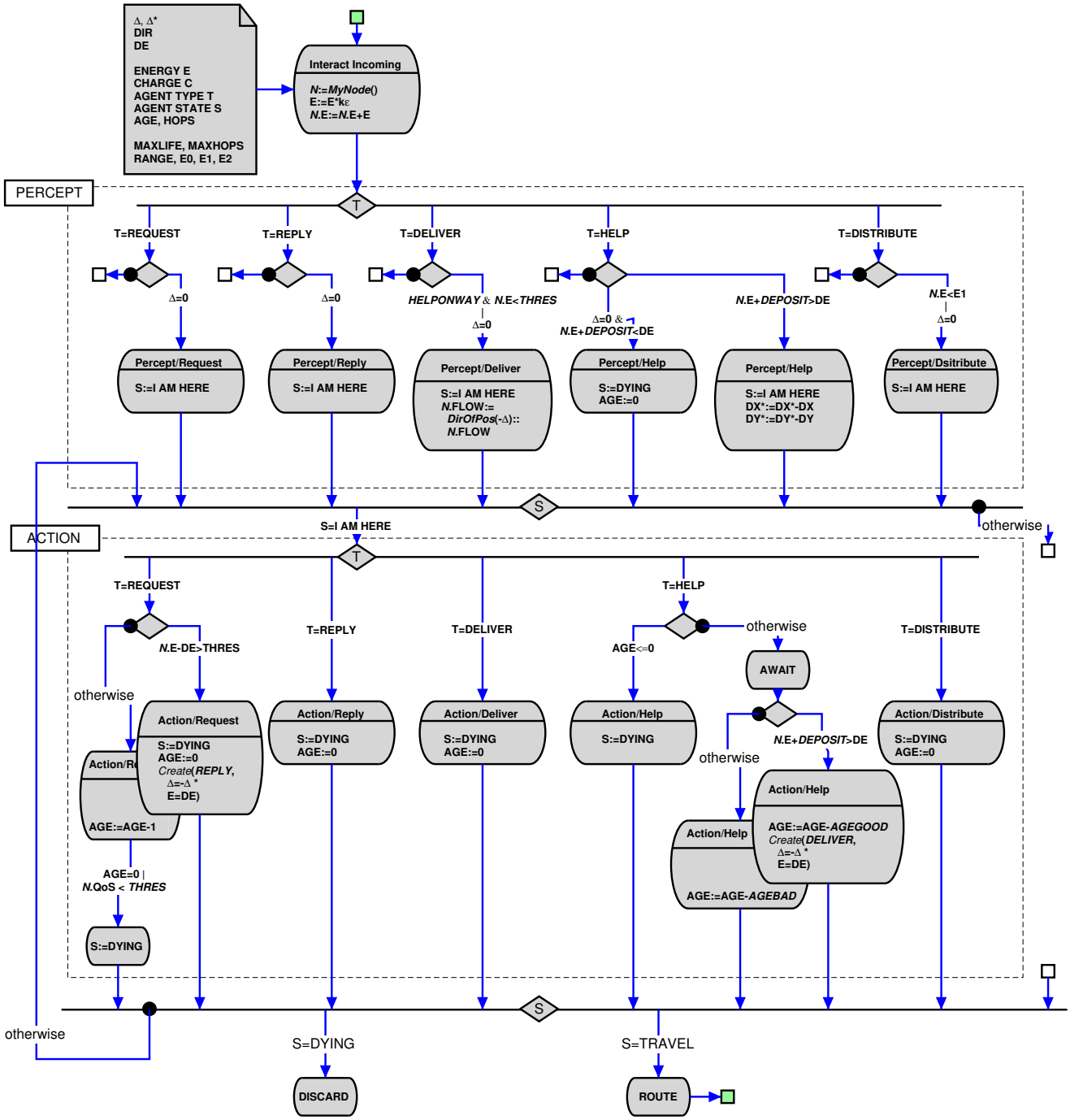


Figure 3: Principle activity diagram of the Energy Agent behaviour. There are five different agent classes (or sub-classes of the energy management agent) differing in their behaviour: Request, Reply, Help, Deliver, and Distribute.

JAM is capable of handling thousands of agents per node, supporting virtualization and resource management. JAM agents can migrate between different (physical) node APPs supporting true agent mobility with process snapshots preserving and embedding the entire agent state with low-resource overhead. Depending on the used*JS* VM, agent processes can be executed with nearly native code speed. *JAM* provides Machine Learning as a service that can be used by agents. The agent only saves a learned model, but not the learner code. Agent interaction and synchronization is provided by exchanging data tuples stored in a tuple space on each*JAM* node.

The agent behaviour is modelled according to an Activity-Transition Graph (ATG) model. The behaviour is composed of different activities representing sub-goals of the agent, and activities perform perception, computation, and inter-action with the environment (other agents).

Agent interaction (communication) is performed by using tuple spaces and mobile signals (point-to-point or point-to-N messages). Using tuple spaces is a common approach for agent communication, as proposed by**[12]**, much simpler than **[13]** proposed with *AgentSpeak*. The transition to another activity depends on internal agent data (body variables). The ATG is entirely programmed in *JavaScript* (*AgentJS*, see **[14]** for details). The ATG can be modified by agents at run-time enabling code morphing and optimization (behaviour adaptation or sub-classing).

JAM agents are mobile, i.e., a snapshot of an agent process containing the entire data an control state including the behaviour program, can migrate to another *JAM* platform. *JAM* provides a broad variety of connectivity, available on a broad range of host platforms. Although *JAM* is used in this work as a simulation platform in the *SeJAM* simulator only, it is ready to use in real-world networks and is capable to execute thousands of agents. The *SeJAM* simulator is built on top of a *JAM* node adding simulation control and visualization, and can be included in a real-world closed-loop simulation with real devices. Since *JAM* can be embedded in any host application and its capability to be easily extended enables the binding of *JAM* to low-level power management and technical energy components (conversion, storage, transfer). A hardware abstraction layer (HAL) enables the access of the power components by agents completely via the tuple space or by extended *AIOS* functions.

VI. Simulation and Evaluation

The simulation was carried out with the *SeJAM* simulator and a three-dimensional grid network consisting of three z-levels (layers), 8 rows, and 5 columns, as already shown in Fig.**1**. Each node of the network is a virtual *JAM* instance connected with up to six neighbouring nodes via serial links. Each node is associated with a virtual energy storage and energy harvester. The links are capable to transfer data and energy as introduced. Each node is populated with at least one node agent. An artificial world agent controls the simulation, perform monitoring, and reforms Monte Carlo simulation of the energy harvesting, and the initial start condition with respect to the initial energy deposit of each node. The randomized energy distribution assign nodes with an initial energy storage in the range [e_1=50,e_2=300](arb. energy units). Depending on the energy threshold settings there is initially a fraction of bad and very bad nodes about 20% of the total number of 120 nodes. In periodic intervals the nodes are charged with randomized energy amounts in the interval [0,e_Δ=0.3].

Each simulation run consists of 3000 simulation steps (in all considered cases the SEM converged either during this simulation range or never).

A typical parameter set used by the MAS is shown below.

```
parameter:{
 energy1:50, Energy range of nodes
 energy2:300,
 energyAlarm:50,   e < e_Alarm: Very Bad Node
 energyThres0:100,  e < e_Thres0: Bad Node
 energyThres1:200,  e > e_Thre1: Very Good Node
 energyDeposit:50,  Reservoir
 energyRequest:50, Def. en. to requeste
 energyDistribute:20, Def. en. to distribute
 explorationRange:4,
 maxLife: 4,
 maxHops: 8,
 Inhibit request/help agent send out
 inhibitTime: 20,
 Internal energy conversing efficiency
 energyK: 0.95,
 energyCommK: 0.8, Energy transfer efficiency
 sem: ['help'],  Energy management strategy
 doHarvest:true,
 harvest:0.3,
 cpuK:0.3,
 createK:3,
}
```

One important outcome of the simulation was the observation that the emergent behaviour on system level depends on the starting condition of the network, i.e., the initial energy distribution. That means the result of the MAS operation can vary significantly under different situations discussed below.

Typical examples of the run and progress of different energy management strategies with respect to the node classification population (very bad, bad, good, very good) are shown in Fig. **5**. Without SEM (not shown), there is commonly no change in the network situation, i.e., the number of bad nodes (typically about 20%) remains unchanged. Using SEM, the MAS is capable to decrease the fraction of all bad nodes below 1%. All three SEM strategies request, help, and help-on-way, show a fast convergence and reaching of the goal to minimize bad nodes but still preserving a high amount of good and very good nodes. The help behaviour has the fastest convergence time, usually eliminating all bad and very bad node states, whereas the request behaviour has a slower convergence time. The help-on-way behaviour can create a remaining fraction of bad and very bad nodes and seems not be appropriate to satisfy the system level goal.

Typical variations of the run and progress of different energy management strategies are shown in Fig. **6**. The request behaviour poses a lower stability in the final outcome of the MAS than the help and help-on-way behaviour. Although help-on-way seems to be more reliable, it tends to create very bad node cluster as shown in Fig. **4**.

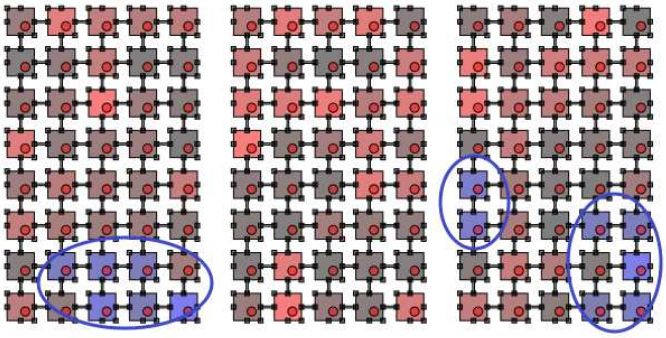


Figure 4: Formation of isolated very bad node clusters

One important measure is the temporal development of energy agents during a SEM run and the total number of created energy agents (help, request, distribute, deliver, reply), shown in Fig. **7**. The help behaviour performs optimally (both concerning the convergence time and the number of agents required). The help-on-way behaviour shows the aforementioned instability and missing convergence.

Fig. **7** summarizes the evaluation of the impact of different SEM agent parameters (parameter sets *B*-*F*). The parameter sets are explained in App. A. Parameter sets *B*-*E* are used to investigate the help MAS behaviour with different maximal help agent life times (staying on a good node and sending out deliver agents). Increasing the lifetime increases the total number of created energy agents without decreasing the fraction of bad nodes significantly. A fraction about 4% still remains (with large variations). But increasing the exploration range and the maximal number of agent hops results always (regardless of the initial start condition) in 0% bad nodes!

Finally, the energy efficiency of the MAS (defined as the fraction of start+harvest energy/final energy) was analyzed and is shown in the center plot of Fig. **8**. All parameter sets show a high efficiency between 87-92%.

The energy equation Eq. **1.** is updated during simulation about every 50 simulation step providing a sufficient smooth change of E based on updated perception and energy harvesting activities.

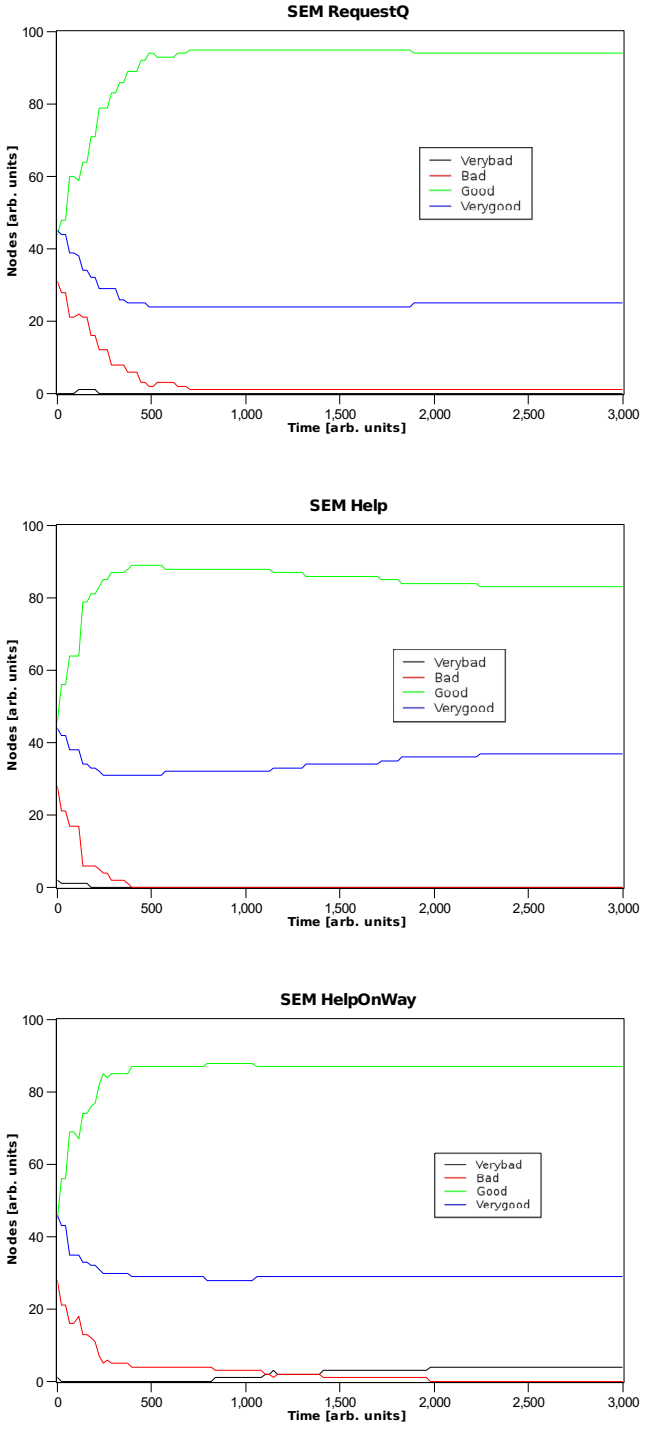


Figure 5: Typical examples of the run and progress of different energy management strategies with respect to the node classification population. (Top) SEM with RequestQ behaviour (Middle) Help behaviour (Bottom) Help On Way behaviour [x-axis: simulation time in arbitrary units, y-axis: node number]

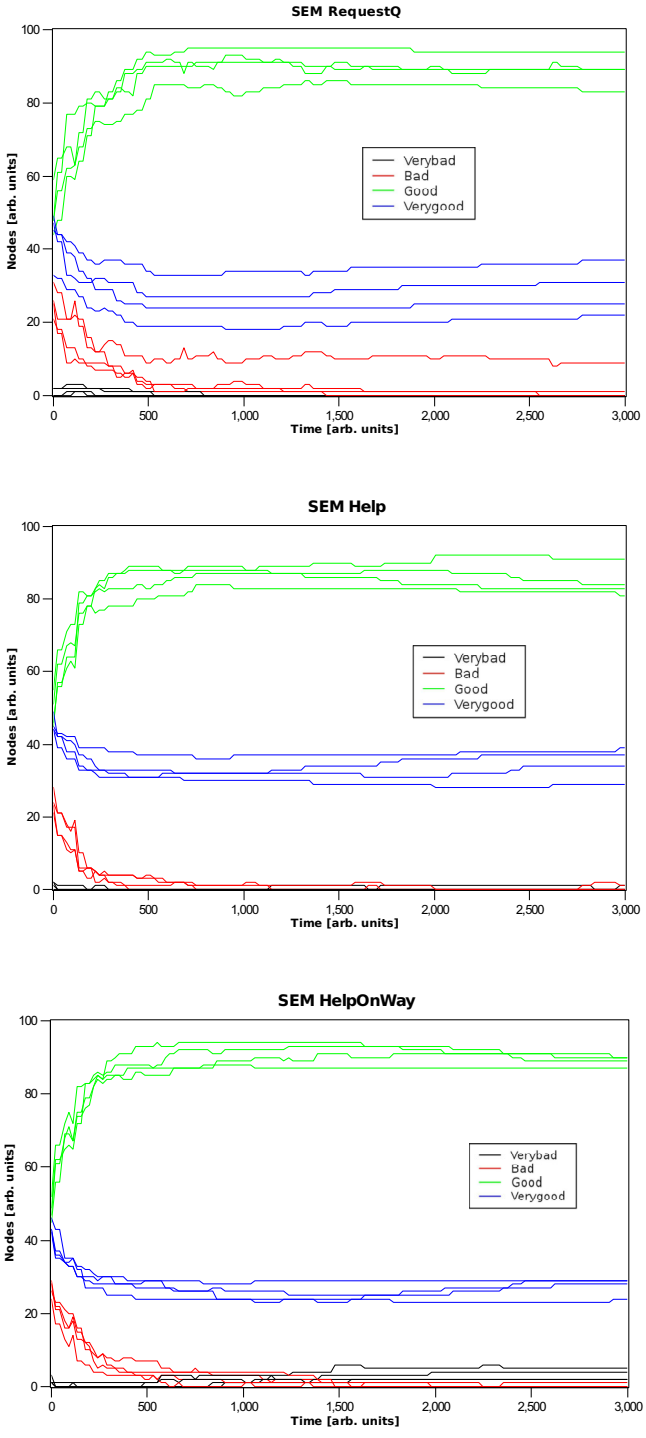


Figure 6: Typical variations of the run and progress of different energy management strategies with respect to the node classification population. (Top) SEM with RequestQ behaviour (Middle) Help behaviour (Bottom) Help On Way behaviour [x-axis: simulation time in arbitrary units, y-axis: node number]

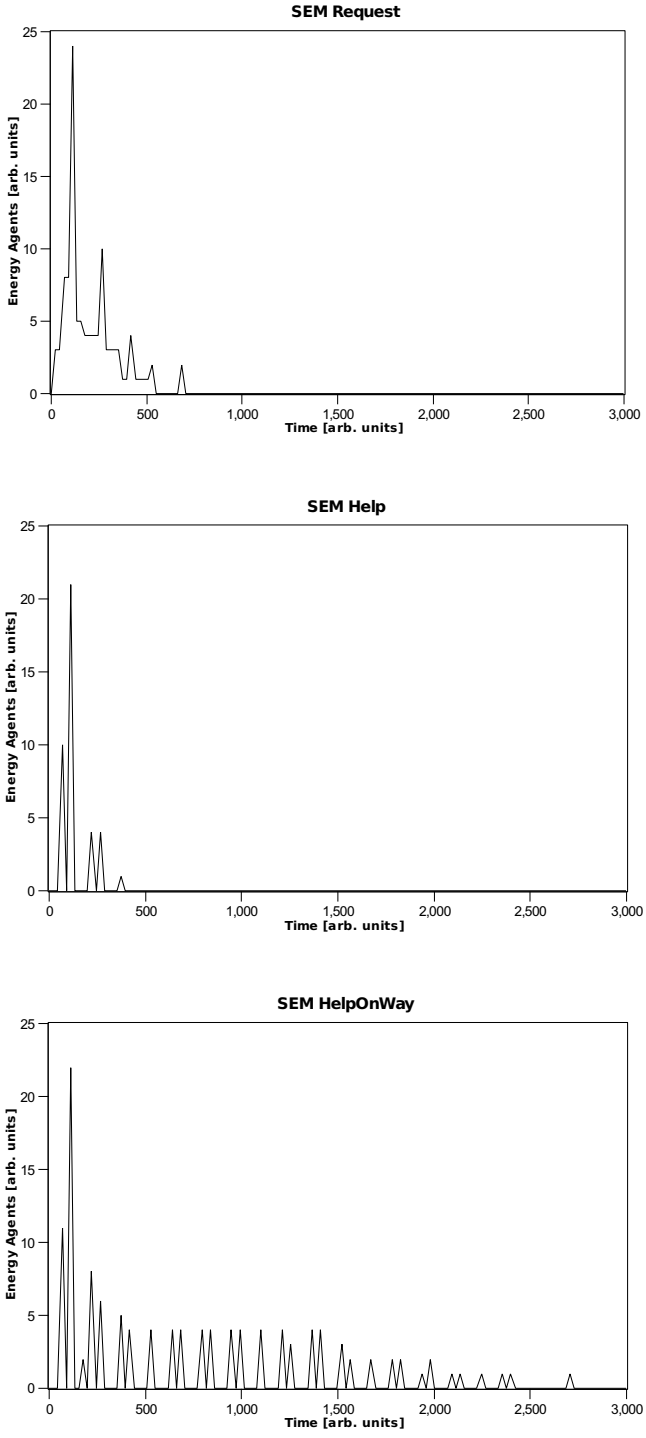


Figure 7: Typical temporal energy agent populations of runs with different energy management strategies. (Left) SEM with RequestQ behaviour (Center) Help behaviour (Right) Help On Way behaviour [x-axis: simulation time in arbitrary units, y-axis: agent number]

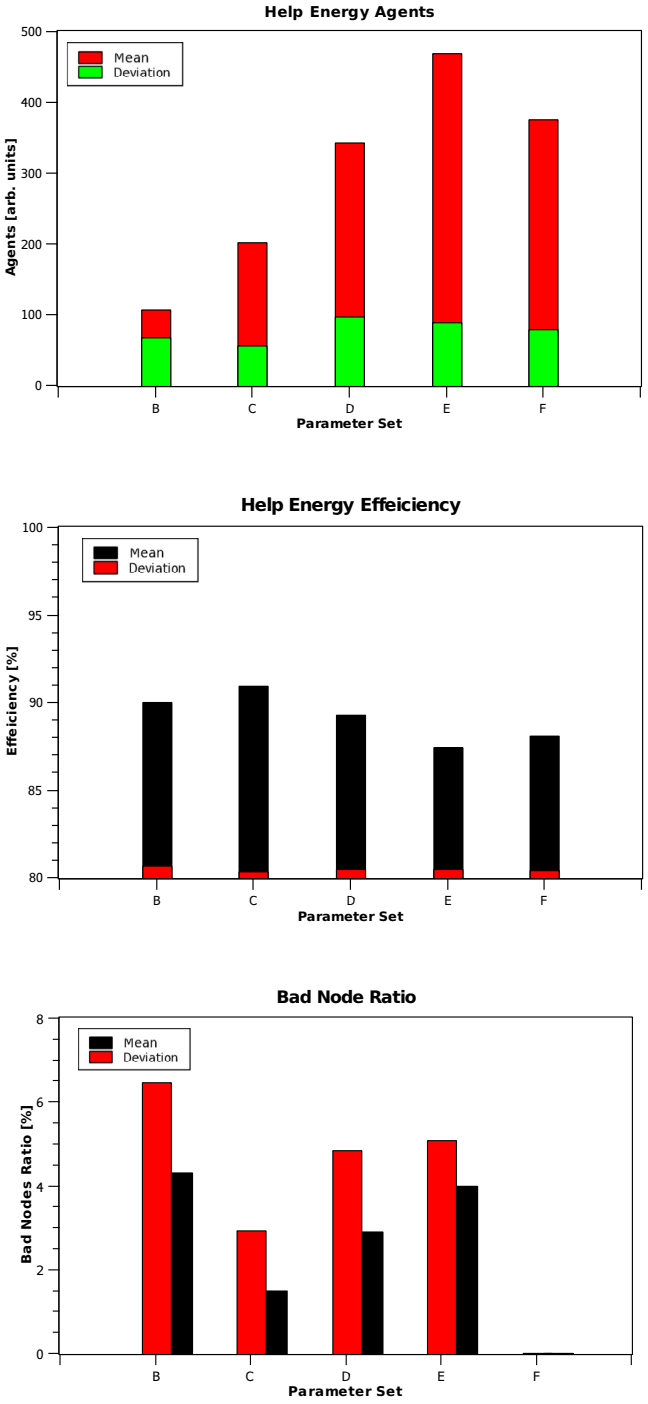


Figure 8: Analysis of the impact of different parameter sets of the SEM help behaviour. (Left) Total number of energy agents created (Center) Energy Efficiency comparing start, harvested, and final energy sum on system level (Right) Bad node ratio (before SEM/after SEM) [x-axis: parameter set, y-axis: agent number, efficiency and bad node ratio in %]

VII. Conclusion and Outlook

Smart energy distribution in self-powered micro-scale networks like sensor networks is a challenge with the goal to satisfy a safe and reliable operational state on system and node level. Under the assumption that nodes are arranged in mesh-like networks with links posing the capability to transfer energy and data between nodes a self-organizing MAS was deployed in this work successfully to distribute energy without a system/world level model and knowledge of the single nodes about the system state. Different agent behaviour were investigated and evaluated. The exploratory help strategy with deliver child agents showed the best and efficient overall behaviour.

The agents were programmed in *JavaScript* using the *JavaScript* Agent Machine Platform (*JAM*) that can be deployed in strong heterogeneous environments on a wide range of devices.

Among the agent behaviour already presented in this work, the issue of rising very bad (non operable) node clusters observed in the current MAS framework must be prevented. One possible solution is directed diffusion propagation around nodes giving energy away, i.e., an agent consuming and transferring energy from a node should trigger energy transfer in the opposite delivery direction (away from the energy valley up to energy hills). Furthermore, distributed supervised learning can be used to improve the emergence of the entire network and MAS on system level and on local region level. The used *JAM* platform already supports agents with an extensive set of learning algorithms posing mobile models that can migrate with agents.

Parameter Sets

```
B={explorationRange:2, maxLife:1, maxHops:4,
 inhibitTime: 20}
C={explorationRange:2, maxLife:2, maxHops:4,
 inhibitTime: 20}
D={explorationRange:2, maxLife:4, maxHops:4,
 inhibitTime: 20}
E={explorationRange:2, maxLife:8, maxHops:4,
 inhibitTime: 20}
F={explorationRange:4, maxLife:4, maxHops:8,
 inhibitTime: 20}
```

Agent Behaviour

The following algorithms describe parts of the agent behaviour in *AAPL* short notation (details [15]) of node and energy agents and their interaction using the tuple space. The stationary node agent controls energy storage, harvesting, and transfer on network nodes (with direct access to energy devices). The mobile energy agent performs energy negotiation and transport.

Notation: Ψ: Agent class, φ: Subclass, Σ: Agent body variables, α: Agent activity, Θ: Agent creation/destruction (+:create, $\times$:destroy, $\rightarrow$:fork), ∇: Tuple space access (+:out, -:inp, %:rd, $\pm$:listen, $\mp$:evaluate), $\Leftrightarrow$: Agent migration, π: activity transitions. Tuple listener receive tuples (with actual and formal parameters) from a corresponing evaluate operation and pass modifieded tuples to the evaluation request.

Ψnode : *options* $\rightarrow$ {
Σ = { *energy*, $energy_{thres*}$, $energy_{Deposit}$, .. }
α*init* : {
• *Agent asks for energy demand* (+) *or grant* (−)
$\nabla^{\pm}(ENERGY?, ?) \rightarrow (*, ask)$ {
$energy < energy_{Thres_0} + energy_{Deposit}$?
• *energy demand*
$ask \leftarrow energy_{Thres_1} - energy$:
• *energy grant*
$ask \leftarrow -(energy - energy_{Thres_0} - energy_{Deposit})$
}
• *Agent requests energy token*
$\nabla^{\pm}(ENERGY-, ?) \rightarrow (*, con)$ {
$energy > con - energy_{Deposit}$?
$energy \leftarrow energy - con$,
$consumed \leftarrow consumend + con$,
• *conversion loss*
$con \leftarrow con * energy_K$:
$con \leftarrow 0$
}
• *Agent delivers energy token*
$\nabla^{\pm}(ENERGY+, ?) \rightarrow (*, del)$ {
• *conversion loss*
$del \leftarrow del * energy_K$
$energy \leftarrow energy + del$
}
• *Agent delivers energy token after migration*
$\nabla^{\pm}(ENERGYC+, ?) \rightarrow (*, del)$ {
• *conversion loss*
$del \leftarrow del * energy_{CommK}$
$energy \leftarrow energy + del$
}
• *Generic energy request*
$\nabla^{\pm}(REQUEST, ?) \rightarrow (*, req)$ {
• *Is this node good, very good, or bad?*
$energy < energy_{Thres_0}$?
• *bad node, needs energy*
$req \leftarrow 0$:
$energy < energy_{Thres_1}$?
• *goog node : grant only half energy request!*
$req \leftarrow req / 2$:
• *very good node : grant full request*
}
}
}

Energy negotiation is peformed by different tuples (energy tokens): *ENERGY?*, *ENERGY-*, *ENERGY+*, *ENERGYC+*, *REQUEST*. Energy agents can replicate based on behaviour.

Ψenergy : *options* $\rightarrow$ {
Σ = { *state*, *charge*, *energy*, *age*, *hops*, Δ, *age*, .. }
α*init* : {
state $\leftarrow$ *SEARCHING*
charge? $\nabla^{\mp}(ENERGY-, charge) \rightarrow$
$(*, *)\{charge \leftarrow 0\}$
}

route () $\rightarrow$ {
range? {
• *Random walk behaviour*
$dir \leftarrow random([NORTH, SOUTH, WEST, ..])$
..
} : {
• *Simple* Δ*routing*
$\Delta_x > 0 \wedge ?\Lambda(EAST)$? Δ_x - -, $\rightarrow EAST$
$\Delta_x < 0 \wedge ?\Lambda(WEST)$? Δ_x ++, $\rightarrow WEST$
$\Delta_y > 0 \wedge ?\Lambda(SOUTH)$? Δ_y - -, $\rightarrow SOUTH$
..
}
}

α*travel* : {
nextdir $\leftarrow$ *route*()
$\neg$ *nextdir*? $\Theta^{\times}(self)$
charge? $\nabla^{\mp}(ENERGY-, charge) \rightarrow$
$(*, req)$ { $energy = req, charge = 0$ }
hops ++, *lastdir* $\leftarrow$ *nextdir*, $\Leftrightarrow$ (*nextdir*)
}

α*arrived* : {
$energy > 0$? $\nabla^{\mp}(ENERGYC+, energy) \rightarrow$
$(*, *)$ {*charge* $\leftarrow$ *energy*, *energy* $\leftarrow 0$}
$\Delta = 0$? *state* $\rightarrow$ *IAMHERE*
}

α*terminate* : { .. }
α*wait* : { .. }

φrequest : {
α*percept* : {
$\Delta = 0 \wedge$ *request*? $\nabla^{\mp}(REQUEST, request) \rightarrow$
$(*, x)$ {*charge* $\leftarrow x$}
}
α*action* : {
charge? $\Theta^{\rightarrow}$(*type* : *REPLY*, $\Delta : -\Delta$,
charge : *charge*, *state* : *SEARCHING*),
• *Stay only if the requested charge was granted*
charge $* 1.1 >$ *request* ? *age* - - : *age* $\leftarrow 0$
charge $\leftarrow 0$
}

π : {
percept $\rightarrow \Delta = 0$? *action* : *travel*
action $\rightarrow$ *age* > 0 ? *wait* : *terminate*
}
}

φreply : {
α*percept* : { $\Delta = 0$? *state* $\rightarrow$ *IMAHERE* }
α*action* : { }
π : {
percept $\rightarrow$ *state* = *SEARCHING* ? *travel* : *terminate*
action $\rightarrow$ *percept*
}
}

φdeliver : {
  $\alpha percept$: {
    • $Help - on - way\ behaviour?$
    $helponway?\ \nabla^{\mp}(ENERGY?, *) \rightarrow$
     $(*, req)\{\ req > 0\ ?\ charge\ \leftarrow$
          $charge - charge/(range + 2)\ \}$
    $\Delta = 0?\ state\ \rightarrow IAMHERE$
  }
  $\alpha action$: { .. }
  π : {
    $percept\ \rightarrow state = SEARCHING\ ?\ travel\ :\ terminate$
    $action\ \rightarrow percept$
  }
}

φhelp : {
  $\alpha percept$: {
    $charge\ \rightarrow 0,\ state\ \rightarrow SEARCHING$
    $\Delta \neq 0\ ?\ \nabla^{\mp}(ENERGY?, *)\ \rightarrow (*, req)$
     $\{\ req < 0 \wedge -req > request?state\ \rightarrow IAMHERE\ \}$
  }
  $\alpha action$: {
    $state = IAMHERE?$
     $\Theta^{\rightarrow}($
       $type : DELIVER,\ state : SEARCHING,$
       $charge : request,\ energy : 0,$
       $delta : -\Delta range : 0$
)
    age - -
  }

  π : {
    $percept\ \rightarrow hops > maxhops\ ?\ terminate\ :$
          $(state = SEARCHING\ ?\ travel\ :\ action)$
    $action\ \rightarrow age > 0\ ?\ wait\ :\ terminate$
  }
}

φdistribute : {
  $\alpha percept$: {
    $charge > 0\ ?\ \nabla^{\mp}(ENERGY?, *) \rightarrow$
     $(*, req)\ \{$
       $req\ >\ charge\ ?$
         • $Deliver\ all\ charge\ on\ this\ node$
         $charge\ \rightarrow 0,\ state\ \rightarrow IAMHERE\ :$
         • $Deliver\ charge\ requested\ from\ node$
         $req\ >\ 0\ ?\ charge\ \rightarrow charge\ -\ req$
     }
  }

  $\alpha action$: { }
  π : {
    $percept\ \rightarrow action$
    $action\ \rightarrow state = SEARCHING\ \wedge hops < maxhops\ ?$
        $travel\ :\ terminate$
  }
 }
}

REFERENCES

[1] J. Lagorse, D. Paire, A. Miraoui, *A multi-agent system for energy management of distributed power sources* , J. of Renewable Energy, Vol. 35, Issue 1, 2010

[2] S. Ghani, M. Mousavi, and A. Movaghar, *Distributed Algorithms for Power Saving Optimization in Sensor Network* , Proceedings of the 8th WSEAS international conference on Data networks communications computers, pp. 109 −115, 2009.

[3] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, *Power management in energy harvesting sensor networks* , ACM Transactions on Embedded Computing Systems, vol. 6, no. 4, p. 32-es, 2007.

[4] Z. Jun, L. Junfeng, W. Jie, and H. W. Ngan, *A multi-agent solution to energy management in hybrid renewable energy generation system* , Renewable Energy 36, vol. 36, pp. 1352-1363, 2011.

[5] G.M. P. O'Hare, D. Marsh, A. Ruzzelli, and R. Tynan, Agents forWireless Sensor Network Power Management, in Parallel Processing, 2005. ICPP 2005 Workshops. International Conference Workshops on, 2005.

[6] P. Leito and S. Karnouskos, *Industrial Agents Emerging Applications of Software Agents in Industry*. Elsevier, 2015.

[7] S. Bosse, F. Kirchner, *Smart Energy Management and Energy Distribution in Decentralized Self-Powered Sensor Networks Using Artificial Intelligence Concepts* , Proceedings of the Smart Systems Integration Conference 2012, Session 4, Zürich, Schweiz, 21 − 22 Mar. 2012, 2012, ISBN: 978-3-8007-3423-8.

[8] E. Rokrok, M. Shaekhah, P. Siano, and J. P. S. Catalao, *A Decentralized Multi-Agent-Based Approach for Low Voltage Microgrid Restoration* , Energies, vol. 10, no. 1491, 2017.

[9] B. Zhao, M. Xue, X. Zhang, C. Wang, and J. Zhao, *An MAS based energy management system for a stand-alone microgrid at high altitude* , Applied Energy, vol. 143, 2015.

[10] S. Bosse, A. Lechleiter,*Structural Health and Load Monitoring with Material-embedded Sensor Networks and Self-organizing Multi-agent Systems* , Procedia Technology, 2014, http://dx.doi.org/10.1016/j.protcy.2014.09.039

[11] E. Gavrin, S.J. Lee, R. Ayrapetyan, R., A.Shitov, (2015) *Ultra lightweight JavaScript engine for internet of things* , in SPLASH Companion 2015 Companion Proceedings of the 2015 ACM SIGPLAN International Conference on Systems, Programming, Languages and Applications: Software for Humanity, 2015, pp. 19-20

[12] L. Chunlina. , L. Zhengdinga., L. Layuanb. , Z. Shuzhia,. (2002) *A mobile agent platform based on tuple space coordination,* Advances in Engineering Software, vol. 33, no. 4, pp. 215−225.

[13] R.H. Bordini., J.F: Hübner. (2006) *BDI agent programming in AgentSpeak using Jason* , Computational Logic in Multi-Agent Systems, Volume 3900 of the series Lecture Notes in Computer Science, Springer,, pp. 143-164.

[14] S. Bosse, *Mobile Multi-Agent Systems for the Internet-of-Things and Clouds using the JavaScript Agent Machine Platform and Machine Learning as a Service*, in The IEEE 4th International Conference on Future Internet of Things and Cloud, 22-24 August 2016, Vienna, Austria, 2016, http://dx.doi.org/10.1109/FiCloud.2016.43.

[15] S. Bosse, A. Lechleiter, *A hybrid approach for Structural Monitoring with self-organizing multi-agent systems and inverse numerical methods in material-embedded sensor networks*, Mechatronics, (2016), http://dx.doi.org/10.1016/j.mechatronics.2015.08.005

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 215–223
DOI: 10.15439/2018F159
ISSN 2300-5963 ACSIS, Vol. 15

Sensitivity in Multi-Ensemble Scheduling

Jörg Bremer
Department of Computing Science
Carl von Ossietzky University
Oldenburg, Germany
joerg.bremer@uni-oldenburg.de

Sebastian Lehnhoff
Department of Computing Science
Carl von Ossietzky University
Oldenburg, Germany
sebastian.lehnhoff@uni-oldenburg.de

***Abstract*—Future smart grid control demands delegation of liabilities to distributed, rather small energy resources in contrast to today's traditional large control power units. Distributed energy scheduling constitutes a complex task for optimization algorithms regarding the underlying high-dimensional, multimodal and non-linear problem structure. For predictive scheduling with high penetration of renewable energy resources, agent-based approaches using classifier-based decoders for modeling individual flexibilities have shown good performance. On the other hand, such decoder-based methods are currently designed for single entities and not able to cope with ensembles of energy resources. Aggregating training sets sampled from individually modeled energy units results in folded distributions with unfavorable properties for training a decoder. Nevertheless, this happens to be a quite frequent use case, e. g. when a hotel, a small business, a school or similar with an ensemble of co-generation, heat pump, solar power, and controllable consumers wants to take part in decentralized predictive scheduling. Recently, an extension to an established agent approach for scheduling individual single energy units has been proposed that is based on second level optimization. The agents' decision routine may be enhanced by a covariance matrix adaption evolution strategy that is hybridized with decoders. In this way, locally managed ensembles of energy units can be included. The applicability has already been demonstrated, but the effects of ensemble composition are so far unknown. Here, we give an widened view on the underlying power level distribution problem and extend the results by conducting a sensitivity analysis on the impact of ensemble size and penetration on communication overhead and residual error.**

I. Introduction

In Germany where a financial security of guaranteed feed-in prices has meanwhile been granted since the early 90th – but also in other countries of the European union and world wide, the share of distributed energy resources (DER) within the electricity grid is constantly and rapidly rising. According to the goal defined by the European Commission [1], concepts for integration into electricity markets will quickly become indispensable to reduce subsidy dependence for both: active power provision and for providing ancillary services like frequency or voltage control [2], [3].

Consequently, combining smart measurement technologies for decentralized information gathering on current operational grid state, new tele-control techniques, communication standards and scalable, decentralized self-organized control schemes will lead to a so called smart grid with decentralized power conditioning and control of the production and distribution of electricity managed without central control; as in the vision of [4] or similar for Europe [5].

As the smart grid will have to delegate many control tasks to small and distributed energy units, new control algorithms are required that are able to cope with large problem sizes and distributed and only locally available information. Virtual power plants (VPP) are a well-known instrument for aggregating and controlling DER [6]. Concepts for several purposes (commercial as well as technical) have been developed. A usual use case commonly emerging within VPP control is the need for scheduling the operation of participating DER. Predictive scheduling [7] describes the optimization problem for day-ahead planning of energy generation in VPPs, where the goal is to select a schedule for each energy unit – from an individual search space of feasible schedules with respect to a future planning horizon – such that a global objective function (e. g. resembling a target power profile for the VPP as close as possible) is optimized.

Recently, distributed approaches gained more and more importance for VPP control. Different works proposed hierarchical and decentralized architectures based on multi-agent systems and market-based computing [8], [9]. Newer approaches try to establish self-organization between actors within the grid [10]–[12]. In contrast, today's commercial VPP are often operated by a single authority that at the same time is the owner of (and responsible for) all distributed energy resources in this rather static unit ensemble. Independently from a concrete implementation for predictive scheduling, the dispatch algorithm has to choose a schedule for each DER in the VPP such that all objectives are met.

In order to choose an appropriate schedule for each participating DER, the algorithm must know from each DER, which schedules are actually operable and which are not. Depending on the type of DER, different constraints restrict possible operations. The information about individual local feasibility of schedules has to be modeled appropriately in (distributed) optimization scenarios, in order to allow unit independent algorithm development. For this purpose, meta-models of constrained spaces of operable schedules have been shown indispensable as a means for independently modeling constraints and feasible regions of flexibility. Each energy unit has its own individual flexibility – i. e. the set of schedules that might be operated without violating any technical operational constraint – based on the capabilities of the unit,

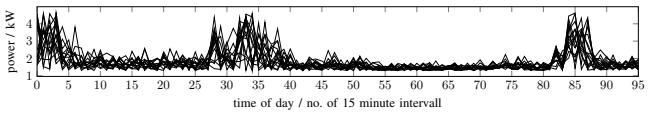


Fig. 1. Example for a training set of schedules for a co-generation plant. A state-of-charge of 50% at night and an increased thermal demand for showering in the morning and dish washing in the evening result in higher flexibilities during these periods.

operation conditions (weather, etc.), cost restrictions and so forth. Integrating these constraints to possible operations of an arbitrary energy unit demands a means for meta-modeling that allows model independent access to feasibility information. [13] introduced a support vector based model that captures individual feasible regions from training sets of operable example schedules. Figures 1 and 2 show example training sets for a co-generation plant and a heat pump respectively.

With an appropriate extension – so called decoders [14] – , these models can also be used for repairing infeasible solution or for systematically generating feasible solutions [15]. Agent-based approaches can derive a so called support vector decoder automatically from the surrogate model and use it as a means for generating feasible solutions without domain knowledge on the (possible, situational) operations of the controlled energy resource [14].

Examples for using decoders in optimization within the smart grid can be found in [16]–[19]. In general, the idea works in two successive stages – a decoder training phase and the actual algorithm/ negotiation execution phase where the decoder is used [7]. During the training phase a decoder is calculated for each unit. These calculations can be done fully parallel. During the succeeding load planning phase, these decoders may be used by an optimization algorithm that determines the optimal partition of a given active power target schedule into schedules for each single unit. The decoder automatically repairs infeasible solutions and thus the solver does not need any domain knowledge about the energy units, their individual constraints, or possible operation.

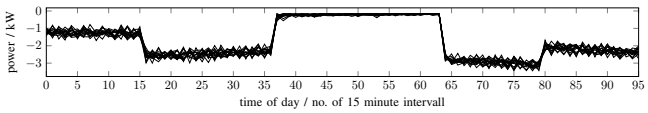


Fig. 2. Example for a training set of schedules for a heat pump with a maximum deviation of 500 Wh from the integral of set thermal demand.

An example for a recently developed agent approach for fully decentralized predictive scheduling is given by the **co**mbinatorial **h**euristics for **d**istributed **a**gents (COHDA). In COHDA [20] each agent is responsible for exactly one energy unit and uses a decoder to locally decide on feasible schedules for the represented unit. The algorithm has shown excellent performance [17], [20], [21]. But, as soon as an agent has to represent a local ensemble of energy units instead of a single device, a problem arises because usually only flexibility models of single units are available. Generating a single decoder for handling all constraints and feasible operations of the whole ensemble is hardly possible due to statistical problems when combining training sets from individually sampled flexibility models. Due to the folded densities only a very small portion from the interior of the feasible region (the dense region) is captured by the machine learning process. But, a combined training set is needed if one wants to train a single decoder for each agent.

For this reason, in [22] a substitute for the single decoder part that generates suitable and feasible schedules for the negotiating agents has been proposed. To achieve this, an evolution strategy is harnessed to do the job of solving the problem using individual decoders (one for each unit in the ensemble). In this way, an optimization problem has to be solved instead of a single mapping with a decoder for each agent decision during the negotiation, but with harnessing the full flexibility of the ensemble. Hence, a new decision method is introduced to the agent approach based on a covariance matrix adaption evolution strategy that widens the applicability to including multiple local ensembles of DER into the VPP without changing the underlying negotiation between the agents.

In [22] the general approach had been scrutinized on basis of gained optimization results regarding effect and performance of the CMA-ES part. A closer look on underlying mechanisms of the agent negotiation and parameter impact is missing so far. Moreover, the influence of ensemble size and composition is unknown. Here, we extend the former work with a study on the impact of folded distributions on different energy units' aggregated flexibility – constituted by the agents' entities – and conduct a sensitivity analysis regarding the impact of group size, composition or group number on the agent approach as well as on the CMA-ES intermediate results.

The rest of the paper is organized as follows. First, an outline on predictive scheduling and related work regarding the decoder approach as well as decentralized, agent-based methods for solving is presented. A strong focus is on the combinatorial heuristics for distributed agents. After scrutinizing the problem of folded power level distributions in aggregated training sets for different types of energy unit ensembles, the necessity of integrating a heuristic approach into the agent method for ensembles is derived. We recap the hybridization of covariance matrix adaption evolutions strategies with support vector decoders and the bi-level approach from [22] for circumventing the problem of folded distributions. The sensitivity of different group traits is analyzed. We conclude with results from several simulation studies showing beyond the effectiveness of the hybrid approach the scalability regarding ensemble size, penetration and communication expenses. Moreover, it is shown that the overall efficiency of the underlying agent approach is not seriously effected by integrating a sub-optimization process into the decision phase.

II. Related Work

A. Predictive scheduling

As related work, solutions to predictive scheduling with decoders have to be discussed in the context of agent-based approaches prior to deriving the root cause that raises the problem when extending scheduling to participants that locally have to control more than one single energy unit. We start with a definition of the general predictive scheduling problem.

As opposed to the usual time series model, we regard a schedule as real valued vector $\boldsymbol{p} = (p_1, \ldots, p_d) \in \mathbb{R}^d$ with each element p_j denoting mean active power generated (positive values) or consumed (negative values) during the j-th of d time intervals. Starting time and width of each time interval are assumed to be known from context information. The feasibility of a schedule $\boldsymbol{p}$ is defined by sets of unit specific technical and economic constraints.

One of the crucial challenges in operating a VPP arises from the complexity of the scheduling task due to the large amount of (small) energy units in the distribution grid [23]. In the following, we consider predictive scheduling, where the goal is to select exactly one schedule $\boldsymbol{p}_i$ for each energy unit U_i from a search space $\mathcal{F}^{(U)}$ of feasible schedules specific to the possible operations and technical constraints of unit U and with respect to a future planning horizon, such that a global objective function (e. g. resembling a target power profile) is optimized by the sum of individual contributions [24]. A basic formulation of the scheduling problem is given by

$$\delta\left(\sum_{i=1}^{m} \boldsymbol{p}_i, \boldsymbol{\zeta}\right) \rightarrow \min;\ \text{s.t.}\quad \boldsymbol{p}_i \in \mathcal{F}^{(U_i)}\ \forall U_i \in \mathcal{U}. \tag{1}$$

In equation (1) δ denotes an (in general) arbitrary distance measure for evaluating the difference between the aggregated schedule of the group and the desired target schedule $\boldsymbol{\zeta}$. W. l. o. g. we assume the Euclidean distance is used.

To each energy unit U_i exactly one schedule $\boldsymbol{p}_i$ has to be assigned. The desired target schedule is given by $\boldsymbol{\zeta}$. $\mathcal{F}^{(U_i)}$ denotes the individual set of feasible schedules that are operable for unit U_i without violating any (technical) constraint. Solving this problem without unit independent constraint handling leads to specific implementations that are not suitable for handling changes in VPP composition or unit setup without having changes in the implementation of the scheduling algorithm [17].

Flexibility modelling can be understood as the task of modelling constraints for energy units. For optimization approaches in smart grid scenarios, black-box models capable of abstracting from the intrinsic model have proved useful [25], [26]. They do not need to be known at compile time. A powerful, yet flexible way of constraint-handling is the use of a decoder that gives a search algorithm hints on where in the search space to look for schedules satisfying local hard constraints (*feasible schedules*) [26], [27].

For our experiments, we used a decoder as described in [15]. Here, a decoder γ is given as mapping function

$$\gamma : \mathbb{R}^d \rightarrow \mathbb{R}^d;\ \gamma(\boldsymbol{p}) \mapsto \boldsymbol{p}^*. \tag{2}$$

With $\boldsymbol{p}^*$ having the following properties:

- $\boldsymbol{p}^*$ can be operated by the respective energy unit without violating any constraint,
- the distance $\|\boldsymbol{p} - \boldsymbol{p}^*\|$ is small; where the term small depends on the problem at hand and often denotes the smallest distance of $\boldsymbol{p}$ to the feasible region.

With such decoder concept for constraint handling one can now reformulate the optimization problem as

$$\delta\left(\sum_{i=1}^{m} \gamma_i(\boldsymbol{p}_i), \boldsymbol{\zeta}\right) \rightarrow \min, \tag{3}$$

where γ_i is the decoder function of unit i that produces feasible, schedules from $\boldsymbol{p} \in [0, p_{max}]^d$ resulting in schedules that are operable by that unit. Please note, that this is a constraint free formulation. With this problem formulation, many standard algorithms for optimization can be easily adapted as there are no constraints (apart from a simple box constraint $\boldsymbol{p} \in [0, p_{max}]^d$) to be handled and no domain specific implementation (regarding the energy units and their operation schedules) has to be integrated. Equation (3) is used as a surrogate objective to find the solution to the constrained optimization problem equation (1).

B. COHDA

The Combinatorial Optimization Heuristics for Distributed Agents (COHDA) was originally introduced in [28], [29]. Since then it has been applied to a variety of smart grid applications [17], [24], [30], [31]. With our explanations we follow [29].

Originally, COHDA has been designed as a fully distributed solution to the predictive scheduling problem (as distributed constraint optimization formulation) in smart grid management [28]. In this scenario, each agent in the multi-agent system is in charge of controlling exactly one distributed energy resource (generator or controllable consumer) with procuration for negotiating the energy. All energy resources are drawn together to a virtual power plant and the controlling agents form a coalition that has to control the VPP in a distributed way. It is the goal for the predictive scheduling problem to find exactly one schedule for each energy unit such that

1) each assigned schedule can be operated by the respective energy unit without violating any hard technical constraint, and
2) the difference between the sum of all targets and a desired given target schedule is minimized.

The target schedule usually comprises 96 time intervals of 15 minutes each with a given amount of energy (or equivalently mean active power) for each time interval, but might also be constituted for a shorter time frame by a given energy product that the coalition has to deliver.

An agent in COHDA does not represent a complete solution as it is the case for instance in population-based approaches [32], [33]. Each agent represents a class within a multiple choice knapsack combinatorial problem [34]. Applied to predictive scheduling each class refers to the feasible region in

the solution space of the respective energy unit. Each agent chooses schedules as solution candidate only from the set of feasible schedules that belongs to the DER controlled by this agent. Each agent is connected with a rather small subset of other agents from the multi-agent system and may only communicate with agents from this limited neighborhood. The neighborhood (communication network) is defined by a small world graph [35]. As long as this graph is at least simply connected, each agent collects information from the direct neighborhood and as each received message also contains (not necessarily up-to-date) information from the transitive neighborhood, each agent may accumulate information about the choices of other agents and thus gains his own local belief of the aggregated schedule that the other agents are going to operate. With this belief, each agent may choose a schedule for the own controlled energy unit in a way that the coalition is put forward best while at the same time own constraints are obeyed and own interests are pursued.

All choices for own schedules are rooted in incomplete knowledge and beliefs in what other agents are probably going to do; gathered from received messages. The taken own choice (together with the basis for decision-making) is communicated to all neighbors and in this way knowledge is successively spread throughout the coalition without any central memory. This process is repeated. Because all spread information about schedule choices is labeled with an age, each agent may decide easily whether the own knowledge repository has to be updated. Any update results in recalculating of the own best schedule contribution and spreading it to the direct neighbors. By and by all agents accumulate complete information and as soon as no agent is capable of offering a schedule that results in a better solution, the algorithm converges and terminates. Convergence has been proven in [20].

More formally, each time an agent receives a message, three successive steps are conducted. First, during the perceive phase an agent a_j updates its own working memory κ_j with the received working memory κ_i from agent a_i. From the foreign working memory the objective of the optimization (i. e. the target schedule) is imported (if not already known) as well as the configuration that constitutes the calculation base of a neighboring agent a_i. An update is conducted if the received configuration is larger or has achieved a better objective value. In this way, schedules that reflect the so far best choices of other agents and that are not already known in the own working memory are imported from the received memory.

During the following decision phase agent a_j has to decide on the best choice for his own schedule based on the updated belief about the system state Γ_k. Index k indicates the age of the system state information. The agent knows which schedules of a subset of other agents (or all) are going to operate. Thus, the schedule that fills the gap to the desired target schedule exactly can be easily identified. Due to operational constraints of the controlled DER, this optimal schedule can usually not be operated. Thus, each agent is equipped with a decoder that automatically maps the identified optimal schedule to a nearby feasible schedule that is operable by the DER and thus feasible. In this way, the decision routine of the agent reduces simply to a mapping call of the decoder. Based on a set of feasible schedules sampled from an appropriate simulation model for flexibility prediction [36], the decoder can be built by learning a support vector model after the approach of [15].

If the objective value for the configuration with this new candidate is better, this new solution candidate is kept as selected one. Finally, if a new solution candidate has been found, the working memory with this new configuration is sent to all agents in the local neighborhood. The procedure terminates, as soon as all agents reach the same system state and no new messages are generated. In this case no agent is able to find a better solution. Finally, all agents know the same final result.

As the whole procedure is based exclusively on local decisions, each agent decides privately which schedules are taken. Private interest and preferences can be included and all information on the flexibility of the local DER is kept private. The same must hold true for agents controlling an ensemble of energy units.

III. Ensemble scheduling

A. Problem

Sometimes the technical equipment of a single participant in a virtual power plant consists of more than just a single generator (or prosumer or controllable consumption). Nevertheless, the owner as operator is usually still represented by a single controlling agent when embedded into a decentralized agent-based control scheme inside a virtual power plant. In this case that agent has to handle the ensemble of energy units as a single unit (in a sense as a single sub VPP) and negotiate to the other agents with the aggregated flexibility. Nevertheless, there is usually no joint model of the whole ensemble, and thus the agent has to use an individual model of each unit and thus a set of individual decoders for deciding on an aggregated schedule for the ensemble.

If an agent covers a set of energy units instead of a single unit, a decoder for the joint feasible region of the group of units has to be used. A model of the operation of the ensemble of units is usually not available. Using the training sets of individual energy units and randomly combining them (adding up exactly one from each training set) to joint schedules in order to gain a training set for the joint behavior is not targeted. The problem is that all source trainings sets are independent random samples and thus the resulting training set exhibits a density (of operable power levels) that results from folding the source distributions.

Figure 3(a) shows a first example. Rather uniformly – except for the gap between zero and minimum engine velocity and a slightly degrading likelihood of higher power levels – distributed values for levels of power as in the case of an co-generation plant with sufficient buffer capacity fold up to an multi-modal Irvin-Hall-distribution [37]. This distribution has some similarities to a sharp normal distribution and the more

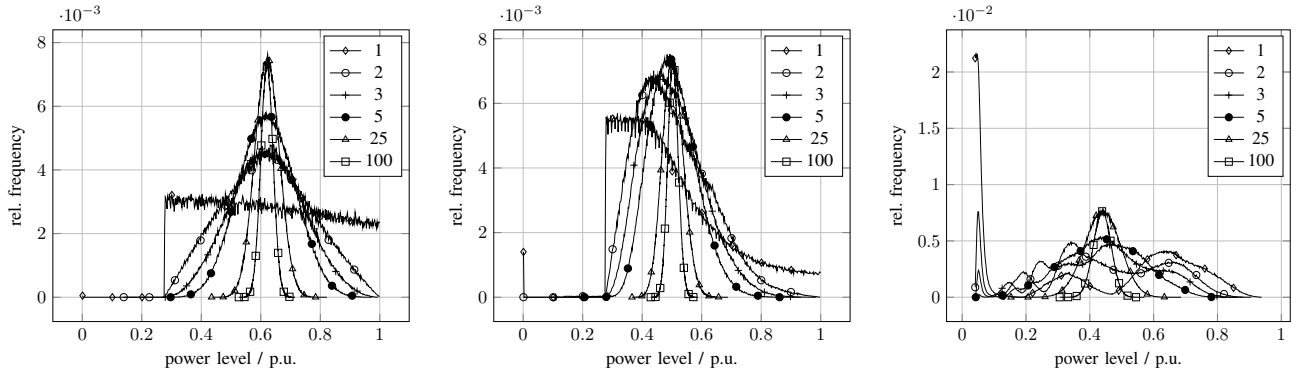


Fig. 3. Probability density of different numbers of folded distributions of operable power levels for co-generation plants for a very cold winters day.

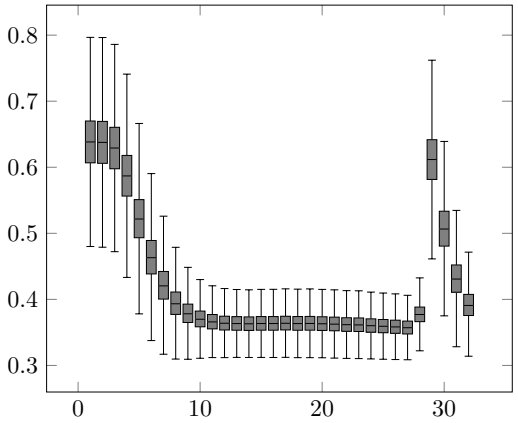


Fig. 4. Probability distribution of power levels at different time intervals of an ensemble of 10 micro CHP units. The training set exhibits a concentration in the inner part of the whole flexibility (grey boxes denoting 3/4 of the samples) making it highly imbalanced.

samples (number of energy units in the ensemble) are folded the more leptokurtic the pdf gets.

Whereas Figure 3(a) considers the distribution of power levels at a certain point in time (7:00 a.m.), shows Figure 3(b) the situation averaged over all time periods of a sample winters day. Due to the integration of the warmer daylight periods, the likelihood of high power levels degrades. Nevertheless, in the case of ensembles of co-generation plants the distributions fold up to a similar aggregated distribution that generates training sets unsuitable for machine learning. If a model is based on an estimation of probability distributions, then it is highly susceptible to the spatial distribution of the samples in feasible space [38] because merely regions of high density are learned. Figure 3(c) shows as a third example the folded distributions in the case of a heat pump.

This folding leads to a sample with a very high density in the middle of the feasible region. At the outskirts the sample is extremely sparse. Thus, almost all instances from the outer parts are neglected as outliers from the support vector approach that generates the surrogate model and the decoder.

For this reason, a decoder trained from such a training sample reproduces only a very small, inner portion of the feasible region. In this way, most of the flexibility that an ensemble could bring in into virtual power plant control is neglected. This can also be seen in Figure 4. The rather small grey boxes represent the data (power levels for different time intervals) that actually should spread over the area denoted by the outer whiskers. Only the small inner part is going to be learned by a model.

B. CMAES with decoder

The covariance matrix adaption evolution strategy [39], [40] (CMA-ES) is a well known evolution strategy for solving multi modal black box problems.

CMA-ES improves its operations by harnessing lessons learned from previously successful evolution steps for future search directions. A new population of solution candidates is sampled from a multi variate normal distribution $\mathcal{N}(0, \boldsymbol{C})$ with covariance matrix $\boldsymbol{C}$ which is adapted such that it that it maximizes the occurrence of improving steps according to previously seen distributions for good steps. Sampling offspring is weighted by a selection of solutions of the parent generation. In a way, the method learns a second order model of the objective function and exploits it for structure information and for reducing calls of objective evaluations. An a priori parametrization with structure knowledge of the problem by the user is not necessary as the method is capable of adapting unsupervised. A good introduction can for example be found in [41]. Especially for non-linear, non-convex black-box problems, the approach has demonstraded excellent performance [41]. CMA-ES is initially not designed for integrated constraint handling in constrained optimization. Nevertheless, some approaches for integrating constraint handling have been developed. In [42] a CMA-ES is introduced that learns constraint function models and rotates mutation distributions accordingly. In [43] an approximation of the directions of the local normal vectors of the constraint boundaries is built by accumulating steps that violate the respective constraints. Then, the variances of these directions are reduced for mutation.

CMA-ES is used for solving the internal optimization problem that arises when an agent has to decide on the best

possible joint schedule to offer during the decision phase of the COHDA negotiation for virtual power plants. With our explanations we follow [22].

We consider an agent negotiation with a stage where each agent has to search the individual flexibility and thus the individual feasible region of operable schedules for the best option (according to given objectives). In case the agent has to control an ensemble with more the one local unit, a decoder that models the feasible region cannot be used as in the case of a single unit. For this reason, a local optimization problem has to be solved in order to decide on a schedule: find the closest aggregated schedule that the local ensemble can operate. This is basically the same problem as for predictive scheduling Eq. 1. As this smaller sub-problem happens to be a local one seen from the agent's perspective, there is no need to harness a distributed solving strategy. Additionally, the problem size is expected to stay rather small with a limited number of devices e. g. inside a household.

Because the operation of several decoders that model the different feasible regions of the local ensemble has to be involved, a heuristic that uses only a small number of objective evaluations is advantageous. CMA-ES is well known for this characteristic [41]. For handling the constraints, the readily available decoders can be used. Thus, the decoder technique also adapted to and employed to the CMA-ES part (cf. . [22]).

In each iteration g of CMA-ES a multivariate distribution is sampled in order to generate a new offspring solution population in the vicinity of good parent solutions:

$$\boldsymbol{x}_k^{(g+1)} \sim \boldsymbol{m}^{(g)} + \sigma^{(g)} \mathcal{N}(0, \boldsymbol{C}^{(g)}),\ k = 1, \ldots, \lambda. \quad (4)$$

$\boldsymbol{C}^{(g)} \in \mathbb{R}^{n \times n}$ constitutes the covariance matrix of the search distribution at generation (iteration) g with overall standard deviation $\sigma^{(g)}$ which can also be interpreted in terms of an adaptive (multivariate) step size. The step size is adapted individually for each dimension to support and favor direction where fast improvement can be expected according to formerly seen results. The mean of the multivariate distribution is denoted by $\boldsymbol{m}^{(g)}$, $\lambda \geq 2$ denotes the population size.

The new mean $\boldsymbol{m}^{(g+1)}$ for generating the sample of the next generation in CMA-ES is calculated as weighted average

$$\boldsymbol{m}^{(g+1)} = \sum_{i=1}^{\mu} w_i \boldsymbol{x}_{i:\lambda}^{(g+1)},\ \sum w_i = 1,\ w_i > 0, \quad (5)$$

of the best (in terms of objective function evaluation) individuals form the current sample $\boldsymbol{x}_i^{(g)}, \ldots, \boldsymbol{x}_\lambda^{(g)}$. In order to introduce the decoder into CMA-ES, ranking is now done with the help of the decoder mapping γ:

$$f(\gamma(\boldsymbol{x}_{1:\lambda}^{(g)})), \ldots, f(\gamma(\boldsymbol{x}_{\lambda:\lambda}^{(g)})),\ \lambda \geq \mu, \quad (6)$$

to define $\boldsymbol{x}_{i:\lambda}^{(g)}$ as the ith ranked best individual.

For the case of the ensemble scheduling example, a solution candidate $\boldsymbol{x}$ is the concatenation of individual schedules

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{p}_1 \boldsymbol{p}_2 \ldots \boldsymbol{p}_m \\ &= (p_{11}, p_{12}, \ldots, p_{1d}, p_{21}, \ldots, p_{2d}, \ldots, p_{md}) \end{aligned} \quad (7)$$

with $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m$ denoting schedules for the respective units in the ensemble.

Finally, the covariance matrix is updated as usual, but also based on the decoder based ranking Eq. 6:

$$\boldsymbol{C}_\mu^{(g+1)} = \sum_{i=1}^{\mu} w_i \left(\boldsymbol{x}_{i:\lambda}^{(g+1)} - m^{(g)} \right) \left(\boldsymbol{x}_{i:\lambda}^{(g+1)} - m^{(g)} \right)^\top. \quad (8)$$

CMA-ES has a set of parameters that can be tweaked to some degree for a problem specific adaption. Nevertheless, default values that are applicable for a wide range of problems are usually available. For our experiments, we used the following default settings for the CMA-ES part. The (external) strategy parameters are $\lambda, \mu, w_{i=1\ldots\mu}$, controlling selection and recombination; c_σ and d_σ for step size control and c_c and μ_{cov} controlling the covariance matrix adaption. We have chosen to set these values after [41]:

$$\lambda = 4 + \lfloor 3 \ln n \rfloor, \quad \mu = \left\lfloor \frac{\lambda}{2} \right\rfloor, \quad (9)$$

$$\boldsymbol{w}_i = \frac{\ln(\frac{\lambda}{2} + 0.5) - \ln i}{\sum_\mu^{j=1} \frac{\lambda}{2} + 0.5) - \ln i},\ i = 1, \ldots, \mu \quad (10)$$

$$C_c = \frac{4}{n+4},\ \mu_{\text{cov}} = \mu_{eff}, \quad (11)$$

$$\begin{aligned} C_{\text{cov}} = &\frac{1}{\mu_{\text{cov}}} \frac{2}{(n+\sqrt{2})^2} \\ &+ \left(1 - \frac{1}{\mu_{\text{cov}}}\right) \min\left(1, \frac{2\mu_{\text{cov}} - 1}{(n+2)^2 + \mu_{\text{cov}}}\right). \end{aligned} \quad (12)$$

An in-depth discussion of these parameters is also given in [44]. These settings are specific to the dimension N of the objective function. In our case is $N = d \cdot m$ related to the number of agents and the dimension of the assigned schedules in the test cases that are discussed in the following section.

IV. Results

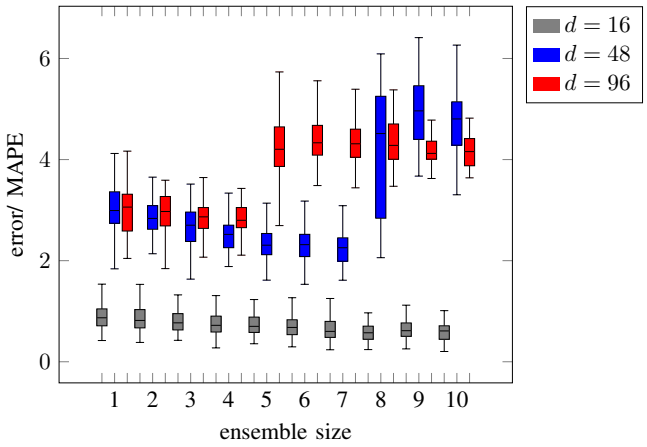


Fig. 5. Sensitivity of the approach to the group size (1 denotes no ensemble in the VPP) of an embedded ensemble in the VPP for different planning time horizons d.

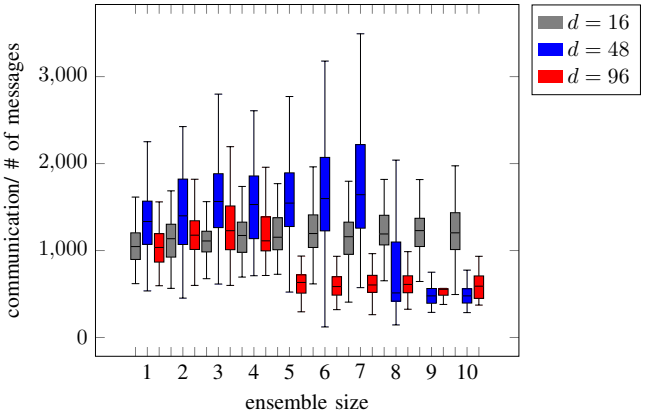


Fig. 6. Impact of the size of an embedded ensemble on the number of exchanges messages and thus on the number of local decisions for different planning horizons d.

TABLE I
SENSITIVITY OF THE APPROACH TO THE NUMBER OF GROUPS (SHARE OF THE WHOLE VPP IN PERCENT) AND IMPACT ON THE COMMUNICATION EFFORT (AND THUS ON LOCAL NUMBER OF DECISIONS) FOR A VPP WITH 10 PARTICIPANTS (SINGLE UNITS AND ENSEMBLES). SCHEDULING HAS BEEN SIMULATED WITH 96-DIMENSIONAL SCHEDULES FOR A PLANNING PERIOD OF A WHOLE DAY.

ensembles/ %	δ_{MAPE}	# messages
0	2.531 ± 1.254	1055.50 ± 244.15
2	2.672 ± 0.626	1271.16 ± 312.46
4	2.502 ± 0.289	1531.94 ± 350.78
6	2.352 ± 0.231	1613.08 ± 389.72
8	2.204 ± 0.210	1629.28 ± 422.94
10	2.102 ± 0.231	1581.00 ± 413.47

Evaluation was again done by simulation with a setup comprising a set of simulated energy resources and a multi-agent system for control [22]. The agent system was implemented after [20]. Each agent is responsible for conducting local decisions and communication with other agents in charge of controlling a small local ensemble of jointly controlled energy resources. Each agent is equipped with the described CMA-ES approach for local decisions on operation.

As a model for distributed energy resources we used a model for co-generation plants that has already served in several studies and projects for evaluation [15], [30], [31], [45], [46]. This model comprises a micro CHP with 4.7 kW of rated electrical power (12.6 kW thermal power) bundled with a thermal buffer store. Constraints restrict power band, buffer charging, gradients, min. on and off times, and satisfaction of thermal demand. Thermal demand is determined by simulating losses of a detached house (including hot water drawing) according to given weather profiles. For each agent the model is individually (randomly) configured with state of charge, weather condition, temperature range, allowed operation gradients, and similar. From these model instances, the respective training sets for building the decoders have been generated with the sampling approach from [36]. In addition, we used models for heat pumps and boilers for hot water provision [47]. A fourth model simulates the flexibilities of a cool storage.

TABLE II
SENSITIVITY OF THE APPROACH TO THE NUMBER OF GROUPS AND IMPACT ON THE COMMUNICATION EFFORT FOR A VPP WITH 50 PARTICIPANTS.

ensembles/ %	δ_{MAPE}	# messages
0	0.935 ± 0.122	87197.24 ± 14149.94
20	0.900 ± 0.354	121473.82 ± 49077.28
40	0.873 ± 0.337	130408.06 ± 63458.05
60	0.770 ± 0.350	145282.74 ± 86664.75
80	0.704 ± 0.378	156506.48 ± 81141.94
100	0.795 ± 0.500	111301.94 ± 78934.29

TABLE III
SENSITIVITY OF THE APPROACH TO THE NUMBER OF GROUPS AND IMPACT ON THE COMMUNICATION EFFORT FOR A VPP WITH 100 PARTICIPANTS.

ensembles/ %	δ_{MAPE}	# messages
0	0.739 ± 0.556	404597.44 ± 192869.14
20	0.776 ± 0.425	443387.92 ± 419493.10
40	0.609 ± 0.355	488633.46 ± 387336.30
60	0.591 ± 0.347	501521.85 ± 573805.91
80	0.632 ± 0.497	493369.70 ± 487284.04
100	0.987 ± 0.660	303466.90 ± 412400.43

The applicability of the hybridized CMA-ES has already been demonstrated in [22]. The approach is able to achieve optimization results with a residual error less than 1 percent. Often well better results with an absolute error of about 30 W for scenarios with a rated power of 470 kW are achieved. Here, again we used the mean absolute percentage error (MAPE)

$$\delta_{MAPE} = \delta(\boldsymbol{x}, \boldsymbol{\zeta}) = \frac{100}{d} \sum_{i=1}^{d} \left| \frac{\zeta_i - x_i}{\zeta_i} \right|, \qquad (13)$$

in order to be able to compare different scenarios with different number of energy units and different rated power.

We simulated the effect of integrating ensembles instead of single energy units into a VPP on the residual error and on the number of exchanged messages between the agents. As the agent system under research is a gossiping type of agent system [48], each message triggers a local decision that translates into a decoder call in the single unit case but into solving a optimization problem with CMA-ES in the ensemble case. Thus, the number of messages is an important indicator for performance scaling with number of integrated ensembles.

Figure 5 shows a first result. The experiment scrutinized a VPP with 10 participants. One participant is an ensemble. The size of the ensemble has been increased from 1 to 10 (an ensemble of size 1 translates again to a single unit) to evaluate means residual error and number of exchanged messages. The experiment has been conducted for differently large planning horizons. For shorter planning horizons the size of the ensemble has almost no impact. Actually, the residual error decreases a little (due to growing flexibility in the VPP). For longer planning horizons the same effect can be observed up to a size where the error escalates to a higher level. At the same time, the number of exchanged messages decreases (cf. Fig. 6). Obviously, the CMA-ES approach starts suffering from some premature convergence problems when the local problem size (ensemble size time schedule dimension) exceeds a certain size; at least when the standard parametrization is

used. Premature convergence at the second level optimization inside an ensemble leads to similar results in successive optimization attempts with similar schedule configurations in the whole VPP. As no better solution is found, the agent sends no message and the first level optimization at agent level ceases earlier with a sub-optimal solution. Hence, integrating methods to prevent premature convergence in the CMA-ES part could largely improve the whole VPP optimization in case of larger ensemble sizes.

Another experiment scrutinizes the number of ensembles in a VPP. Tables I to III show the result. Now, the share of ensembles (with a fixed size of 3 CHP) in a VPP is varying from 0 to 100 percent and the effect is scrutinized. The result quality increases in most of the cases due to a growing flexibility with a growing number of ensembles. For smaller VPP sizes the communication effort grows with the number of ensembles, for larger VPPs the number of sent messages stays on the same level compared with the case of 100 percent single units.

With these results, one can conclude that the introduction of ensembles does not deteriorate the performance of the agent-based predictive scheduling. Performance shortcomings for larger ensembles seem to be due to premature convergence and should be overcome with future integration of e. g. a better adapted step size control.

V. Conclusion

Using machine learning approaches for flexibility modeling and automatically deriving decoders from these models for efficient and domain knowledge independent implementation of (distributed) optimization methods has proven a useful tool in managing the future smart grid. So far, these models can only be applied to single energy units, because distributions of power levels in the training sets of single units fold up when aggregating them to ensemble training sets. Thus, the training set renders useless for appropriately learning a model for the joint flexibility of a group of energy units.

[22] presented a hybrid approach to overcome the problem of folded densities when training decoders for ensembles of energy resources in predictive scheduling. To achieve this, we embedded a CMA-ES solver in the decision routine of an established agent-based solution.

With this approach also households, hotels, small businesses, schools or similar with an ensemble of co-generation, heat pump, solar power, and controllable consumers can take part in agent-based decentralized predictive scheduling for providing energy services in future smart grid architectures without a need for an (expensive) individual link of each single device in the ensemble. By using a hybrid approach of evolution strategy and support vector based decoder, such ensemble based participants in virtual power plants can easily be represented by a single agent. Moreover, agents with our decision method still implement the same interface as single unit agents and can thus be easily integrated with the standard COHDA protocol. Applicability had already been demonstrated in [22].

Our new simulations showed that CMA-ES is well suitable for being hybridized with a decoder in order to build a system that may operate with arbitrary energy units regardless of individual constraints that restrict feasible operation. CMA-ES performs satisfactorily on reasonable large ensembles. Additional simulations showed that size and number of ensembles within a VPP scale well up to reasonable sizes. Communication does not suffer from an increase in number of exchanged messages. Based on these results the inclusion of secondary, local optimization objectives like cost or preferences are a consequentially next step in future work.

References

[1] European Parliament & Council, "Directive 2009/28/ec of 23 april 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing directives 2001/77/ec and 2003/30/ec."

[2] O. Abarrategui, J. Marti, and A. Gonzalez, "Constructing the active european power grid," in *Proceedings of WCPEE09*, Cairo, 2009.

[3] A. Nieße, S. Lehnhoff, M. Trschel, M. Uslar, C. Wissing, H. J. Appelrath, and M. Sonnenschein, "Market-based self-organized provision of active power and ancillary services: An agent-based approach for smart distribution grids," in *Complexity in Engineering (COMPENG), 2012*, June 2012, pp. 1–5.

[4] K. Vinay Kumar and R. Balakrishna, "Smart grid: Advanced metering infrastructure (ami) & distribution management systems (dms)," *International Journal of Computer Science and Engineering*, vol. 3, no. 11, 2015.

[5] I. Colak, G. Fulli, S. Sagiroglu, M. Yesilbudak, and C.-F. Covrig, "Smart grid projects in europe: Current status, maturity and future scenarios," *Applied Energy*, vol. 152, pp. 58 – 70, 2015.

[6] S. Awerbuch and A. M. Preston, Eds., *The Virtual Utility: Accounting, Technology & Competitive Aspects of the Emerging Industry*, ser. Topics in Regulatory Economics and Policy. Kluwer Academic Publishers, 1997, vol. 26.

[7] M. Sonnenschein, O. Lünsdorf, J. Bremer, and M. Tröschel, "Decentralized control of units in smart grids for the support of renewable energy supply," *Environmental Impact Assessment Review*, no. 0, pp. –, 2014, in press.

[8] R. Kamphuis, C. Warmer, M. Hommelberg, and K. Kok, "Massive coordination of dispersed generation using powermatcher based software agents," in *19th International Conference on Electricity Distribution*, 05 2007.

[9] K. Kok, Z. Derzsi, J. Gordijn, M. Hommelberg, C. Warmer, R. Kamphuis, and H. Akkermans, "Agent-based electricity balancing with distributed energy resources, a multiperspective case study," *Hawaii International Conference on System Sciences*, vol. 0, p. 173, 2008.

[10] A. Kamper and A. Esser, "Strategies for decentralised balancing power," in *Biologically-inspired Optimisation Methods: Parallel Algorithms, Systems and Applications*, ser. Studies in Computational Intelligence, M. R. A. Lewis, S. Mostaghim, Ed. Berlin, Heidelberg: Springer, Juni 2009, no. 210, pp. 261–289.

[11] R.-C. Mihailescu, M. Vasirani, and S. Ossowski, "Dynamic coalition adaptation for efficient agent-based virtual power plants," in *Proceedings of the 9th German conference on Multiagent system technologies*, ser. MATES'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 101–112.

[12] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings, "Agent-based control for decentralised demand side management in the smart grid," in *AAMAS*, L. Sonenberg, P. Stone, K. Tumer, and P. Yolum, Eds. IFAAMAS, 2011, pp. 5–12.

[13] J. Bremer, B. Rapp, and M. Sonnenschein, "Support vector based encoding of distributed energy resources' feasible load spaces," in *IEEE PES Conference on Innovative Smart Grid Technologies Europe*, Chalmers Lindholmen, Gothenburg, Sweden, 2010.

[14] J. Bremer and M. Sonnenschein, "A distributed greedy algorithm for constraint-based scheduling of energy resources," in *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012, Proceedings*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012, pp. 1285–1292.

[15] ——, "Constraint-handling for optimization with support vector surrogate models – a novel decoder approach," in *ICAART 2013 – Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, J. Filipe and A. Fred, Eds., vol. 2. Barcelona, Spain: SciTePress, 2013, pp. 91–105.

[16] A. Nieße and M. Sonnenschein, "A fully distributed continuous planning approach for decentralized energy units," in *Multiagent System Technologies*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2015.

[17] A. Nieße, S. Beer, J. Bremer, C. Hinrichs, O. Lünsdorf, and M. Sonnenschein, "Conjoint dynamic aggrgation and scheduling for dynamic virtual power plants," in *Federated Conference on Computer Science and Information Systems - FedCSIS 2014, Warsaw, Poland*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 9 2014.

[18] J. Bremer and M. Sonnenschein, "Parallel tempering for constrained many criteria optimization in dynamic virtual power plants," in *Computational Intelligence Applications in Smart Grid (CIASG), 2014 IEEE Symposium on*, Dec 2014, pp. 1–8.

[19] A. Schiendorfer, J.-P. Steghöfer, and W. Reif, "Synthesised constraint models for distributed energy management," in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2014, pp. 1529–1538.

[20] C. Hinrichs, "Selbstorganisierte Einsatzplanung dezentraler Akteure im Smart Grid," Ph.D. dissertation, Carl von Ossietzky Universitt Oldenburg, 2014.

[21] J. Bremer and S. Lehnhoff, "Decentralized coalition formation in agent-based smart grid applications," in *Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection*, ser. Communications in Computer and Information Science, vol. 616. Springer, 2016, pp. 343–355.

[22] ——, *Hybrid Multi-ensemble Scheduling*. Cham: Springer International Publishing, 2017, pp. 342–358.

[23] S. McArthur, E. Davidson, V. Catterson, A. Dimeas, N. Hatziargyriou, F. Ponci, and T. Funabashi, "Multi-agent systems for power engineering applications – Part I: Concepts, approaches, and technical challenges," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1743–1752, 2007.

[24] M. Sonnenschein, C. Hinrichs, A. Nieße, and U. Vogel, "Supporting renewable power supply through distributed coordination of energy resources," in *ICT Innovations for Sustainability*, ser. Advances in Intelligent Systems and Computing, L. M. Hilty and B. Aebischer, Eds. Springer International Publishing, 2015, vol. 310, pp. 387–404.

[25] F. Gieseke and O. Kramer, "Towards non-linear constraint estimation for expensive optimization," in *Applications of Evolutionary Computation*, ser. Lecture Notes in Computer Science, A. Esparcia-Alcázar, Ed. Springer Berlin Heidelberg, 2013, vol. 7835, pp. 459–468.

[26] J. Bremer and M. Sonnenschein, "Model-based integration of constrained search spaces into distributed planning of active power provision." *Comput. Sci. Inf. Syst.*, vol. 10, no. 4, pp. 1823–1854, 2013.

[27] C. A. Coello Coello, "Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art," *Computer Methods in Applied Mechanics and Engineering*, vol. 191, no. 11-12, pp. 1245–1287, Jan. 2002.

[28] C. Hinrichs, M. Sonnenschein, and S. Lehnhoff, "Evaluation of a Self-Organizing Heuristic for Interdependent Distributed Search Spaces," in *International Conference on Agents and Artificial Intelligence (ICAART 2013)*, J. Filipe and A. L. N. Fred, Eds., vol. Volume 1 – Agents. SciTePress, 2013, pp. 25–34.

[29] C. Hinrichs, S. Lehnhoff, and M. Sonnenschein, "A Decentralized Heuristic for Multiple-Choice Combinatorial Optimization Problems," in *Operations Research Proceedings 2012*. Springer, 2014, pp. 297–302.

[30] C. Hinrichs, J. Bremer, and M. Sonnenschein, "Distributed Hybrid Constraint Handling in Large Scale Virtual Power Plants," in *IEEE PES Conference on Innovative Smart Grid Technologies Europe (ISGT Europe 2013)*. IEEE Power & Energy Society, 2013.

[31] A. Nieße and M. Sonnenschein, "A fully distributed continuous planning approach for decentralized energy units," in *Informatik 2015. GI-Edition - Lecture Notes in Informatics (LNI)*, D. W. Cunningham, P. Hofstedt, K. Meer, and I. Schmitt, Eds., vol. 246. Bonner Köllen Verlag, 2015, pp. 151–165.

[32] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm Intelligence*, vol. 1, no. 1, pp. 33–57, 2007.

[33] D. Karaboga and B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, Nov. 2007.

[34] T. Lust and J. Teghem, "The multiobjective multidimensional knapsack problem: a survey and a new approach," *CoRR*, vol. abs/1007.4063, 2010.

[35] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[36] J. Bremer and M. Sonnenschein, "Sampling the search space of energy resources for self-organized, agent-based planning of active power provision," in *27th International Conference on Environmental Informatics for Environmental Protection, EnviroInfo 2013*, B. Page, A. G. Fleischer, J. Göbel, and V. Wohlgemuth, Eds. Shaker, 2013, pp. 214–222.

[37] P. Hall, "The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable," *Biometrika*, vol. 19, no. 3/4, pp. pp. 240–245, 1927.

[38] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Mach. Learn.*, vol. 54, no. 1, pp. 45–66, 2004.

[39] A. Ostermeier, A. Gawelczyk, and N. Hansen, "A derandomized approach to self-adaptation of evolution strategies," *Evolutionary Computation*, vol. 2, no. 4, pp. 369–380, 1994.

[40] N. Hansen, "The CMA evolution strategy: a comparing review," in *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, Eds. Springer, 2006, pp. 75–102.

[41] ——, "The CMA Evolution Strategy: A Tutorial," Tech. Rep., 2011.

[42] O. Kramer, A. Barthelmes, and G. Rudolph, "Surrogate constraint functions for cma evolution strategies," in *Proceedings of the 32Nd Annual German Conference on Advances in Artificial Intelligence*, ser. KI'09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 169–176.

[43] D. V. Arnold and N. Hansen, "A (1+1)-cma-es for constrained optimisation," in *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '12. New York, NY, USA: ACM, 2012, pp. 297–304.

[44] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, Jun. 2001.

[45] J. Bremer, B. Rapp, and M. Sonnenschein, "Encoding distributed search spaces for virtual power plants," in *IEEE Symposium Series on Computational Intelligence 2011 (SSCI 2011)*, Paris, France, 4 2011.

[46] J. Neugebauer, O. Kramer, and M. Sonnenschein, "Classification cascades of overlapping feature ensembles for energy time series data," in *Proceedings of the 3rd International Workshop on Data Analytics for Renewable Energy Integration (DARE'15)*. Springer, 2015.

[47] M. Sonnenschein, H.-J. Appelrath, W.-R. Canders, M. Henke, M. Uslar, S. Beer, J. Bremer, O. Lünsdorf, A. Nieße, J.-H. Psola *et al.*, "Decentralized provision of active power," in *Smart Nord - Final Report*. Hartmann GmbH, Hannover, 2015.

[48] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis and applications," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3. IEEE, 2005, pp. 1653–1664.

11th International Workshop on Computational Optimization

MANY real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

TOPICS

The list of topics includes, but is not limited to:

- combinatorial and continuous global optimization
- unconstrained and constrained optimization
- multiobjective and robust optimization
- optimization in dynamic and/or noisy environments
- optimization on graphs
- large-scale optimization, in parallel and distributed computational environments
- meta-heuristics for optimization, nature-inspired approaches and any other derivative-free methods
- exact/heuristic hybrid methods, involving natural computing techniques and other global and local optimization methods

The applications of interest are included in the list below, but are not limited to:

- classical operational research problems (knapsack, traveling salesman, etc)
- computational biology and distance geometry
- data mining and knowledge discovery
- human motion simulations; crowd simulations
- industrial applications
- optimization in statistics, econometrics, finance, physics, chemistry, biology, medicine, and engineering.

BEST PAPER AWARD

The best WCO'18 paper will be awarded during the social dinner of FedCSIS 2018.

The best paper will be selected by WCO'18 co-Chairs by taking into consideration the scores suggested by the reviewers, as well as the quality of the given oral presentation.

EVENT CHAIRS

- **Fidanova, Stefka,** Bulgarian Academy of Sciences, Bulgaria
- **Mucherino, Antonio,** INRIA, France
- **Zaharie, Daniela,** West University of Timisoara, Romania

PROGRAM COMMITTEE

- **Abud, Germano,** Universidade Federal de Uberlândia, Brazil
- **Bonates, Tibérius,** Universidade Federal do Ceará, Brazil
- **Breaban, Mihaela**
- **Chira, Camelia,** Technical University of Cluj-Napoca, Romania
- **Gruber, Aritanan**
- **Hosobe, Hiroshi,** Hosei University, Japan
- **Iiduka, Hideaki,** Kyushu Institute of Technology, Japan
- **Lavor, Carlile,** IMECC-UNICAMP, Brazil
- **Micota, Flavia,** West University of Timisora, Romania
- **Muscalagiu, Ionel,** Politehnica University Timisoara, Romania
- **Pintea, Camelia,** Tehnical University Cluj-Napoca, Romania
- **Siarry, Patrick,** Universite Paris XII Val de Marne, France
- **Stefanov, Stefan,** South-West University ""Neofit Rilski, Bulgaria
- **Stoean, Catalin,** University of Craiova, Romania
- **Stuetzle, Thomas,** Université Libre de Bruxelles (ULB), Belgium
- **Zilinskas, Antanas,** Vilnius University, Lithuania

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 227–231
DOI: 10.15439/2018F26
ISSN 2300-5963 ACSIS, Vol. 15

A Graph-Theoretic Approach to the Train Marshalling Problem

Jens Dörpinghaus
Fraunhofer Institute for Algorithms and Scientific Computing,
Schloss Birlinghoven, Sankt Augustin, Germany
Email: jens.doerpinghaus@scai.fraunhofer.de

Rainer Schrader
Institut für Informatik,
Universität zu Köln, Germany
Email: schrader@zpr.uni-koeln.de

***Abstract*—Rearranging cars of an incoming train in a hump yard is a widely discussed topic. We focus on the train marshalling problem where the incoming cars of a train are distributed to a certain number of sorting tracks. When pulled out again to build the outgoing train, cars sharing the same destination should appear consecutively. The goal is to minimize the number of sorting tracks. We suggest a graph-theoretic approach for this $\mathcal{NP}$-complete problem. The idea is to partition an associated directed graph into what we call pseudochains of minimum length. We describe a greedy-type heuristic to solve the partitioning problem which, on random instances, performs better than the known heuristics for the train marshalling problem.**

I. Introduction

A HUMP yard usually consists of a hump and a set of classification or sorting tracks and one or more roll-in and pull-out tracks [1]. In hump yards freight cars are arranged or rearranged into a specific sequence of cars. The outgoing trains will deliver goods to new destinations. A practical introduction to hump yards with examples can be found in the work of Hiller [2].

This can be very complex. For example the hump yard in in Zürich-Limmattal (CH) consists of 18 roll-in tracks, 64 sorting tracks with a length of 650-850 meters and 16 roll-out tracks, see [2][3].

Every incoming car arriving at the hump yard will be assigned to a sorting track. At the end of this process all cars of every sorting track will be placed as a block on the roll-out track. For an optimization approach the number and length of sorting tracks, the number of roll-in and pull-out operations can be minimized.

Hansmann provided a general class of *Sorting of rolling Stock Problems* (SRSP) in [4]. We will focus on the *Train Marshalling Problem* (TMP): using a minimum number of tracks, rearrange the cars in a hump yard in such a way that cars sharing the same destinations appear consecutively in the rearranged train.

During the process only two movements are allowed: the sorting of cars to the tracks and one pull-out movement for all cars. The tracks are not limited in length, so we can think of the tracks as stacks. We only allow one roll-in operation per car and one pull-out operation per track. No further shunting is allowed.

Apparently, TMP was first introduced by Zhu and Zhu [5] in 1983 who considered it under additional constraints and gave first results and polynomial algorithms. In 2000, Dahlhaus et al. [6] proved that TMP is $\mathcal{NP}$-complete and introduced new bounds. Brueggeman et al. show in [7] that the problem is fixed parameter tractable. In another work by Dahlhaus, Manne, Miller and Ryan [8] they described similar problems. More bounds and algorithms can be found in the work of Beygang [9] and Beygang et al. [1]. They introduced a graph-theoretic approach by considering the interval graph of a given instance. The problem also occurs in the works of Hansmann [4]. Other approaches can be found in the work of Rinaldi and Rizzi [10] who focused on dynamic programming and Haahr and Lusby [11].

First of all we will give a short formal problem description and all relevant definitions. After introducing pseudochains and discussion splittable destinations we will derive a novel greedy heuristic to solve the TMP. We will evaluate the results on some random instances and finish with a conclusion.

II. Problem description

With every *car* i in the hump yard we associate a natural number $\sigma_i \in \mathbb{N}^+$ representing the destination of the car. A *train* σ of length n then is a sequence

$$\sigma = (\sigma_1, ..., \sigma_n)$$

of cars with $\sigma_i \in \{1, ..., d\}$ for $i \in \{1, ...n\}$.

Example II.1. *Let $\sigma = (1, 2, 1, 3, 2)$. There are three destinations, where the first and third car and, resp., the second and the last have the same destination.*

We want to rearrange the cars in a departing train such that all cars are sorted in blocks according to their destination. For this, only two shunting operations are permitted: the roll-in movement of a car to one of the sorting tracks and the pull-out of all cars on a sorting track. The goal is to minimize the number of sorting tracks, denoted by $K(\sigma)$. Since only one shunting operation per sorting track is allowed the minimization of shunting operations is equivalent to the minimization of sorting tracks.

For a given sequence σ let S_k be the elements of σ with destination k. Then we may describe the incoming sequence by a partition $S = \{S_1, \ldots, S_d\}$ of $\{1, 2, ...n\}$. Dahlhaus et al.

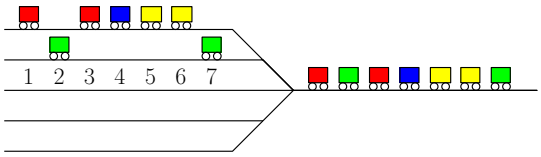


Fig. 1: Illustration for example II.2.

[6] have shown that TMP may be rephrased as follows: find the smallest number $K(S)$ and a permutation π of $1,\ldots,d$ such that the sequence of numbers

$$\underbrace{1,2,...,n,1,2,...,n,1,2,...,n}_{K(S)\text{-times}}$$

contains the elements of $S_{\pi(1)}$ as a subsequence followed by all elements of $S_{\pi(2)}$ and so on.

Example II.2. *Let $n = 7$, $d = 4$ und $S = \{S_1, S_2, S_3, S_4\}$ with $S_1 = \{1,3\}$, $S_2 = \{2,7\}$, $S_3 = \{4\}$ and $S_4 = \{5,6\}$. Then $K(S) = 2$ and $\pi(1,3,4,2)$, see figure 1 for an illustration:*

$$\underbrace{\mathbf{1}\,2\,\mathbf{3}}_{S_1}\ \underbrace{\mathbf{4}}_{S_3}\ \underbrace{\mathbf{5}\,\mathbf{6}}_{S_4}\ 7\ 1\ \underbrace{\mathbf{2}\,3\,4\,5\,6\,\mathbf{7}}_{S_2}$$

We now define some necessary preliminaries following the work of Beygang in [9].

III. Preliminaries

Let $\mathbb{S}^n$ be the set of all problem instances of TMP with n cars. For $S \in \mathbb{S}^n$ let $d = d(S)$ be the number of destinations in this instance. For an instance $S \in \mathbb{S}^n$, a *track assignment* is function $tr : \{1,...,n\} \to \mathbb{N}$ which assigns a track to every car. A track assignment is feasible if it gives a feasible solution for TMP.

For a given sequence σ and a destination k let $first(k)$ denote the position of the first occurrence of k and $last(k)$ be its last occurrence. Let $I_k = [first(k), last(k)]$ be the associated interval. Then the intervals induce a partial order on the set of destinations via $i < j$ if $last(i) < first(j)$. We consider the associated comparability graph and its complement, the interval graph.

Definition III.1. *(Comparability Graph associated with an Input Instance) For a given instance S of TMP, the associated comparability graph graph is given by $D(S) = (V, A)$ such that $V = (I_1, ..., I_d)$ and $(I_k, I_j) \in A$ if $k < j$.*

Definition III.2. *(Interval Graph associated with an Input Instance) For a given instance S of TMP, the associated interval graph is given by $G(S) = (V, E)$ such that $V = (I_1, ..., I_d)$ and $(I_k, I_j) \in E$ if $I_k \cap I_j \neq \emptyset$.*

Beygang already introduced some bounds and two important heuristics for the TMP. The deterministic SPLIT-Algorithm was introduced in [9] and computes a feasible solution by splitting destinations whenever possible in $O(n)$. The GREEDY-Algorithm was also introduced in [9]. It finds a feasible solution by partitioning the interval graph $G(S)$ into a minimum number $\chi(G_S)$ of stable sets, each assigned to one track. Recall that this is equivalent to partitioning $D(S)$ into a minimum number of chains. We will generalize this approach to partition $D(S)$ into pseudochains.

IV. Pseudochains

Let $D = (V, A \cup B)$ be a directed graph with a set B of blue arcs, $A \cap B = \emptyset$. We allow that $B = \emptyset$ or $A = \emptyset$. Recall that a chain in a transitively oriented graph is a subset $v_1, \ldots, v_k$ of vertices such that $(v_i, v_j) \in A$ for all $1 \leq i < j \leq k$.

Definition IV.1. *(Pseudochain) Let $D = (V, A \cup B)$ as above such that the subgraph D_A induced by the arcs in A is transitively orientable. $\mathcal{C} \subseteq V$ is a* pseudochain *of length $\ell(\mathcal{C}) = k \geq 1$ if $\mathcal{C}$ can be written as*

$$\mathcal{C} = C_1, b_2, C_2, b_3, C_3, ..., b_k, C_k,$$

where the C_i's are mutually disjoint chains in D_A with last element a_i and first Element c_i and $(a_{i-1}, b_i), (b_i, c_i) \in B$ for $2 \leq i \leq k$.

Figure 2 illustrates a pseudochain of length three.

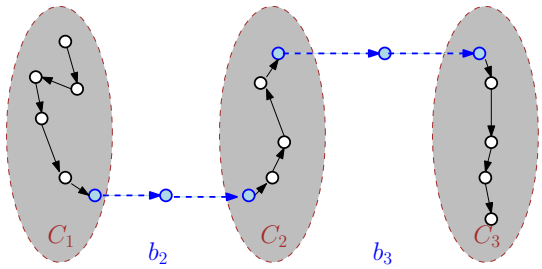


Fig. 2: A pseudostable chain of length 3. Transitive arcs are omitted, dashed arcs correspond to blue arcs.

Now we can define the minimization problem as follows.

Definition IV.2. *(minPC) Given a directed graph $D = (V, A \cup B)$ with a set B of blue edges, $A \cap B = \emptyset$ such that D_A is transitively orientable. Partition V into pseudochains $P = \mathcal{C}_1, \ldots, \mathcal{C}_k$ such that the total length $\ell(P) = \sum_{i=1}^{k} \ell(\mathcal{C}_i)$ of the partition is minimal.*

Lemma IV.3. *Given a directed graph $D = (V, A \cup B)$ with a set B of blue edges such that D_A is transitively orientable and a minimum partition V into pseudochains $P = \mathcal{C}_1, \ldots, \mathcal{C}_k$. Then there is a partition P' with $\ell(P') = \ell(P)$ such that for all centers $c(b_i)$ of all blue paths b_i in P' exist two nodes u in C_{i-1} and v in C_i so that $(c(b_i), v)$, $(u, c(b_i))$ and $(u, v) \notin A$.*

V. Splittable destinations

Observe that we can always produce a feasible track assignment by opening a track for each destination. But we may be able to do better by distributing the cars of one destination to two tracks. For this, we consider three destinations (a, b, c) with $I_a \cap I_b \neq \emptyset$ and $I_b \cap I_c \neq \emptyset$. Thus we need at least two tracks for $S(a) \cup S(b)$ and two tracks for $S(b) \cup S(c)$. We call the destination b *splittable* with predecessor a and successor c

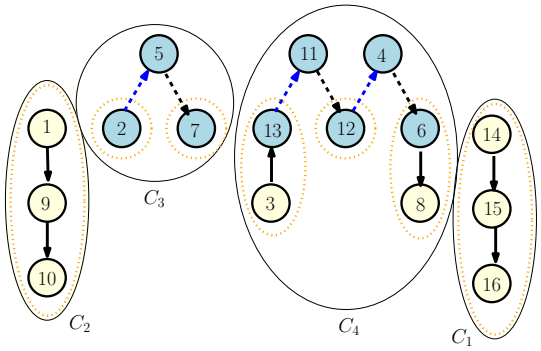


Fig. 3: A pseudochain partition found in example VI.1

if there is a feasible track assignment which assigns b to two different tracks. For short, we say that a triple (a, b, c) with the properties above is splittable.

In order to be feasible, some cars of $S(b)$ must then form a block at the end of one track containing the cars of $S(a)$ and a block at the beginning of some other track containig the cars of $S(c)$. It is easy to show the following Theorem:

Theorem V.1. *(Splittable Destinations) Given an instance $S \in \mathbb{S}^n$ of TMP. Let (a, b, c) be a triple of destinations with $I_a \cap I_b \neq \emptyset$ and $I_b \cap I_c \neq \emptyset$. Then the triple (a, b, c) is splittable if and only if there is no car of destination b between $first(c)$ and $last(a)$.*

Proof. Given a feasible track assignment and a destination b which is assigned to two tracks. We may assume that both tracks contain cars of other destinations. Let a be the destination preceding cars of b on the first track and c the destination following the cars on the other track. Then it is easy to see, that in the incoming train either $last(a) < first(c)$ or no car of destination b occurred between $first(c)$ and $last(a)$.

Conversely, let (a, b, c) a triple of destinations (a, b, c) as above. We start by assigning $S(a)$ to an open track, track 1 say, and cars of $S(c)$ to track 2. Cars of destination b will also be assigned to track 2 if they occur before $last(a)$, and to track one track 1 otherwise. All other cars are assigned to a destination-specific track. By assumption, either $last(a) < first(c)$, and the assignment is feasible. In the other case, after the first car of $S(c)$ is assigned to track 2, all remaining cars of $S(b)$ are assigned to the end of track 1. So in both cases the assignment is feasible. □

Observe that splittable triples cannot be read off from the comparability graph D itself. So in the next section we will enhance D to capture this extra information.

VI. MINPC AND TMP ARE EQUIVALENT

Let D be the comparability graph of the intervals given by an instance $S \in \mathbb{S}^n$ and (a, b, c), a splittable triple. Observe that by definition I_b overlaps both I_a and I_c. So $(a, b), (b, c) \notin A$. Let $B = \{(a, b), (b, c) \mid (a, b, c)$ is a splittable triple$\}$ and $D^* = D^*(S) = (V, A \cup B)$ be the *extended comparability graph* of S.

Example VI.1. *Given an instance $S \in \mathbb{S}^{50}$ with 16 destinations and*

$$\sigma = (1, 1, 2, 1, 2, 3, 3, 3, 4, 2, 2, 1, 5, 3, 3, 4, 2, 1, 1, 6, 6, 2, 5, 7, 8, 1, 9, 10, 8, 11, 12, 13, 2, 5, 8, 10, 14, 14, 15, 16, 16, 12, 7, 4, 10, 5, 7, 8, 13, 11)$$

A partition P of the extended comparability graph $D^*(S)$ in pseudochains is given by

- $\mathcal{C}_1 = \{14, 15, 16\}$ *with* $\ell(\mathcal{C}_4) = 1$.
- $\mathcal{C}_2 = \{1, 9, 10\}$ *with* $\ell(\mathcal{C}_2) = 1$.
- $\mathcal{C}_3 = \{2, 5, 7\}$ *with* $C_1 = \{2\}$, $b_2 = 5$, $C_2 = \{7\}$ *and* $\ell(P_1) = 2$.
- $\mathcal{C}_4 = \{3, 4, 6, 8, 11, 12, 13\}$ *with* $C_1 = \{3, 13\}$, $b_2 = 11$, $C_2 = \{12\}$, $b_3 = 4$, $C_3 = \{6, 8\}$ *and* $\ell(P_3) = 3$.

See Figure 3. The weight is $\ell(P) = 7$.

Lemma VI.2. *Let $S \in \mathbb{S}^n$ and P, a pseudochain partition of the extended comparability graph $D^*(S)$. Then P induces a feasible track assignment using $\ell(P)$ tracks.*

Proof. It suffices to show that we can assign a pseudochain $\mathcal{C} = C_1, b_2, C_2, b_3, C_3, ..., b_k, C_k$ of length k to k tracks. Let chain C_i begin with c_i and end with a_i. Let $B'_i = \{c \in b_i : c < last(c_{i-1})\}$ and $B''_i = \{c \in b_i : c > last(c_{i-1})\}$. We claim that, for $1 \leq i \leq k-1$, we can schedule the pseudochain such that track i contains B'_{i-1} followed by C_i again followed by B''_i. Suppose this is true for some $1 \leq j < k$. Since, by definition, the triple (a_j, b_{j+1}, c_{j+1}) is splittable, we may fill track j with B''_{j+1}, open track $j+1$ with B'_j and fill it with C_{j+1}. □

Lemma VI.3. *Let $S \in \mathbb{S}^n$ and tr, a feasible track assignment using k trains. Then tr induces a pseudochain partition P of the extended comparability graph with $\ell(P) = k$.*

Proof. Since tr is feasible, the cars of a destination d are assigned to at most two tracks and form a consecutive subsequence on their tracks. If they are assigned to two tracks they must be placed at the end of one track and at the beginning of some other track. Define a directed graph H on the set of destinations. Two destinations i, j are linked by an edge (i, j) if cars of i are placed immediately before cars of j on the same track. Then each connected component of H induces a pseudochain C of $D(S)$. Since $\ell(C)$ corresponds to the number of tracks used by the component, the claim follows. □

Example VI.4. *Consider the instance of example VI.1. The pseudochain partition P induces the following track assignment:*

Track 1 :	$2_3, 2_{33}\ 5_{34}, 5_{51}$
Track 2 :	$5_1, 5_{23}\ 7_{24}, 7_{47}$
Track 3 :	$1_1, 1_{26}\ 9_{27}\ 10_{28}, 10_{45}$
Track 4 :	$3_6, 3_{15}\ 13_{32}, 13_{49}\ 11_{50}, 11_{51}$
Track 5 :	$11_1, 11_{30}\ 12_{31}, 12_{42}\ 4_{44}, 4_{51}$
Track 6 :	$4_9, 4_{16}\ 6_{20}, 6_{21}\ 8_{25}, 8_{48}$
Track 7 :	$14_{37}, 14_{38}\ 15_{39}\ 16_{40}, 16_{41}$

Here, the lower indices represent the position of the car in the input sequence. It is a feasible track assignment using $\ell(P) = 7$ tracks.

Theorem VI.5. *Let $S \in \mathbb{S}^n$ and $D^*(S)$, the extended comparability graph. Then minPC on $D^*(S)$ is equivalent to TMP.*

Proof. Follows from Lemma VI.2 and VI.3. □

Thus for every optimal solution of an instance $S \in \mathbb{S}^n$ of the TMP using $K(S)$ tracks, there exists a corresponding partition of the extended comparability graph $D(S)$ into pseudochains.

VII. A new Greedy-Approach: Greedy-PC

This greedy approach is based on the above observations on pseudochain partitions. Let $S \in \mathbb{S}^n$ be an instance of TMP and $T = (t_1, ..., t_{t_S})$ be the list of all splittable destination in $\mathcal{S}$ sorted increasingly by their left boundary. Our approach will return a partition of $D^*(S)$ into pseudochains.

Given a splittable destination $(a, b, c) \in T$, the function add applied to a pseudochain P returns `true` if the triple can be added to P and `false` otherwise. We follow the idea to have the best solution within this chain P and try to add every possible splittable triple in T to a pseudochain. We will redo this as long as nodes remain in $\mathcal{S}$.

The worst-case runtime of this heuristic is $f(n) = (\frac{1}{2}n^3 + n^2) = O(n^3)$. See algorithm 1 for an implementation in pseudocode.

VIII. Experimental Results

We used Python 3.4 with NetworkX for creating random instances and implement the greedy heuristic as well as the Linear Programming relaxation introduced by Beygang [9]. Four 2.4 GHz processors and 8 GB RAM were available running Linux Kernel 3.10. We used GLPK (GNU Linear Programming Kit) 4.52 to solve the linear program. To get comparable results, we followed [9] to create random instances. This function takes the number n of cars and computes uniform and independent problem instances.

The greedy heuristic introduced by Beygang et al. ([9] and [1]) leads to the upper bound denoted by *Coloring*. It is equivalent to a graph coloring approach for the interval graph $G(S)$. The runtime is $O(n^2)$. Algorithm 1 has also polynomial runtime in $O(n^3)$.

We approximate the optimal solution according to the bounds u_{lp} and l_{lp}, the upper and lower bound given by the solution of the linear program introduced by [9]. It was observed in [12] that the lower bound very often coincides with the value of the optimal solution.

Figures 4 and 5 summarize the output of the heuristics on 50 random instances with a fixed number of cars. For a small number of cars the distance between *Coloring* and u_{greedy} is small, but notable, see figure 4. We notice the greedy approach can lead to solutions using more tracks than the *Coloring* approach. The situation changes significantly for instances with more cars. Figure 5 shows that Greedy-PC performs better than *Coloring* on instances with 300 cars.

Algorithm 1 Greedy-PC

Require: Extended comparability graph $D^*(S)$ with its maximal stable set $\mathcal{S}$ in $D(S)$ and a list $T = (t_1, ..., t_{t_S})$ of splittable destinations in $\mathcal{S}$.
Ensure: Partition of $D^*(S)$ in pseudochains

```
visited = ∅
count = 0
while |T| > 0 do
    count++
    P.add(pseudochain P_count)
    for every (a, b, c) = t_i ∈ T do
        if P_count.add(a, b, c) = true then
            visited.add a, b, c
        end if
    end for
    for every v ∈ visited do
        delete every t_i containing v from T
    end for
end while
for every node v ∈ V(G) do
    if v ∉ visited then
        for i = 1, ..., count do
            if P_i.add(v) = true then
                visited.add v
                exit
            end if
        end for
        count++
        P.add(pseudochain P_count)
        P_count.add(v)
    end if
end for
return P
```

IX. Conclusions

We have introduced and discussed pseudochain partitions and their relation to the Train Marshalling Problem. There is only little discussion about TMP in the literature, but the problem has an intimate relationship to other sorting problems of rolling stock, see [4]. Thus it is an important step to provide a better understanding of the underlying graph structures. Pseudochain partitions directly lead to a new heuristic providing a improved upper bounds for optimal solutions of TMP. We could proof that every optimal solution of TMP is equivalent to a minimal partition of the corresponding extended comparability graph $D^*(S)$ into pseudochains.

The greedy approach has been evaluated for 2 instances with 100 and 300 cars, each consisting of 50 random instances each. The computational results show that the model is useful and the proposed Greedy-approach performs in general significantly better than other state-of-the art approaches.

To sum up, although we achieved encouraging results, there are still questions which are not answered or even discussed in this paper. For example, can the inherent structure of

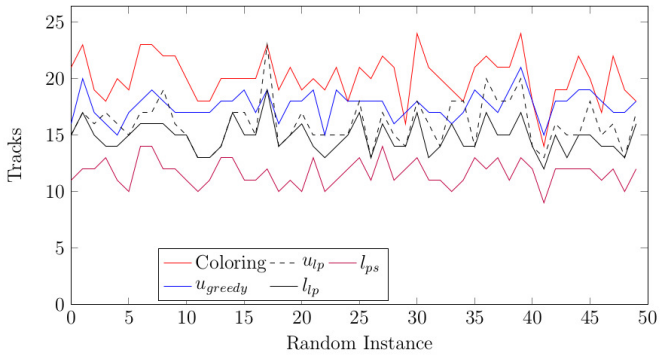


Fig. 4: Results for random instances with $n = 100$ cars. *Colouring* is a Greedy-approach introduced by Beygang, u_{lp} and l_{lp} are upper and lower bounds of the integer linear program approach, see [9]. The lower bound l_{ps} was introduced in [12]. u_{greedy} shows the results of our novel algorithm 1.

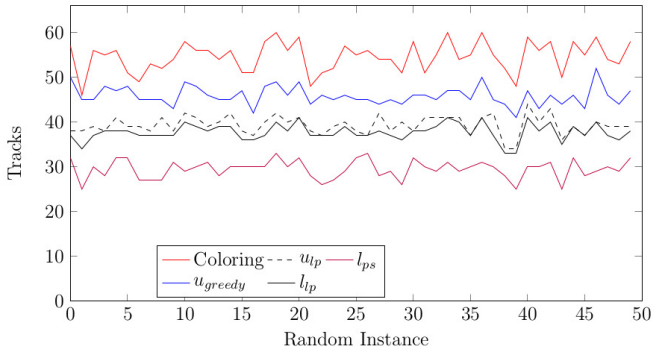


Fig. 5: Results for random instances with $n = 300$ *Colouring* is a Greedy-approach introduced by Beygang, u_{lp} and l_{lp} are upper and lower bounds of the integer linear program approach, see [9]. The lower bound l_{ps} was introduced in [12]. u_{greedy} shows the results of our novel algorithm 1.

pseudoschains be used to find even better heuristics than those discussed in this paper? Are there any instances of the TMP that can be solved in polynomial time?

The results encourages the further improvement on heuristics to solve minPC and the application of this method to other sorting of rolling Stock Problems.

REFERENCES

[1] K. Beygang, F. Dahms, and S. O. Krumke, "Train marshalling problem - algorithms and bounds," University of Kaiserslautern, Tech. Rep. 132, 2010.

[2] W. Hiller, *Rangierbahnhöfe*. Berlin: VEB Verlag für Verkehrswesen, 1983.

[3] M. Giger, "Rangierbahnhof Limmattal," *SWISS ENGINEERING STZ AUTOMATE NOW!*, vol. 1-2, pp. 93–94, 2010.

[4] R. S. Hansmann, *Optimal Sorting of Rolling Stock*. Göttingen: Cuvillier, 2011.

[5] Y. Zhu and R. Zhu, "Sequence reconstruction under some order-type constraints," *Scientia Sinica (A)*, vol. 26(7), pp. 702–713, 1983.

[6] E. Dahlhaus, P. Horak, M. Miller, and J. Ryan, "The train marshalling problem," *Discrete Applied Mathematics*, vol. 103(1-3), pp. 41–54, 2000. [Online]. Available: https://doi.org/10.1016/s0166-218x(99)00219-x

[7] L. Brueggeman, M. Fellows, R. Fleischer, M. Lackner, C. Komusiewicz, Y. Koutis, A. Pfandler, and F. Rosamond, "Train marshalling is fixed parameter tractable," in *Fun with Algorithms*, E. Kranakis, D. Krizanc, and F. Luccio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 51–56. [Online]. Available: https://doi.org/10.1007%2F978-3-642-30347-0_8

[8] E. Dahlhaus, F. Manne, M. Miller, and J. Ryan, "Algorithms for combinatorial problems related to train marshalling," *Proceedings of AWOCA 2000, Hunter Valley*, pp. 7–16, 2000.

[9] K. Beygang, *On the Solution of Some Railway Freight Car optimization Problems*. München: Dr. Hut, 2011.

[10] F. Rinaldi and R. Rizzi, "Solving the train marshalling problem by inclusion–exclusion," *Discrete Applied Mathematics*, vol. 217, Part 3, pp. 685 – 690, 2017.

[11] J. T. Haahr and R. M. Lusby, "A matheuristic approach to integrate humping and pullout sequencing operations at railroad hump yards," *Networks*, vol. 67, no. 2, pp. 126–138, 2016.

[12] J. Dörpinghaus, *Pseudostabile Mengen in Graphen*. Fraunhofer Verlag, 2018. [Online]. Available: https://kups.ub.uni-koeln.de/8066/

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 233–236
DOI: 10.15439/2018F47
ISSN 2300-5963 ACSIS, Vol. 15

Hybrid Ant Colony Optimization Algorithm for Workforce Planning

Stefka Fidanova
IICT, BAS
Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Gabriel Luque
DLCS University of Mlaga
29071 Mlaga, Spain
E-mail: gabriel@lcc.uma.es

Olympia Roeva
IBPhBME, BAS
Sofia, Bulgaria
E-mail: olympia@biomed.bas.bg

Marcin Paprzycki
SRI, PAS
Warsaw, Poland
E-mail: marcin.paprzycki@ibspan.waw.pl

Pawel Gepner
Intel Corporation
Swindon, UK
E-mail: pawel.gepner@intel.com

***Abstract*—Every organization and factory optimize their production process with a help of workforce planing. The aim is minimization of the assignment costs of the workers, who will do the jobs. The problem is very complex and needs exponential number of calculations, therefore special algorithms are developed to be solved. The problem is to select employers and to assign them to the jobs to be performed. This problem has very strong constraints and it is difficult to find feasible solutions. The objective is to fulfil the requirements and to minimize the assignment cost. We propose a hybrid Ant Colony Optimization (ACO) algorithm to solve the workforce problem, which is a combination between ACO and an appropriate local search procedure.**

Keywords: Workforce Planning, Ant Colony Optimization, Metaheuristics, Local search

I. Introduction

One of the most important decision making problem, common for all branches of industry, is workforce planning. The workforce planing is a part of the human resource management. It includes multiple level of complexity, therefore it is a hard optimization problem (NP-hard). This problem consists of two decision sets: selection and assignment. The first set shows selected employees from available workers. The assignment set shows which worker which job will perform. The aim is to fulfil the work requirements with minimal assignment cost.

As we mentioned the problem is a hard optimization problem with strong constraints and is impossible to be solved with exact methods or traditional numerical methods for instances with realistic size. These kind of methods can be apply only on some simplified variants of the problem. A deterministic workforce planing problem is studied in [11], [17]. Workforce planing models are reformulated as mixed integer programming in [11]. The authors show that the mixed integer program is much easier to solve than the non-linear program. In [17] the model includes workers differences and the possibility of workers training and upgrading. In [4] and [18] a variant with random demands of the problem is proposed. Two stage program of scheduling and allocating with random demands is considered in [4]. Other variant of the problem is to include uncertainty [12], [14], [16], [23], [24]. Most of the authors simplify the problem by omitting some of the constraints. In [6] a mixed linear programming is applied and in [18] a decomposition method is applied. For the more complex non-linear workforce planning problems, the convex methods are not applicable.

Nowadays, nature-inspired metaheuristic methods receive great attention [2], [15], [19], [21], [22]. In considered here problem some heuristic method including genetic algorithm [1], [13], memetic algorithm [20], scatter search [1] etc., are applied.

So far the Ant Colony Optimization (ACO) algorithm is proved to be very effective solving various complex optimization problems [7], [10]. In our previous work [8] we propose ACO algorithm for workforce planning. We have considered the variant of the workforce planning problem proposed in [1]. Current paper is the continuation of [8]. We propose a hybrid ACO algorithm which is a combination of ACO with a local search procedure. The aim is to improve the algorithm performance.

The rest of the paper is organized as follows. In Section 2 the mathematical description of the problem is presented. In Section 3 hybrid ACO algorithm for workforce planing problem is proposed. Section 4 shows computational results, comparisons and discussion. A conclusion and directions for future work are done in Section 5.

II. Definition of the Workforce Planning Problem

In this paper we solve the workforce planning problem proposed in [1] and [9]. The set of jobs $J = \{1, \ldots, m\}$ must be completed during a fixed period of time. The job j requires d_j hours to be completed. $I = \{1, \ldots, n\}$ is the set of workers, candidates to be assigned. Every worker must perform every of assigned to him job minimum h_{min} hours to can work in efficient way. Availability of the worker i is s_i hours. One worker can be assigned to maximum j_{max} jobs. The set A_i shows the jobs, that worker i is qualified. Maximum t workers can be assigned during the planed period, or at

most t workers may be selected from the set I of workers. The selected workers need to be capable to complete all the jobs. The aim is to find feasible solution, that optimizes the objective function. c_{ij} is the cost of assigning the worker i to the job j. The mathematical model of the workforce planing problem can be described as follows:

The objective function is the minimization of the total assignment cost. The number of hours for each selected worker is limited . The work must be done in full . The number of the jobs, that every worker can perform is limited . There is minimal number of hours that every job must be performed by every assigned worker to can work efficiently . The number of assigned workers is limited.

Same model can be used with different objective functions. Minimization of total assignment cost is the aim of this paper. If $\tilde{c}_{ij}$ is the cost the worker i to performs the job j for one hour, than the objective function can minimize the cost of the hall jobs to be finished.

$$f(x) = \text{Min} \sum_{i \in I} \sum_{j \in A_i} \tilde{c}_{ij}.x_{ij} \tag{1}$$

The workforce planning problem is difficult to be solved because of very restrictive constraints especially the relation between the parameters h_{min} and d_j. When the problem is structured (d_j is a multiple of h_{min}), in this case it is more easier to find feasible solution, than for unstructured problems (d_j and h_{min} are not related).

III. Hybrid Ant Colony Optimization Algorithm

The ACO is a nature inspired method. It is metaheuristics methodology following the behaviour of real ants looking for a food. Real ants use chemical substance, called pheromone, to mark their path ant to can return back. An ant moves in random way and when it detects a previously laid pheromone it decides whether to follow it and reinforce it with a new added pheromone. Thus the more ants follow a given trail, the more attractive that trail becomes. Using their collective intelligent the ants can find a shorter path between the source of the food and the nest.

A lot of problems coming from real life and industry needs exponential number of calculations. Therefore the only option is to be applied some metaheuristics. The goal is to find a good solution for a reasonable time [5].

A. ACO Algorithm for Workforce Planning

In this section we will apply the ACO algorithm for workforce planing from our previous work [8], which is without local search procedure. One of the main points of the ant algorithm is the proper representation of the problem by graph. In our case the graph of the problem is 3 dimensional and the node (i, j, z) corresponds worker with number i to be assigned to the job j for time z. The graph of the problem is asymmetric, because the maximal value of z depends of the value of j, different jobs needs different time to be completed. At the beginning of every iteration every ant starts to construct their solution, from random node of the graph of the problem. For every ant are generated three random numbers. The first random number corresponds to the worker we assign and is in the interval $[0, \ldots, n]$. The second random number corresponds to the job which this worker will perform and is in the interval $[0, \ldots, m]$. We verify if the worker is qualified to perform the job, if not, we chose in a random way another job. The third random number corresponds to the number of hours worker i is assigned to performs the job j and is in the interval $[h_{min}, \ldots, \min\{d_j, s_i\}]$. After, the ant applies the transition probability rule to include next nodes in the partial solution, till the solution is completed, or there is not a possibility to include new node.

We propose the following heuristic information:

$$\eta_{ijl} = \begin{cases} l/c_{ij} & l = z_{ij} \\ 0 & otherwise \end{cases} \tag{2}$$

When there are several candidate nodes with a same probability, the next node is chosen between them in a random way. When some move of the ant do not meets the problem constraints, then the probability of this move is set to be 0. If for all possible nodes the value of the transition probability is 0, it is impossible to include new node in the solution and the solution construction stops. When the constructed solution is feasible the value of the objective function is the sum of the assignment cost of the assigned workers. If the constructed solution is not feasible, the value of the objective function is set to be equal to -1. The ants constructed feasible solutions deposed a new pheromone on the elements of their solutions. The new added pheromone is equal to the reciprocal value of the objective function.

$$\Delta\tau_{i,j} = \frac{\rho - 1}{f(x)} \tag{3}$$

Thus the nodes of the graph belonging to solutions with less value of the objective function, receive more pheromone than others and become more desirable in the next iteration.

The end condition used in our algorithm is the number of iterations.

B. Local Search Procedure

The our main contribution in this paper is the hybridization of the ACO algorithm with a local search procedure. The aim of the local search is to decrease the time to find the best solution and eventually to improve the achieved solutions. We apply local search procedure only on infeasible solutions and only one time disregarding the new solution is feasible or not. Thus, our local search is not time consuming. If the solution is not feasible we remove part of the assigned workers and after that we assign in their place new workers. The workers which will be removed are chosen randomly. On this partial solution we assign new workers applying the rules of ant algorithm. The ACO algorithm is a stochastic algorithm, therefore the new constructed solution is different from previous one with a high probability.

TABLE I: Test instances characteristics

Parameters	Value
n	20
m	20
t	10
s_i	[50, 70]
j_{max}	[3, 5]
h_{min}	[10, 15]

TABLE II: ACO parameter settings

Parameters	Value
Number of iterations	100
ρ	0.5
τ_0	0.5
Number of ants	20
a	1
b	1

IV. Computational Results

In this section test results are reported and compared with ACO algorithm without local search procedure. We use the artificially generated problem instances considered in [1]. The test instances characteristics are shown in Table I.

The set of test problems consists of 20 used in [1], [8]. In our previous work [8] we show that our ACO algorithm outperforms the genetic and scatter search algorithms from [1]. The number of iterations is fixed to be maximum 100. In Table II the parameter settings of our ACO algorithm are shown. The values are fixed experimentally.

The workforce problem has very restrictive constraints. Therefore only 2-3 of the ants, per iteration, find feasible solution. Sometimes exist iterations without any feasible solution. Its complicates the search process. Our aim is to decrease the number of unfeasible solutions and thus to increase the possibility ants to find good solutions and so to decrease needed number of iterations to find good solution. We observe that after the local search procedure applied on the first iteration, the number of unfeasible solutions in a next iterations decrease. It is another reason the calculation time does not increase significantly. We are dealing with four cases: without local search procedure (ACO); local search procedure when the number of removed workers is quarter from the number of all assigned workers (ACO quarter); local search procedure when the number of removed workers is half from the number of all assigned workers (ACO half); local search procedure when all assigned workers are removed and the solution is constructed from the beginning (ACO restart).

We perform 30 independent runs with every one of the four cases, because the algorithm is stochastic ant to guarantee the robustness of the average results. We apply ANOVA test for statistical analysis to guarantee the significance of the achieved results. We are interested of the number of iterations for finding the best result. It can be very different for different test problems, so we will use ranking of the algorithms. The variant of our hybrid algorithm is on the first place, if it achieves the best solution with less average number of iterations over 30 runs, according other cases and we assign to it 1, we assign 2 to the case on the second place, 3 to the case on the third place and 4 to the case with most number of iterations. On some cases can be assigned same numbers if the number of iterations to find the best solution is the same. We sum the ranking of the cases over all 20 test problems to find final ranking of the different cases of the hybrid algorithm.

TABLE III: Hybrid ACO ranking

	ACO	ACO quarter	ACO half	ACO restart
first place	4 times	4 times	8 times	8 times
second place	4 times	4 times	7 times	6 times
third place	8 times	6 times	4 times	3 times
forth place	4 times	6 times	1 times	3 times
ranking	52	54	38	41

We observe that the local search procedure decreases the number of unfeasible solutions, found by traditional ACO algorithm in the next iterations, thus when the number of iterations increase, the need of local search procedure decreases. On Table III we report the achieved ranking of different cases of our hybrid algorithm. As we mentioned above, with ACO quarter we call the case when quarter of the workers are removed. ACO half is the case when half of the workers are removed. ACO restart is the case when all workers are removed. It is like to restart the solution construction, to construct the solution from the beginning.

We calculate the ranking, regarding the average number of iterations to find best solution over 30 runs of the test. When more than half of the ants find unfeasible solutions, the deviation from the average is larger compared to the tests when the most of the ants achieve feasible solutions. The Table III shows that the local search procedure decreases the number of iterations needed to find the best solution, when more than half of the workers are removed. The traditional ACO algorithm and hybrid ACO with removing quarter of the workers are 4 times on the first place when either, by chance, the algorithm find the best solution on the first iteration, or all ants find feasible solutions. We observe that the both cases are on the third and forth place 12 times. This means that removing less than half of the workers is not enough to construct feasible solution. The ACO algorithm with removed half of the workers 15 times is on the first or second place and only one time is on the fourth place, which means that it performs much better than previous two cases. When all workers are removed the achieved ranking is similar to the case when half of the workers are removed. Let the maximal number of assigned workers is t. Thus the every one of the solutions consists about t workers. If all of the workers are removed, the ant need to add new workers on their place which number is about t. When half of the workers are removed, then the ant will add about $t/2$ new workers. The calculation time to remove and

TABLE IV: Hybrid ACO comparison according calculation time

	ACO	ACO quarter	ACO half	ACO restart
first place	4 times	3 times	10 times	6 times
second place	7 times	5 times	5 times	5 times
third place	8 times	4 times	3 times	4 times
forth place	1 times	8 times	2 times	5 times
average time	82.244 s	93.98 s	79.63 s	103.012 s

add about $t/2$ workers is about two times less than to remove and add about t workers. Thus we can conclude that the local search procedure with removing half of the workers performs better than other cases.

Another way for comparison is the calculation time. For every test problem and every case we calculate the average time to achieve best solution over 30 runs. In Table IV we did similar ranking as in Table III, but taking in to account the calculation time instead number of iterations. Regarding the Table IV the ranking according the time is similar to the ranking according to the number of iterations from the Table III. The best performance is when half of the worker are removed in the local search procedure and the worst performance is when quarter of the workers are removed. The local search procedure with removing half of the worker is on the first place 10 times and on the forth place only 2 times. The local search procedure with removing quarter of the workers is on the first place 3 times and on the forth place 8 times. Regarding the calculation time the local search procedure with removing half of the workers again is the best, but the worst is the local search procedure with removing all workers. Reconstructing a solution from the beginning takes more time than to reconstruct partial solution, therefore ACO algorithm with local search procedure removing all workers performs worst. The results from Table IV show that removing only quarter of the workers from the solution is not enough for construction of good solution and is time consuming comparing with traditional ACO algorithm.

V. Conclusion

In this paper we propose Hybrid ACO algorithm for solving workforce assignment problem. The ACO algorithm is combined with appropriate local search procedure. The local search procedure is applied only on unfeasible solutions. The main idea is to remove part of the workers in the solution in a random way and to include new workers in their place. Three variants of the local search procedure are compared with traditional ACO algorithm, removing quarter of the assigned workers, removing half of the assigned workers and removing all assigned workers. The local search procedure with removing half of the assigned workers performs better than other algorithms.

Acknowledgment

Work presented here is partially supp, and by the Polish-Bulgarian collaborative grant “Practical aspects for scientific computing”.

References

[1] Alba E., Luque G., Luna F., *Parallel Metaheuristics for Workforce Planning*, J. Mathematical Modelling and Algorithms, Vol. 6(3), Springer, 2007, 509-528.

[2] Albayrak G., zdemir ., *A state of art review on metaheuristic methods in time-cost trade-off problems*, International Journal of Structural and Civil Engineering Research, Vol. 6(1), 2017, 30-34.

[3] Bonabeau E., Dorigo M. and Theraulaz G., *Swarm Intelligence: From Natural to Artificial Systems*, New York,Oxford University Press, 1999.

[4] Campbell G., *A two-stage stochastic program for scheduling and allocating cross-trained workers*, J. Operational Research Society 62(6), 2011, 10381047.

[5] Dorigo M, Stutzle T., *Ant Colony Optimization*, MIT Press, 2004.

[6] Easton F., *Service completion estimates for cross-trained workforce schedules under uncertain attendance and demand*, Production and Operational Management 23(4), 2014, 660675.

[7] Fidanova S., Roeva O., Paprzycki M., Gepner P., *InterCriteria Analysis of ACO Start Startegies*, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, 2016, 547-550.

[8] Fidanova S., Luquq G., Roeva O., Paprzycki M., Gepner P., *Ant Colony Optimization Algorithm for Workforce Planning*, FedCSIS'2017, IEEE Xplorer, IEEE catalog number CFP1585N-ART, 2017, 415-419.

[9] Glover F., Kochenberger G., Laguna M., Wubbena, T. *Selection and assignment of a skilled workforce to meet job requirements in a fixed planning period*. In:MAEB04, 2004, 636641.

[10] Grzybowska K., Kovcs, G., *Sustainable Supply Chain - Supporting Tools*, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Vol. 2, 2014, 13211329.

[11] Hewitt M., Chacosky A., Grasman S., Thomas B., *Integer programming techniques for solving non-linear workforce planning models with learning*, European J of Operational Research 242(3),2015, 942950.

[12] Hu K., Zhang X., Gen M., Jo J., *A new model for single machine scheduling with uncertain processing time*, J Intelligent Manufacturing, Vol 28(3), Springer, 2015, 717-725.

[13] Li G., Jiang H., He T., *A genetic algorithm-based decomposition approach to solve an integrated equipment-workforce-service planning problem*, Omega, Vol. 50, Elsevier, 2015, 117.

[14] Li R., Liu G., *An uncertain goal programming model for machine scheduling problem*. J. Inteligent Manufacturing, Vol. 28(3), Springer, 2014, 689-694.

[15] Mucherino A., Fidanova S., Ganzha M., *Introducing the environment in ant colony optimization*, Recent Advances in Computational Optimization, Studies in Computational Intelligence, Vol. 655, 2016, 147158.

[16] Ning Y., Liu J., Yan L., *Uncertain aggregate production planning*, Soft Computing, Vol. 17(4), Springer, 2013, 617624.

[17] Othman M., Bhuiyan N., Gouw G., *Integrating workers' differences into workforce planning*, Computers and Industrial Engineering, Vol. 63(4), 2012, 10961106.

[18] Parisio A, Jones CN., *A two-stage stochastic programming approach to employee scheduling in retail outlets with uncertain demand*, Omega, Vol. 53, Elsevier, 2015, 97-103.

[19] Roeva O., Atanassova V., *Cuckoo search algorithm for model parameter identification*, Int. J. Bioautomation, Vol. 20(4), 2016, 483492.

[20] Soukour A., Devendeville L., Lucet C., Moukrim A., *A Memetic algorithm for staff scheduling problem in airport security service*, Expert Systems with Applications, Vol. 40(18), 2013, 75047512.

[21] Tilahun S.L., Ngnotchouye J.M.T., *Firefly algorithm for discrete optimization problems: A survey*, Journal of Civil Engineering, Vol. 21(2), 2017, 535545.

[22] Toimil D., Gmes A., *Review of metaheuristics applied to heat exchanger network design*, International Transactions in Operational Research, Vol. 24(1-2), 2017, 726.

[23] Yang G., Tang W., Zhao R., *An uncertain workforce planning problem with job satisfaction*, Int. J. Machine Learning and Cybernetics, Springer, 2016. doi:10.1007/s13042-016-0539-6 http://rd.springer.com/article/10.1007/s13042-016-0539-6

[24] Zhou C., Tang W., Zhao R., *An uncertain search model for recruitment problem with enterprise performance*, J Intelligent Manufacturing, Vol. 28(3), Springer, 2014, 295-704. doi:10.1007/s10845-014-0997-1

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 237–243
DOI: 10.15439/2018F201
ISSN 2300-5963 ACSIS, Vol. 15

Community based influence maximization in the Independent Cascade Model

László Hajdu*, Miklós Krész†, András Bóta‡,
*University of Szeged
Institute of Informatics
Árpád tér 2,
6720 Szeged, Hungary
Email: hajdul@inf.u-szeged.hu
†University of Szeged
Gyula Juhász Faculty of Education
Boldogasszony sgt. 6
6720 Szeged, Hungary
also at
Innorenew CoE
Livade 6,
6310 Izola, Slovenia
also at
University of Primorska
Andrej Marušič Institute
Muzejski trg 2
SI-6000 Koper, Slovenia
E-mail: miklos.kresz@innorenew.eu
‡University of Szeged
Gyula Juhász Faculty of Education
Boldogasszony sgt. 6
6720 Szeged, Hungary
E-mail: bandras@inf.u-szeged.hu

***Abstract*—Community detection is a widely discussed topic in network science which allows us to discover detailed information about the connections between members of a given group. Communities play a critical role in the spreading of viruses or the diffusion of information. In [1], [8] Kempe et al. proposed the Independent Cascade Model, defining a simple set of rules that describe how information spreads in an arbitrary network. In the same paper the influence maximization problem is defined. In this problem we are looking for the initial vertex set which maximizes the expected number of the infected vertices. The main objective of this paper is to further improve the efficiency of influence maximization by incorporating information on the community structure of the network into the optimization process. We present different community-based improvements for the infection maximization problem, and compare the results by running the greedy maximization method.**

I. Introduction

BUILDING networks between people, companies, or other individuals based on their activities or properties became a common task in the previous decade. These networks describe the connection structure of their members, and show us a bigger and more detailed picture about their behavior. One of the most useful methods applied to networks is the detection of dense subgraphs, known as the detection of *communities*. Community detection is a well researched area of network science and a large variety of methods exists in its literature[2], but validating the results of an arbitrary community detection method, especially in an application-oriented way, remains an open problem.

Strong connections between individuals belonging to the same community make it easy for viruses, information or influence to spread between members. The Independent Cascade [1] model provides a possible scenario of how an actual spreading event can happen. The inputs of this model are: a graph, an assignment of edge infection probabilities to its edges and the set of initially active vertices. The process is iterative and in each iteration, every active vertex tries to activate its neighbors with the probability assigned to the edge connecting them. Each vertex remains active for exactly one iteration, afterwards it is removed from the spreading process. The process stops if there are no more active vertices. In the same paper the influence maximization problem is defined. In this problem we are looking for the initial vertex set which maximizes the expected number of the infected vertices. While the optimization problem is NP-complete, Kempe et al.

proposed a greedy method that gives a guaranteed precision result. A variety of other algorithms and heuristics were proposed to improve the efficiency of influence maximization [3][20][21]. A good overview of the maximization problem can be found in [22].

The greedy method gives us a good and guaranteed solution for infection maximization, but in real-sized networks it is unable to solve the task within a acceptable time. Here we introduce a new method, where the search space of the original greedy method is reduced based on different scores. Another objective is to compare the output of different community detection algorithms. Community detection methods are hard to validate if real life, since information about the members is not available.

In this paper we present new community based infection maximization methods which can improve the basic greedy method and increase the size of the solvable network. The methodology is also suitable to validate and compare different community detection methods.

II. Community detection

The main objective of community detection is to find dense subgraphs. The largest fraction of detection methods in the literature defines communities as disjunct sets of nodes. A significant number of works, however, follow a different approach, allowing overlaps between the groups of nodes. In this paper we take the latter, overlapping approach.

First of all we define different community detection algorithms to extract information for the infection maximization algorithm. For this purpose we chose a directed community detection method, and converted an undirected method from the literature to directed. The first algorithm which is used in this paper, is the directed version [6] of the original Clique percolation method [5]. The second method is the Hub Percolation method (HPM) [4], which was extended to work on directed networks.

A. Directed Hub Percolation

The original hub percolation [4] method is based on cliques and hubs. Maximal cliques are maximal fully connected subgraphs of an arbitrary graph, while hubs are locally important nodes in community detection. We choose the method because during the detection process, the method provides additional information which can be useful for the maximization problem. At first the algorithm finds undirected maximal cliques containing at least 3 nodes in the network. In our case the clique detection algorithm is replaced by a directed clique detection algorithm, and an additional parameter is introduced in the end of the method because providing higher resolution of the results. First of all we define the concept of a directed maximal clique.

Let d_{v_c} be the restricted out-degree of a node v in clique c: the out-degree of a given node inside the clique. The definition of the directed maximal clique is the following:

- The clique contains all directed edges from v_1 to v_2 where $d_{v1_c} > d_{v2_c}$

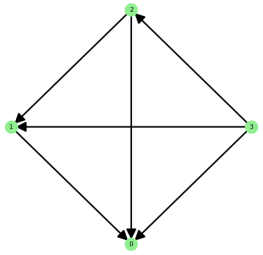


Fig. 1. An example of a directed clique. The restricted out degree of the nodes are 3, 2, 1 and 0.

- The clique contains no directed loops
- Every node in a clique has a different restricted out-degree
- It is maximal so it can not be expanded to a bigger clique

The Figure 1 shows an example of a directed clique. The restricted out degree of the nodes are different from each other. In the literature this structure also called transitive tournament [18][19]. Based on the clique definition, the algorithm of the directed hub percolation method is the following:

1) Find all at least 3 sized maximal cliques in the network. Let C contain these cliques.
2) A HubValue is defined for every node as follows: $\forall v \in V(G)$ let $h_v = |H_v|$ where $H_v = \{h | v \in h, h \in C\}$.
3) Base on h_v and a Hub Selection strategy we decide whether vertices are chosen as hub. Let H be the set of the hubs.
4) Let C_h be the set of the cliques which contains only hubs.
5) Let C_e be the set of extended cliques built in the following way: expand all $c_h \in C_h$ with the cliques containing at least 2 common vertices with c_h, that is with $c \in C$ where $|c_h \cap c| \geq 2$. Let c_e be the subgraph of the expanded vertices.
6) Merge every $c_{e_0}, c_{e_1} \in C_e$ if they have at least x common hubs.
7) The given C_e set contains the communities of the network.

B. Hub Selection

The third step of the algorithm introduces a Hub Selection strategy, which defines how vertices are chosen as a hubs. The hub selection strategies are the following:

- **Median of 1 neighborhood:** A vertex v is hub, if the value of h_v is greater than the median of the h_v values of its neighbors.
- **Mean of 1 neighborhood with parameter**: A vertex v is hub, if the value of h_v is greater than the average of the h_v values of the neighbors, multiplied by a $q > 0$ parameter.
- **Weighted mean of 1 neighborhood:** The value of the h_v is multiplied by the weights on the out-edges. A vertex

v is hub, if the value of the computed h_v is bigger than the average of the h_v values in one neighborhood.

The third strategy was changed compared to the original, emphasizing the direction of the edges, because a hub is better if it has more out edges. In our experience, it improves the quality of the output, if the hub selection strategy contains information about the edge weights.

III. Independent Cascade Model

There are numerous models of infection spreading in the literature, and these models were adopted to many different scientific fields including epidemics, sociology and economics. The two models most relevant to this paper were proposed in [8] by Domingos and Richardson and [7] by Granovetter. The former was used to improve the efficiency of virus marketing, the latter was the first method used to model the spreading of behavior. These models were later adopted to networks in [1], [9] by Kleinberg and Kempe. The infection model discussed in this paper is the Independent Cascade Model. The rest of this section describes this model in detail.

Let $G = (V, E)$ be a directed network, where $\forall(v, u) \in E$ edge has a $p(v, u)$ probability where $0 < p(v, u) \leq 1$. We assign states to the nodes: they are either susceptible, infected or removed. Let A_0 be the initial infected set of nodes $A_0 \subset V(G)$, all other nodes are susceptible at the beginning. The infection process takes place in discrete time steps or iterations. Through the iterations let A_i denote the set of the nodes which become infected in the i-th iteration. Each node stays infected for exactly one iteration, afterwards it is removed from the process. The process terminates in finite steps, and let A denote the set of removed nodes at the end of the process. In each iteration each infected node may make one attempt to infect its susceptible neighbors according to the value $p(v, u)$ on the edge connecting them. Algorithm 1 summarizes the Independent Cascade Model.

Algorithm 1 Independent Cascade

```
Let A_0 denote the set of initially infected nodes
While A_i ≠ ∅
    A_i ← newly infected nodes
    ∀v ∈ A_i tries to infect their neighbors with p(v, u)
    If the infection is successful
        A_{i+1} = A_{i+1} ∪ u
    End If
End While
```

If the A_i set is empty the infection process stops. Let $\sigma(A_0)$ denote the expected number of infected nodes with A_0 as the initial set. Let $w_f(v)$ be the final infection probability of a given node. The value of the $\sigma(A_0)$ formally is the following:

$$\sigma(A_0) = \sum_{v \ in \ G(V)} w_f(v) \tag{1}$$

There are numerous examples in the literature to compute the $\sigma(A_0)$ [3], [11]. The exact computation of $\sigma(A_0)$ is a #P-Complete problem [21].

A. Complete Simulation

In this paper the complete simulation algorithm proposed in [3], [1] is used to compute the expected number of infected vertices. A generalized version of the model can be found in [3]. In Complete Simulation algorithm sample size is an important parameter because it sets the number of independent simulations, as such the precision of the result. The Complete Simulation algorithm for the Independent Cascade Model is shown on Algorithm 2.

Algorithm 2 Complete Simulation

```
Input: Graph G, sample size s
A_0 ← initially infected nodes
j ← 0
∀v ∈ G(V) : f_v = 0
While j < s
    ∀e ∈ G(E) let the edge active or passive based on p(e)
    Modified DFS from ∀v ∈ A_0
    If n ∈ G(V) node is accessible from v ∈ A_0
        n : f_v ← f_v + 1
    End If
j ← j + 1
End While
∀v ∈ G(V) : fv ← fv / s
```

The simulation generates s different networks, each having different, randomized edge infection probabilities, and in every independent simulation every edge is either in an active or a passive state. The modified Depth First Search uses only active edges to visit the nodes, and increases the f_v values of the nodes if they are visited in the simulation instance. Finally, the f_v values are divided by the number of the independent simulations, which gives us an expected value for every node. In this paper complete simulation is used to get the $\sigma(A_0)$ value for a given initially infected set.

B. Infection maximization

The infection maximization problem is an optimization problem where the main objective is to maximize the spread of infection in the network. The problem is to find the set of k initial infectors which give the maximal expected infection, so in other words we are looking for an A_0 vertex set for any $|A_0| = k$ which maximizes the value of $\sigma(A_0)$.

To try different varieties of these sets, several repeated computations of the simulation is needed. If we want to try all possible initially infected sets for $k = 2$ of the example on Figure 2 we need 56 different simulations for this small network, but in a real-sized network it is not computable in acceptable time. The original infection maximization problem was published by Kempe et. al [1]. In the same paper they have proven the NP-hardness of the problem, and gave a greedy optimization method which can give at least 63% of the optimum for any case.

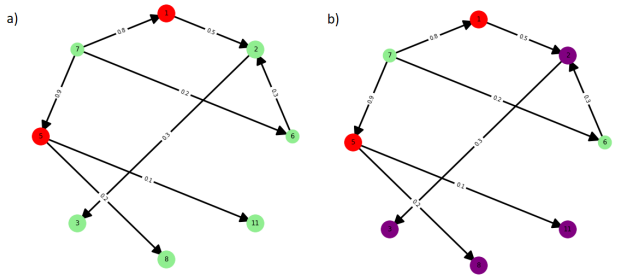


Fig. 2. On figure a) the $A_0 = \{1, 5\}$ so nodes 1 and 5 are infected initially and $k = 2$. The figure b) shows the result of the simulation with samplesize of 100 000. The red nodes are the initially infected nodes, the purple nodes have greater infection than zero, and the green nodes are uninfected. In this example $\sigma(A_0) = 2.94546$.

C. Greedy method

The greedy method starts from an empty set and increases the number of the initially infected nodes until it reaches the given k. In every iteration the algorithm chooses the node that currently seems to be the best choice. The algorithm does not give the optimal solution but it has a guaranteed precision of 63% of the optimum, but in most real-life cases it gives much better solution. At first the algorithm chooses the most infectious node from the network which can maximize the spread alone in the most efficient way. After that in every iteration one node is added to A_0 which gives the greatest improvement of the spread of infection with the other selected nodes. In the end, the algorithm gives an infected node set which maximizes the expected value of the infection. Algorithm 3 shows the greedy method.

Algorithm 3 Greedy method

1: **Input:** Graph G, size of the infected set k
2: **Output:** A_0 infected set
3: $A_0 \leftarrow \emptyset$
4: **While** $|A_0| \leq k$
5: $\quad A_0 = A_0 \cup \ arg\ max_{v \in G(V) \setminus A_0)} \sigma(A_0 \cup \{v\})$

In every iteration of the algorithm the next node is chosen from a $\{G(V) \setminus A_0\}$ set. The idea of the paper is to reduce the size of the set of the possible nodes so we will minimize the search space in every iteration based on some computed value which comes from a community detection algorithm.

IV. Reduction methods

The original greedy method gives us a quite good solution, but in real-sized networks the running time of the method can be too high. If the search space of the greedy method is reduced, it cannot guarantee the 63% precision of the optimal solution, but with a well chosen heuristic it can give a better solution. The main advantage of our methodology is the running time. In this section different reduction methods are demonstrated based on a computed value assigned to every node describing the quality of a node as an infector. We aim to improve the performance of the method by incorporating community-based information taken from one of the detection methods discussed above. Let $V*$ be the reduced selection set, and in every iteration the reduced greedy algorithm chooses from $\{G(V*) \setminus A_0\}$ resulting in decreased runtime. We give two different values based on the directed hub percolation method and the directed clique percolation methods. We introduce two different $f(v)$ functions which scores the nodes based on a different community or clique based statistic.

A. Hub Value

Cliques indicate the strongest connection between groups of nodes because in a clique every node is connected with each other. Let $f(v)$ $f : v \rightarrow Z$ be a function which assigns a number to every node. Let $f_{hv}(v)$ be a function that assigns the hub value h_v introduced in section 2 to the nodes of the network indicating how many directed cliques contain the node. The score is based on the idea, that a node can be a good infector if multiple cliques contain it, because in this way the node can spread the infection between cliques.

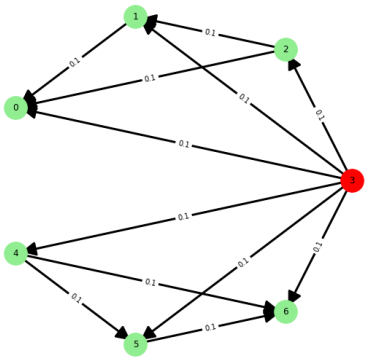


Fig. 3. Example of hub value calculation. The hub value of node 3 is $f_{hv}(3) = 2$ because two directed cliques contain the red node. All of the green nodes have one as a h_v. In this case the node 3 is a very good infector because it can spread the infection in both directed cliques.

After every nodes get the score, the $G(V)$ set is sorted according to h_v. The reduced $V*$ set contains the top nodes of the ordered $G(V)$ set. Since the hub value doesn't contain information on the edges of the graph, we introduce two different approaches.

- *unweighted hub value:* The nodes are sorted based only on h_v values.
- *mean weighted hub value:* The h_v is multiplied by the mean of the probabilities on every out-edge of the actual node.

Since the second technique contains information on the edge weights and the out degree of the node, it gives a higher score if a node is in many cliques, has many out edges, and the out probabilities are high. If the network is undirected, the method can be more efficient because an undirected clique indicates a stronger connection than a directed.

B. Community Value

The second technique can be based on the results of different community detection methods, providing the ability to compare these methods. In this case the score for a given node is how many communities contain the actual node. Every overlapping community detection method can be compared using this methodology providing a comprehensive community comparing technique.

The basic idea is the same in the previous section, the difference is in the $f(v)$ function. Let c_v be the community value and let the $f_{cv}(v)$ $f : v \to Z$ be a function which scores the nodes based on their community values. The reduced set works in the same way as in the case of hub values. The communities in real life have additional meanings: they can group the nodes into different sets, but if the main objective is infection maximization, a node can be a good infector if it is a member of many communities. The nodes with large community values can work as an infection bridges between different communities, since in real life a person or a company can be a good infector if it appears in many different areas of life.

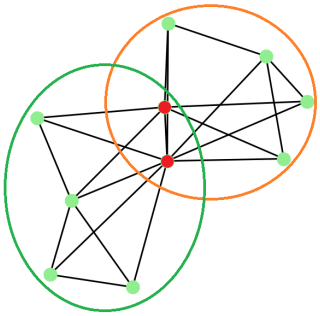


Fig. 4. Community values in overlapping communities. Since two different communities contain the red nodes, the community value of the red nodes is 2. The community value of the green node is 1.

Figure 4 shows an example of the community value. Considering only the individual communities, the red nodes and the green nodes are the same because they are just simple members of the group. From a global viewpoint however, the green nodes can be less effective in disease spreading because they are only connected to nodes inside their communities. The red nodes have access to both communities. The community value can also be used in two ways based on the computation of the value.

- *unweighted community value:* The c_v values denote how many communities contain the node.
- *mean weighted community value:* The c_v values are multiplied by the mean of the probabilities on every out-edge from v.

The nodes with zero or low out degrees are also eliminated from the reduced set. In the results section, the above techniques were compared to the results of the original greedy method.

V. Results

In this section we present the results of our modified infection maximization method, and test our methodology which provides a way to compare different overlapping community detection techniques. We ran the algorithm on a PC with I7 4790 CPU (3.6 Ghz) and 16GB of RAM. The Complete Simulation and the optimization framework is implemented in Java. We tested our methodology on six different randomly generated and seven real-life networks. For the random networks we used the igraph package of the R language and for the figures we used a Python version of our framework. With the greedy method the sample size was set high to get the best precision.

A. Precision of the results

All test were run with $s = 1000$ in every iteration because in the greedy method the algorithm has to compute the expected value for all possible nodes which is very time-consuming. Higher s values do not give more precise results as presented in this section, but their computation takes much longer. At the end of the testing process the final solution was rerun with $s = 100000$. Results show that complete simulation has lower precision compared to the greedy method by 1.14% measured on the final set of infected nodes. Let $\sigma(A_0)_{greedy}$ be the expected value of the given infected set in the greedy method, and $\sigma(A_0)_{final}$ be the expected value of the infected set with a high precision complete simulation. Table I shows the precision loss of complete simulation compared to the greedy method on the final infection values ordered according to the density of the network.

TABLE I
PRECISION OF INFECTION MAXIMIZATION

Diff	Precision	Density
0.70245	1.148%	1.774
0.8183	1.082%	1.802
0.87047	0.403%	1.833
0.86507	0.905%	1.834
0.77251	0.545%	1.838
0.06035	0.161%	3.104
0.81857	0.654%	3.473
0.05284	0.252%	3.492
0.03828	0.038%	3.708
0.69613	0.562%	3.872
0.51405	0.550%	4.602
2.53454	0.270%	6.393
1.05984	0.158%	25.44

The column diff shows the difference between $\sigma(A_0)_{greedy}$ and $\sigma(A_0)_{final}$. The precision column denotes the percentage of loss compared to the expected value of the final infections. Results were computed on the random networks below.

B. Random networks

The random graphs were generated in 6 different sizes with the forest fire model [12]. The properties of the random networks are the following:

- Number of nodes from 250 to 1500
- Number of edges from 873 to 5205

- Forward burning probability was 0.34
- Edge probabilities were randomly drawn from an uniform distribution between 0 and 0.2

The test results of the original greedy algorithm, and the size of the networks are shown on Table II.

TABLE II
RESULTS OF THE ORIGINAL GREEDY ALGORITHM ON RANDOM NETWORKS

Graph	nodes	edges	k	Greedy	Time
rand_1	250	873	3	20.922	15.81s
rand_2	500	1552	5	37.338	84.31s
rand_3	750	3452	8	93.321	438.52s
rand_4	1000	3708	10	99.462	796.65s
rand_5	1250	4841	13	123.756	1704.85s
rand_6	1500	5205	15	125.044	2534.69s

During testing the size of the initial infected set was 1% of the number of nodes in every scenario. The randomly generated networks are not too big, just enough to show our concept works. Furthermore, a real-sized network has millions of nodes and edges or more and it is not possible to test the greedy algorithm on it due to its time complexity. The size of the reduced set $V*$ was 10% of the number of nodes. Table III shows the result of the greedy algorithm with the reduced $V*$ based on hub and community based methods. As the size of the networks increases, the running times follow the size of the reduced set. The time column shows that the running times are approximately 10% of the original.

TABLE III
RESULTS FOR THE HUB AND COMMUNITY BASED REDUCED SET ALGORITHM ON RANDOM NETWORKS. (**HV:** HUB VALUE, **DHP:** COMMUNITY VALUE BASED ON DIRECTED HUB PERCOLATION, **DCP**: COMMUNITY VALUE BASED ON DIRECTED CLIQUE PERCOLATION, **DIFF:** PERCENTAGE OF THE SOLUTION COMPARED TO THE GREEDY ALGORITHM, **TIME:** TIME OF THE SOLUTIONS COMPARED TO THE TIME OF THE GREEDY METHOD)

Graph	HV	Diff	DHP	Diff	DCP	Diff	Time
rand_1	20.87	99.7%	21.03	100.5%	4.65	22%	18%
rand_2	37.35	**100%**	37.67	**100.8%**	9.45	25%	13%
rand_3	93.07	99.7%	93.35	**100.1%**	26.45	28%	11%
rand_4	99.46	**100%**	100.25	**100.5%**	22.63	22%	11%
rand_5	124.06	**100%**	123.1	99.5%	31.15	25%	10%
rand_6	124.9	99.9%	121.4	97%	34.55	28%	10%

In four cases the hub or community value based reduced set method gives a similar or better solution than the original greedy algorithm. However in the rest of the cases it cannot reach the reference solution but it still gives acceptable results with a much better running time than the simple greedy method. If we compare our two community detection methods, the table shows that the directed hub percolation method gives much better solution than the directed clique percolation. The DHPM detects the overlapping nodes, and the strongly connected dense subgraphs better than the DCPM in these networks. Besides random networks we also tested our methodology on real-life networks.

C. Real networks

The first five real-life networks considered in this paper are word association graphs based on a survey connected to the website "Agykapocs.hu" created by László Kovács[17]. The nodes of these graphs are words and the edges are associations between the words based on the user answers. The different networks come from the different versions of the word-association network. The rest of the real-life networks are from a well known data set from Stanford University [13]. The first network from this data set is an email network which describes email connections of a large European research institution [14][15]. The second is the bitcoin alpha trust network which describes trust connections between bitcoin traders [16]. The edge weights of the network were generated in the same way as with the random networks: they were drawn from an uniform distribution between 0 and 0.2.

TABLE IV
RESULTS OF THE ORIGINAL GREEDY ALGORITHM ON REAL NETWORKS

Graph	nodes	edges	k	Greedy	Time
real_1	2751	5043	28	215.955	8969.39s
real_2	2088	3839	21	141.533	3868.12s
real_3	1680	3082	17	95.563	2005.33s
real_4	1460	2632	15	75.568	1298.91s
real_5	1305	2315	13	61.137	889.64s
email	1005	25571	10	670.187	5742.73s
bitcoin	3783	24186	38	936.535	83303,53s

We can find a lot of real-life networks larger than these, but according to Table IV even on these quite small networks the running times can be very high, indicating that the normal greedy algorithm may not be able to find a good solution especially with a high k parameter. The results of the reduced set algorithm are shown in Table V.

TABLE V
RESULTS FOR THE HUB AND COMMUNITY BASED REDUCED SET ALGORITHM ON REAL-LIFE NETWORKS.(**HV:** HUB VALUE, **DHP:** COMMUNITY VALUE BASED ON DIRECTED HUB PERCOLATION, **DCP**: COMMUNITY VALUE BASED ON DIRECTED CLIQUE PERCOLATION, **DIFF:** PERCENTAGE OF THE SOLUTION COMPARED TO THE GREEDY ALGORITHM, **TIME:** TIME OF THE SOLUTIONS COMPARED TO THE TIME OF THE GREEDY METHOD)

Graph	HV	Diff	DHP	Diff	DCP	Diff	Time
real_1	215.05	99%	215.8	**100%**	189.6	87%	10%
real_2	139.92	99%	138.6	98%	122.8	87%	10%
real_3	99.46	**104%**	97.78	**102%**	87.69	91%	10%
real_4	74.87	99%	74.34	98%	71.64	94%	10%
real_5	61.18	**100%**	59.54	97%	58.54	95%	11%
email	662.5	99%	662.4	99%	663.4	99%	10%
bitcoin	924.7	98%	916.4	98%	928.1	99%	11%

On real-life networks the results are not as pronounced as with the random networks. In three cases our method reaches very good solutions with a satisfactory running time, but the other solutions are still satisfactory. If we compare the two reduced set methods, we can say that the DCPM works much better on real-life networks. In the general case however, the DHPM still works better.

VI. CONCLUSION AND FUTURE WORK

In this paper we proposed a new community based infection maximization algorithm to reduce the running time of the

greedy algorithm of Kempe et al, allowing the application of the algorithm to real-life networks. The methodology also allows us to measure the quality of any overlapping community detection method. Our approach is based on a community or hub based $f(v)$ function that scores the nodes according to their ability to infect other nodes, and builds a reduced candidate set for the greedy method.

The main result of this paper is based on the hypothesis, that in real-life infections spread easier inside communities. Apart from the main result, the improved infection maximization method, we present a comparing methodology which can support the development of different high resolution community detection algorithms. In the future we want to improve the presented community-based approaches and try out different $f(v)$ functions to score the nodes. While our current methodology is based on and supports the greedy algorithm, we want to develop a clearly community based infection maximization method using the results of this paper.

ACKNOWLEDGMENT

László Hajdu acknowledges the support of the National Research, Development and Innovation Office - NKFIH Fund No. SNN-117879.

Miklós Krész acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the support of the EU-funded Hungarian grant EFOP-3.6.2-16-2017-00015.

REFERENCES

[1] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). ACM, New York, NY, USA, 137-146. DOI: 10.1145/956750.956769

[2] Santo Fortunato. Community detection in graphs. Physics Reports, 486(3-5):75-174, February 2010. ISSN 03701573. DOI: 10.1016/j.physrep.2009.11.002.

[3] A. Bóta, A. Pluhár, M. Krész: Approximations of the Generalized Cascade Model. Acta Cybernetica Volume 21 (2013) 37–51. DOI: 10.14232/actacyb.21.1.2013.4

[4] M. K. A. Bota. A high resolution clique based overlapping community detection algorithm for small world networks. Informatica, 39:177-186, 2015.

[5] G. Palla, I. Derényi, I. Farkas, T. Vicsek: Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435 (2005), pp. 814-818 DOI: 10.1038/nature03607

[6] G. Palla, et al., Directed network modules, New J. Phys. 9 (2007) 186. DOI: 10.1088/1367-2630/9/6/186

[7] M. Granovetter. Threshold models of collective behavior. Am. J. Sociol, 83:1420-1443, 1978. DOI: 10.1080/0022250X.1983.9989941

[8] P. Domingos, M. Richardson, Mining the Network Value of Customers, Proc. Seventh ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2001) 57-66. DOI: 10.1145/502512.502525

[9] D. Kempe, J. Kleinberg, E. Tardos, Influential Nodes in a Diffusion Model for Social Networks. Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP), Springer-Verlag (2005) 1127- 1138. DOI: 10.1007/11523468_91

[10] Wei Chen, Chi Wang and Yajun Wang, Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2010) 1029-1038. DOI: 10.1145/1835804.1835934

[11] Srivastava, Ajitesh and Chelmis, Charalampos and Prasanna, Viktor K. (2015) The unified model of social influence and its application in influence maximization. Social Network Analysis and Mining 5(1):66:1-66:15 DOI: 10.1007/s13278-015-0305-x

[12] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations. Proceedings of the1th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2005) 177-187 DOI: 10.1145/1081870.1081893

[13] Jure Leskovec and Andrej Krevl, SNAP Datasets: Stanford Large Network Dataset Collection, http://snap.stanford.edu/data, jun. 2014

[14] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. "Local Higher-order Graph Clustering." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. DOI: 10.1145/3097983.3098069

[15] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007. DOI: 10.1145/1217299.1217301

[16] S. Kumar, F. Spezzano, V.S. Subrahmanian, C. Faloutsos. Edge Weight Prediction in Weighted Signed Networks. IEEE International Conference on Data Mining (ICDM), 2016. DOI: 10.1109/ICDM.2016.0033

[17] Kovács, L., Conceptual Systems and Lexical Networks in the Mental Lexicon, (In Hungarian: Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban) Tinta Konyvkiadó, Budapest, 2013.

[18] Szabó Sándor, Zaválnij Bogdán Coloring the nodes of a directed graph Acta Universitatis Sapientiae Informatica 6:(6) pp. 117-131. (2014) DOI: 10.2478/ausi-2014-0021

[19] Szabó Sándor, Zaválnij Bogdán Coloring the edges of a directed graph Indian Journal of Pure & Applied Mathematics 45:(2) pp. 239-260. (2014) DOI: 10.1007/s13226-014-0061-z

[20] Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. 2010. Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). ACM, New York, NY, USA, 1039-1048. DOI: https://doi.org/10.1145/1835804.1835935

[21] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09). ACM, New York, NY, USA, 199-208. DOI: https://doi.org/10.1145/1557019.1557047

[22] Yuchen Li, Ju Fan, Yanhao Wang, Kian-Lee Tan. 2018. Influence Maximization on Social Graphs: A Survey. IEEE Transactions on Knowledge and Data Engineering PP(99):1-1 DOI: 10.1109/TKDE.2018.2807843

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 245–248
DOI: 10.15439/2018F29
ISSN 2300-5963 ACSIS, Vol. 15

Feature Selection in Time-Series Motion Databases

Florian Elain*, Antonio Mucherino†, Ludovic Hoyet‡, Richard Kulpa§

*INSA, Univ Rennes, Rennes, France. Email: florian.elain@gmail.com
†IRISA, Univ Rennes, Rennes, France. Email: antonio.mucherino@irisa.fr
‡INRIA, Univ Rennes, Rennes, France. Email: ludovic.hoyet@inria.fr
§M2S, Univ Rennes, Rennes, France. Email: richard.kulpa@irisa.fr

***Abstract*—The selection of relevant features in large databases is one of the most important and challenging problems in data mining. Samples forming a given database are generally described by a predefined set of features, and the situation where not all such features can be used for classification purposes needs very often to be faced in real applications. This situation is very typical when the database is related to a phenomenon whose characteristics are not well known. In this context, the extraction of relevant features can therefore also provide additional information on the studied phenomena. We tackle the feature selection problem from an optimization point of view, by reducing it to the problem of finding a maximal consistent "clustering" grouping together the samples and the features of the database. In this work, we extend this approach to dynamical databases, where features are not represented by only one real value, but they are rather given as sequences of a predefined number of real values. Our main contribution consists in proposing an alternative representation of the database so that it fits with a tridimensional matrix with no missing entries, from which a consistent triclustering can be obtained.**

I. Introduction

More and more attention is given nowadays to techniques for mining data, because of the growing amount of information that can be obtained from different resources and that needs to be analyzed [11]. The main aim of such techniques is to identify suitable partitions of a given set of data, where *similar* data can be grouped together. Such partitions can in fact help in finding important relationships in the original data. In some applications, a subset exists for which a classification of the data is already available; the classification associated to this subset can therefore be exploited for *learning* how to classify data for which a classification is not yet known. In this context, our work aims at looking for optimal selections of the features of a dataset in order to improve the quality of the performed classifications.

Let $\mathbb{S}$ be a set of n samples, where every $S_i \in \mathbb{S}$ is represented by an ordered set of m time-series Q_j^i. The number m of time-series per sample is fixed, whereas the length of every time-series can vary. We suppose that a classification of the samples S_i of $\mathbb{S}$, in a given number of classes, is available.

More formally, we suppose that every time-series Q_j^i is a sequence of ℓ_i real values $q_{j,k}^i$, with k counting from 1 to ℓ_i. The length of every time-series depends on the sample S_i, and, since all features of a sample are generally recorded at the same time, we can suppose, without losing generality, that ℓ_i is a constant for all time-series forming the same sample. In brief, we have:

$$\begin{array}{rl} \mathbb{S} = (S_1, S_2, \ldots, S_n) & \text{a set of samples,} \\ S_i = (Q_1^i, Q_2^i, \ldots, Q_m^i) & \text{ordered set of time-series,} \\ Q_j^i = (q_{j,1}^i, q_{j,2}^i, \ldots, q_{j,\ell_i}^i) & \text{time-series.} \end{array}$$

We consider the problem of selecting the subset of time-series that can better describe the phenomena under study. To this purpose, we propose a three-dimensional matrix representation of the original dataset $\mathbb{S}$ that is independent from the length of the time-series, and look for a consistent clustering in sub-matrices where the maximal number of time-series is preserved. Our approach finds its inspiration and extends some previous works (the reader is referred to [7] for a complete description) where non-dynamical problems were considered (every feature was represented by one real value per sample, and not by a time-series).

This short paper is organized as follows. In Section II, we will briefly recall previous works on static problems where the matrix representation of $\mathbb{S}$ is possible with a two-dimensional full matrix. In Section III, we will introduce our three-dimensional matrix representation for datasets where features are represented by time-series. In Section IV, we will propose an extension of the approach recalled in Section II to the data representation introduced in Section III. Finally, Section V will discuss on how to create datasets of human motions to be analyzed by the presented technique, and Section VI will conclude the paper.

II. Feature Selection by Consistent Biclustering

Feature selection is widely studied in the context of data mining. In case the samples of a given dataset do not have a temporal component, the feature selection problem can be tackled by consistent biclustering [7], [9], [10]. This approach works particularly well for problems where measurements are available for every sample, and where the number of features is generally larger than the number of samples in the dataset. The aim, in fact, is to select only important and relevant features from the dataset, whereas others may not be adequate for describing the samples. This gives two immediate consequences. First, if only pertinent features are used and all others are rejected, the memory space necessary for storing the data is optimized. Secondly, a strict relationship between

samples and features can this way be identified, which may reveal important information about the problem under study.

If a set of data contains n samples which are described by m features, then the dataset can be represented by a $m \times n$ matrix $\mathcal{A}$, where the samples are organized column by column, and the features are organized row by row. In this context, we refer to a *bicluster* of $\mathcal{A}$ as a sub-matrix of $\mathcal{A}$, whose elements are a subset of samples and features. Equivalently, a bicluster can be seen as a pair of subsets (S_r, F_r), where S_r is a class (or cluster) of samples, and F_r is a class (or cluster) of features. A *biclustering* [1] is a partition of $\mathcal{A}$ in p biclusters:

$$\mathbb{B} = \{(S_1, F_1), (S_2, F_2), \ldots, (S_p, F_p)\},$$

such that the following conditions are satisfied:

$$\bigcup_{r=1}^{p} S_r \equiv \mathcal{A}, \qquad S_\zeta \cap S_\xi = \emptyset \quad 1 \leq \zeta \neq \xi \leq p,$$

$$\bigcup_{r=1}^{p} F_r \equiv \mathcal{A}, \qquad F_\zeta \cap F_\xi = \emptyset \quad 1 \leq \zeta \neq \xi \leq p,$$

where $p \leq \min(n, m)$ is the number of biclusters.

If a classification for the samples of $\mathcal{A}$ is available, as well as a classification for its features, a biclustering $\mathbb{B}$ can be trivially constructed. Inversely, classifications of samples and features can be obtained from $\mathbb{B}$.

In some data mining applications, there exist sets of data for which a classification of its samples is already given: we say in this case that a *training set* is available. However, the classification of the features used for describing the samples is generally not known, or, equivalently, there is no biclustering $\mathbb{B}$ associated to this training set. Therefore, no a priori information about possible relationships between samples and features is in general given.

A way to obtain a classification for the features from a training set $\mathcal{A}$ is to assign each feature to the class where it is "mostly expressed" (see [7] for a wider discussion). This idea comes from the study of biclusterings related to gene expression data [4], but it can be applied as well to problems arising in other fields (see for example [8]). Once a classification for the features is obtained, a biclustering $\mathbb{B}$ for $\mathcal{A}$ can be computed by simply applying the definition of biclustering. If the found biclustering is *consistent* (in the sense given in [7] for the bidimensional case, the reader is referred to Section IV for additional details), then the selected features are most likely the ones that better describe the samples. In this work, this approach is extended in Section IV to consistent *tri*clusterings.

The feature selection problem can subsequently be formulated as a 0–1 linear fractional optimization problem, which was proved to be NP-hard [5]. We consider a bilevel reformulation of this optimization problem, whose inner problem is linear. For its solution, we employ a heuristic that is based on the meta-heuristic Variable Neighborhood Search (VNS) [3] where, at each iteration, the inner problem is solved exactly.

III. Constructing 3D comparison matrices

As stated in the Introduction, our focus in this work is on datasets whose samples S_i are described by a predefined number of time-series Q^i_j. As in the previous works on consistent biclustering, it is supposed that, for every sample S_i, the same number of time-series Q^i_j are available. Moreover, every pair of time-series Q^A_j and Q^B_j, sharing the same index j but belonging to two different samples, must be related to the same kind of information (e.g. we cannot compare angle variations with the concentration level of a chemical compound). These requirements, which basically ensure that the matrix representation of the biclustering has no missing entries in dimension 2, does not imply a similar property when working with time-series and clustering in 3D. In fact, while the number of samples and the number of features are two constants of the problem (the first two dimensions), the number of elements ℓ_i forming a time-series depends on the sample S_i. Therefore, the corresponding three-dimensional matrix may, in general, have missing entries. Moreover, elements $q^i_{j,k}$ sharing the same index k may have no relationship (whereas common index i means "same sample", and common index j means "same time-series", or equivalently "same feature").

Consider two samples A and B, and two homologous time-series Q^A_j and Q^B_j:

$$(q^A_{j,1}, q^A_{j,2}, \ldots, q^A_{j,\ell_A}), \qquad (q^B_{j,1}, q^B_{j,2}, \ldots, q^B_{j,\ell_B}).$$

In order to obtain a coherent three-dimensional matrix representation, we construct a new matrix where the entries represent comparison scores between pairs of time-series Q^i_j. We consider Dynamic Time Warping (DTW) for a global and temporal alignment of every pair of time-series (see for example [12]). Together with DTW, we also consider the more recent Correlation DTW (CoDTW) [2], which is able to perform better quality alignments in more difficult situations. From the original dataset $\mathbb{S}$, we can therefore compute a full three-dimensional matrix consisting of DTW scores between pairs of samples A and B, for a given feature j:

$$\mathrm{DTW}(A, B; j).$$

A graphical representation of this three-dimensional matrix is given in Fig. 1.

The rows of such a matrix (as well as its columns) contain all (Co)DTW values of one sample S_i in comparison with all the others, for a fixed set of homologous time-series. For this reason, it is reasonable to represent a sample S_i with either a row or a column of such a matrix. This three-dimensional matrix is the result of extending this sample representation to all sets of homologous time-series.

IV. Consistent triclustering

The matrix representation of the original dataset $\mathbb{S}$ that we propose consists of all scores obtained from the time-series comparisons (see previous section). Let DTW be the $n \times n \times m$ matrix containing all such scores. Once the binary vector x is

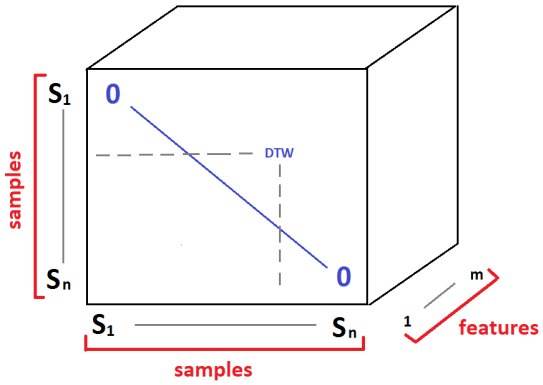


Fig. 1. Score comparison matrix obtained from the original dataset $\mathbb{S}$.

defined over the index set $\{1,2,\ldots,m\}$ so that

$$x_j = \begin{cases} 1 & \text{if the feature } j \text{ is selected} \\ 0 & \text{otherwise,} \end{cases}$$

we can define the sub-matrix DTW[x] obtained by removing from the matrix DTW all features j such that $x_j = 0$.

We suppose that a classification C_S for the samples of $\mathbb{S}$ in p classes is already available. Let $C_S(r)$, with $r \in \{1,2,\ldots,p\}$, indicate the subset of samples belonging to r^{th} class, and let s_{Ar} be a binary parameter indicating whether the sample S_A belongs to the class of samples r. A classification C_F for the homologous sets of time-series can be identified by applying the following rule. For a fixed $\hat{r} \in \{1,2,\ldots,p\}$, the homologous set indexed by $j \in \{1,2,\ldots,m\}$ is assigned to the $\hat{r}^{th}$ class if, and only if, by definition:

$$\forall \xi \in \{1,2,\ldots,p\} \,|\, \xi \neq \hat{r},$$
$$\sum_{A,B \in C_S(\hat{r}) \,|\, A \neq B} \frac{\mathrm{DTW}(A,B;j)}{|C_S(\hat{r})|} < \sum_{A,B \in C_S(\xi) \,|\, A \neq B} \frac{\mathrm{DTW}(A,B;j)}{|C_S(\xi)|}.$$

We suppose working on datasets for which the equation above cannot be satisfied with the equality, otherwise the classification of the features would not be unique.

Let f_{jr} be a binary parameter indicating whether the time-series with index j belongs to the class of features r. A *triclustering* of DTW[x] is *consistent* if

$$\forall \hat{r}, \xi \in \{1,\ldots,p\}, \hat{r} \neq \xi, \forall A,B \in C_S(\hat{r}), A \neq B$$
$$\frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} < \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j}. \quad (1)$$

It is important to remark that the matrix DTW does not admit, in general, a consistent triclustering if all features are considered. Notice that, differently from the previous works, we are interested here in the *less expressed* scores, because they correspond to time-series showing higher similarities.

The problem of selecting the relevant features by consistent triclustering can be stated as follows:

$$\max_x \left(f(x) = \sum_{j=1}^{m} x_j \right)$$
$$\text{subject to} \quad \forall \hat{r}, \xi \in \{1,\ldots,p\}, \hat{r} \neq \xi, \forall A,B \in C_S(\hat{r}), A \neq B$$
$$\frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} < \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j}. \quad (2)$$

As already pointed out, we consider the bilevel reformulation proposed in [7], and we solve the problem by employing a VNS-based heuristic. To perform such a reformulation, we transform the denominators of the optimization problem constraints (see equ.(2)) into continuous variables $y_r, r = 1,\ldots,p$, where y_r represents the number of selected features in the feature class $C_F(r)$:

$$\forall r \in \{1,\ldots,p\}, \; y_r = \sum_{j=1}^{m} f_{jr} x_j.$$

Using the newly introduced variables, the constraint in the original optimization problem can be rewritten by replacing $\sum_{j=1}^{m} f_{j\hat{r}} x_j$ and $\sum_{j=1}^{m} f_{j\xi} x_j$ by $y_{\hat{r}}$ and y_ξ, respectively. We normalize the values:

$$\overline{y}_r = \frac{\sum_{j=1}^{m} f_{jr} x_j}{m},$$

so that the following constraint is satisfied:

$$\sum_{r=1}^{p} \overline{y}_r \leq 1.$$

Our bilevel program is therefore:

$$\textbf{outer pb} \begin{cases} \min_{\overline{y}} \left(g(x,\overline{y}) = \sum_{r=1}^{p} \left[(1 - \overline{y}_r) + \sum_{\xi=1:\xi\neq r}^{p} c(x,r,\xi) \right] \right) \\ \text{subject to} \\ \textbf{inner pb} \begin{cases} x = \arg\max_x \left(f(x) = \sum_{j=1}^{m} x_j \right) \\ \text{subject to consistency constraint (1)} \end{cases} \\ \sum_{r=1}^{p} \overline{y}_r \leq 1, \end{cases}$$

where $c(x,\hat{r},\xi)$ is

$$\sum_{j \in C_S(\hat{r})} \left| \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\hat{r}} x_j}{\sum_{j=1}^{m} f_{j\hat{r}} x_j} - \frac{\sum_{j=1}^{m} DTW(A,B,j) f_{j\xi} x_j}{\sum_{j=1}^{m} f_{j\xi} x_j} \right|_+,$$

with $|\cdot|_+$ denoting the function that returns its argument if positive, and 0 otherwise. Hence, $c(x,\hat{r},\xi)$ is strictly positive if and only if at least one constraint is not satisfied.

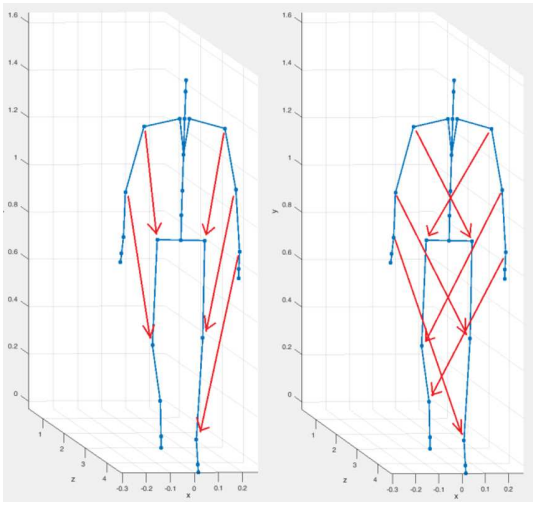


Fig. 2. A graphical representation of some newly introduced features aiming at measuring the symmetries in the movements (see distance sets represented by the red arrows).

V. Creating a Dataset of Human Motions

Motion capture makes it possible to track human movements over time. Markers are generally placed on the body surface of an actor, from which main bones and joint positions of the corresponding skeletal structure, over time, are derived. In general, absolute positions, together with bone rotation angles, are given for every frame of the motion capture (see skeletal representation in Fig. 2, in blue). The data are often stored in a specific format named BVH (BioVision Hierarchy), where joints and bones representing the actor are organized in a hierarchical way.

By collecting a certain number of captured human motions in BVH format, we can define a dataset of motions having particular properties (e.g. all motions are related to a certain human movement, but performed by different classes of humans, such as experts and novices, or male and female). This dataset can serve as a basis for our analyses, but it needs to be manipulated before its effective use.

In fact, the position of each skeleton joint in space may not give very useful information about the movements. Therefore, we propose to enrich the dataset with additional information as follows. For every motion, together with some information related to rotation angles between body parts (which can be easily extracted from BVH files), we also consider relative distances between joint pairs. Some recent studies, in fact, have shown that relative distances can play an important role in the representation of human motions [6]. As Fig. 2 shows, subset of distances can provide information about the symmetry of the movements. All these additional features are represented by time-series.

We have performed some very preliminary experiments where some relevant features were extracted from a so-constructed dataset of human motions (walking motions, with male and female actors) by using our optimization-based approach for feature selection by consistent triclustering. For lack of space, we cannot include any of them in this short paper. As for a future work, we will create a larger collection of motion datasets, having various properties, and we will use them to validate the theory presented in this paper.

VI. Conclusions

We extended an optimization-based approach to feature selection to datasets containing dynamical data. To do so, we proposed an alternative matrix representation of the data, where the original time-series are replaced by similarity scores. This made it possible to extend a previous approach for consistent biclustering to our new three-dimensional matrix representation.

Future works will be aimed at performing supervised classifications by exploiting the information that can be derived from obtained consistent triclusterings. Moreover, we plan to extend this approach to fuzzy sets, so that samples and features can actually belong to more than one class. It is our opinion that this would help better describing real phenomena, such as human motions.

References

[1] S. Busygin, O.A. Prokopyev, P.M. Pardalos, *Feature Selection for Consistent Biclustering via Fractional 0–1 Programming*, Journal of Combinatorial Optimization **10**, 7-21, 2005.

[2] S.A. Etemad, A. Arya, *Correlation-Optimized Time Warping for Motion*, The Visual Computer: International Journal of Computer Graphics **31**(12), 1569–1586, 2015.

[3] P. Hansen and N. Mladenovic, *Variable Neighborhood Search: Principles and Applications*, European Journal of Operational Research **130**(3), 449-467, 2001.

[4] L.-L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z. Weng, G.L. Mutter, M.P. Frosch, M.E. MacDonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, Gr. Stephanopoulos, Ge. Stephanopoulos, S.R. Gullans, *A Compendium of Gene Expression in Normal Human Tissues*, Physiological Genomics **7**, 97-104, 2001.

[5] O.E. Kundakcioglu, P.M. Pardalos, *The Complexity of Feature Selection for Consistent Biclustering*. In: Clustering Challenges in Biological Networks, S. Butenko, P.M. Pardalos, W.A. Chaovalitwongse (Eds.), World Scientific Publishing, 257–266, 2009.

[6] A. Mucherino, D.S. Gonçalves, A. Bernardin, L. Hoyet, F. Multon, *A Distance-Based Approach for Human Posture Simulations*, IEEE Conference Proceedings, Federated Conference on Computer Science and Information Systems (FedCSIS17), Workshop on Computational Optimization (WCO17), Prague, Czech Republic, 441–444, 2017.

[7] A. Mucherino, L. Liberti, *A VNS-based Heuristic for Feature Selection in Data Mining*. In: "Hybrid Meta-Heuristics", Studies in Computational Intelligence **434**, E-G. Talbi (Ed.), 353–368, 2013.

[8] A. Mucherino, A. Urtubia, *Consistent Biclustering and Applications to Agriculture*, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop on Data Mining in Agriculture (DMA10), Berlin, Germany, 105–113, 2010.

[9] A. Mucherino, P. Papajorgji, P.M. Pardalos, *Data Mining in Agriculture*, 274 pages, Springer, 2009.

[10] A. Mucherino, P.J. Papajorgji, P.M. Pardalos, *A Survey of Data Mining Techniques Applied to Agriculture*, Operational Research: An International Journal **9**(2), 121–140, 2009.

[11] G. Piatetsky-Shapiro, *Advances in Knowledge Discovery and Data Mining*. Usama M. Fayyad, Padhraic Smyth, Ramasamy Uthurusamy (Eds.), vol. 21. Menlo Park: AAAI press, 1996.

[12] X. Xi, E. Keogh, Ch. Shelton, L. Wei, C.A. Ratanamahatana, *Fast Time Series Classification Using Numerosity Reduction*, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 8 pages, 2006.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 249–255
DOI: 10.15439/2018F179
ISSN 2300-5963 ACSIS, Vol. 15

Computing Edit Distance between Rooted Labeled Caterpillars

Kohei Muraka
Graduate School of Computer Science and Systems Engineering
Kyushu Institute of Technology
Kawazu 680-4, Iizuka 820-8502, Japan
Email: muraka@dumbo.ai.kyutech.ac.jp

Takuya Yoshino, Kouichi Hirata*
Department of Artificial Intelligence
Kyushu Institute of Technology
Kawazu 680-4, Iizuka 820-8502, Japan
Email: {yoshino,hirata}@dumbo.ai.kyutech.ac.jp

***Abstract*—A *rooted labeled caterpillar* is a rooted labeled tree transformed to a path after removing all the leaves in it. In this paper, we design the algorithm to compute the edit distance between rooted labeled caterpillars in $O(\lambda^2 h^2)$ time, where λ and h are the maximum number of leaves and the maximum height in two caterpillars, respectively.**

I. Introduction

COMPARING tree-structured data such as HTML and XML data for web mining or RNA and glycan data for bioinformatics is one of the important tasks for data mining. The most famous distance measure [2] between *rooted labeled unordered trees* (*trees*, for short) is the *edit distance* [9]. The edit distance is formulated as the minimum cost of *edit operations*, consisting of a *substitution*, a *deletion* and an *insertion*, applied to transform a tree to another tree. It is known that the edit distance is always a metric and coincides with the minimum cost of *Tai mappings* [9].

Unfortunately, the problem of computing the edit distance between trees is MAX SNP-hard [13]. This statement also holds even if trees are binary or the maximum height of trees is at most 3 [1], [4].

Many variations of the edit distance have developed as more structurally sensitive distances, by introducing the restriction of Tai mappings (*cf.*, [7], [11]). All the variations except those of an alignment distance [5] are metrics and the problem of computing them is tractable [10], [11], [12], [14]. In particular, the *isolated-subtree distance* (or *constrained distance*) [12], which is defined as the minimum cost of isolated-subtree mappings, is the most general tractable variation of the edit distance [11].

On the other hand, a *caterpillar* (*cf.* [3]) is a tree transformed to a path after removing all the leaves in it. Whereas the caterpillars are very restricted and simple, there are some cases containing many caterpillars in real dataset, see Table II in Appendix.

As a method to compare two caterpillars, we can adopt a *complete subtree histogram distance*, which is an L_1-distance between histograms consisting of complete subtrees in two trees [1]. The complete subtree histogram is computable in linear time and always a metric but it is greater than the edit distance in general [1]. In particular, as an extreme case, there exists two caterpillars such that the edit distance between them is one but the complete subtree histogram distance is the number of nodes in two caterpillars, consider two paths with the same length such that the labels of leaves are different.

As another method, we can also adopt a *path histogram distance*, which is an L_1-distance between histograms consisting of paths from the root to leaves in two trees [6]. The path histogram distance is computable in linear time and always a metric for caterpillars, which is not a metric for trees, but it is incomparable with the edit distance [6].

Since a caterpillar is an unordered tree, it remains open whether or not the problem of computing the edit distance between caterpillars is tractable. Hence, we discuss this problem.

First, we point out that there exists a Tai mapping between two caterpillars that is not an isolated-subtree mapping. Then, we cannot apply the algorithm to compute the isolated-subtree distance or its variations [10], [11], [12], [14] that are tractable variations of the edit distance, to compute the edit distance between caterpillars.

On the other hand, a caterpillar has the structural property that the children of a non-leaf node in a caterpillar consist of at most one caterpillar and leaves (possibly empty). Then, by deleting a non-leaf node in a caterpillar, we obtain at most one caterpillar and the set of leaves as a forest. Furthermore, once such leaves are obtained, then we can add them to the previous set of leaves.

Based on this property, in this paper, we design the algorithm to compute the edit distance between caterpillars in $O(\lambda^2 h^2)$ time, where λ and h are the maximum number of leaves and the maximum height in two caterpillars, respectively. Furthermore, we point out that the structural restriction of caterpillars provides the limitation of tractable computing of the edit distance for unordered trees.

II. Caterpillars and Edit Distance

A *tree* T is a connected graph (V, E) without cycles, where V is the set of vertices and E is the set of edges. We denote V and E by $V(T)$ and $E(T)$. The *size* of T is $|V|$ and denoted by $|T|$. We sometime denote $v \in V(T)$ by $v \in T$. We denote an empty tree $(\emptyset, \emptyset)$ by $\emptyset$. A *rooted tree* is a tree with one

*The author would like to express thanks for support by Grant-in-Aid for Scientific Research 17H00762, 16H02870 and 16H01743 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

node r chosen as its *root*. We denote the root of a rooted tree T by $r(T)$.

Let T be a rooted tree such that $r = r(T)$ and $u, v, w \in T$. We denote the unique path from r to v, that is, the tree (V', E') such that $V' = \{v_1, \ldots, v_k\}$, $v_1 = r$, $v_k = v$ and $(v_i, v_{i+1}) \in E'$ for every i $(1 \leq i \leq k-1)$, by $UP_r(v)$. The *parent* of $v(\neq r)$, which we denote by $par(v)$, is its adjacent node on $UP_r(v)$ and the *ancestors* of $v(\neq r)$ are the nodes on $UP_r(v) - \{v\}$. We say that u is a *child* of v if v is the parent of u and u is a *descendant* of v if v is an ancestor of u. We call a node with no children a *leaf* and denote the set of all the leaves in T by $lv(T)$.

The *degree* of v, denoted by $d(v)$, is the number of children of v, and the *degree* of T, denoted by $d(T)$, is $\max\{d(v) \mid v \in T\}$. The *height* of v, denoted by $h(v)$, is $\max\{|UP_v(w)| \mid w \in lv(T[v])\}$, and the *height* of T, denoted by $h(T)$, is $\max\{h(v) \mid v \in T\}$.

We use the ancestor orders $<$ and $\leq$, that is, $u < v$ if v is an ancestor of u and $u \leq v$ if $u < v$ or $u = v$. We say that w is the *least common ancestor* of u and v, denoted by $u \sqcup v$, if $u \leq w$, $v \leq w$ and there exists no node $w' \in T$ such that $w' \leq w$, $u \leq w'$ and $v \leq w'$.

Let T be a rooted tree (V, E) and v a node in T. A *complete subtree of T at v*, denoted by $T[v]$, is a rooted tree $T' = (V', E')$ such that $r(T') = v$, $V' = \{u \in V \mid u \leq v\}$ and $E' = \{(u, w) \in E \mid u, w \in V'\}$.

We say that u is *to the left of* v in T if $pre(u) \leq pre(v)$ for the preorder number *pre* in T and $post(u) \leq post(v)$ for the postorder number *post* in T. We say that a rooted tree is *ordered* if a left-to-right order among siblings is given; *unordered* otherwise. We say that a rooted tree is *labeled* if each node is assigned a symbol from a fixed finite alphabet Σ. For a node v, we denote the label of v by $l(v)$, and sometimes identify v with $l(v)$. In this paper, we call a rooted labeled unordered tree a *tree* simply. Furthermore, we call a set of trees a *forest*.

As the restricted form of trees, we introduce a *rooted labeled caterpillar* (*caterpillar*, for short) as follows, which this paper mainly deals with.

Definition 1 (Caterpillar (cf., [3])): We say that a tree is a *caterpillar* if it is transformed to a path after removing all the leaves in it. For a caterpillar C, we call the remained path a *backbone* of C and denote it by $bb(C)$.

Next, we introduce an *edit distance* and a *Tai mapping*.

Definition 2 (Edit operations [9]): The *edit operations* of a tree T are defined as follows, see Figure 1.

1) *Substitution*: Change the label of the node v in T.
2) *Deletion*: Delete a node v in T with parent v', making the children of v become the children of v'. The children are inserted in the place of v as a subset of the children of v'. In particular, if v is the root in T, then the result applying the deletion is a forest consisting of the children of the root.
3) *Insertion*: The complement of deletion. Insert a node v as a child of v' in T making v the parent of a subset of the children of v'.

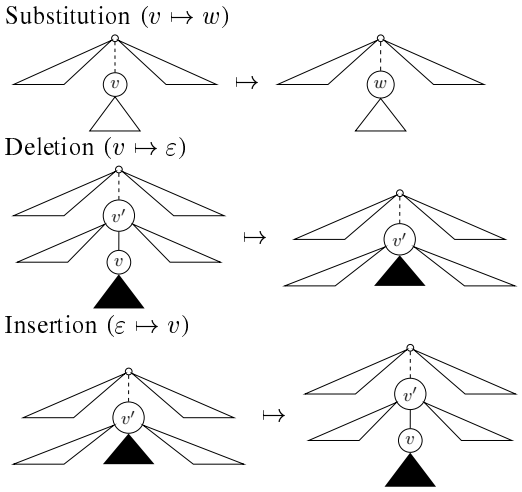


Fig. 1. Edit operations for trees.

Let $\varepsilon \notin \Sigma$ denote a special *blank* symbol and define $\Sigma_\varepsilon = \Sigma \cup \{\varepsilon\}$. Then, we represent each edit operation by $(l_1 \mapsto l_2)$, where $(l_1, l_2) \in (\Sigma_\varepsilon \times \Sigma_\varepsilon - \{(\varepsilon, \varepsilon)\})$. The operation is a substitution if $l_1 \neq \varepsilon$ and $l_2 \neq \varepsilon$, a deletion if $l_2 = \varepsilon$, and an insertion if $l_1 = \varepsilon$. For nodes v and w, we also denote $(l(v) \mapsto l(w))$ by $(v \mapsto w)$. We define a *cost function* $\gamma : (\Sigma_\varepsilon \times \Sigma_\varepsilon \setminus \{(\varepsilon, \varepsilon)\}) \mapsto \mathbf{R}^+$ on pairs of labels. We often constrain a cost function γ to be a *metric*, that is, $\gamma(l_1, l_2) \geq 0$, $\gamma(l_1, l_2) = 0$ iff $l_1 = l_2$, $\gamma(l_1, l_2) = \gamma(l_2, l_1)$ and $\gamma(l_1, l_3) \leq \gamma(l_1, l_2) + \gamma(l_2, l_3)$. In particular, we call the cost function that $\gamma(l_1, l_2) = 1$ if $l_1 \neq l_2$ a *unit cost function*.

Definition 3 (Edit distance [9]): For a cost function γ, the *cost* of an edit operation $e = l_1 \mapsto l_2$ is given by $\gamma(e) = \gamma(l_1, l_2)$. The *cost* of a sequence $E = e_1, \ldots, e_k$ of edit operations is given by $\gamma(E) = \sum_{i=1}^{k} \gamma(e_i)$. Then, an *edit distance* $\tau_{\text{TAI}}(T_1, T_2)$ between trees T_1 and T_2 is defined as follows:

$$\tau_{\text{TAI}}(T_1, T_2) = \min \left\{ \gamma(E) \;\middle|\; \begin{array}{l} E \text{ is a sequence} \\ \text{of edit operations} \\ \text{transforming } T_1 \text{ to } T_2 \end{array} \right\}.$$

Definition 4 (Tai mapping [9]): Let T_1 and T_2 be trees. We say that a triple (M, T_1, T_2) is a *Tai mapping* (a *mapping*, for short) from T_1 to T_2 if $M \subseteq V(T_1) \times V(T_2)$ and every pair (v_1, w_1) and (v_2, w_2) in M satisfies the following conditions.

1) $v_1 = v_2$ iff $w_1 = w_2$ (one-to-one condition).
2) $v_1 \leq v_2$ iff $w_1 \leq w_2$ (ancestor condition).

We will use M instead of (M, T_1, T_2) when there is no confusion denote it by $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$.

Let M be a mapping from T_1 to T_2. Let I_M and J_M be the sets of nodes in T_1 and T_2 but not in M, that is, $I_M = \{v \in T_1 \mid (v, w) \notin M\}$ and $J_M = \{w \in T_2 \mid (v, w) \notin M\}$. Then, the *cost* $\gamma(M)$ of M is given as follows.

$$\gamma(M) = \sum_{(v,w) \in M} \gamma(v, w) + \sum_{v \in I_M} \gamma(v, \varepsilon) + \sum_{w \in J_M} \gamma(\varepsilon, w).$$

Trees T_1 and T_2 are *isomorphic*, denoted by $T_1 \equiv T_2$, if there exists a mapping $M \in \mathcal{M}_{\text{TAI}}(T_1, T_2)$ such that $I_M =$

$J_M = \emptyset$ and $\gamma(M) = 0$.

Theorem 1 (Tai [9]): $\tau_{\mathrm{TAI}}(T_1, T_2) = \min\{\gamma(M) \mid M \in \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)\}$.

Unfortunately, the following theorem is known for the problem of computing τ_{TAI}.

Theorem 2 ([1], [4], [13]): Let T_1 and T_2 be trees. Then, the problem of computing $\tau_{\mathrm{TAI}}(T_1, T_2)$ is MAX SNP-hard. This statement also holds even if both T_1 and T_2 are binary or the maximum height of T_1 and T_2 is at most 3.

Finally, we introduce an *isolated-subtree mapping* and an *isolated-subtree distance* as the variations of the Tai mapping and the edit distance.

Definition 5 (Isolated-subtree mapping and distance [12]): Let T_1 and T_2 be trees and $M \in \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)$. We say that M is an *isolated-subtree mapping*, denoted by $M \in \mathcal{M}_{\mathrm{ILST}}(T_1, T_2)$, if M satisfies the following condition for every $(v_1, w_1), (v_2, w_2), (v_3, w_3) \in M$:

$$v_3 < v_1 \sqcup v_2 \iff w_3 < w_1 \sqcup w_2.$$

Furthermore, we define an *isolated-subtree distance* $\tau_{\mathrm{ILST}}(T_1, T_2)$ as follow.

$$\tau_{\mathrm{ILST}}(T_1, T_2) = \min\{\gamma(M) \mid M \in \mathcal{M}_{\mathrm{ILST}}(T_1, T_2)\}.$$

It is obvious that $\mathcal{M}_{\mathrm{ILST}}(T_1, T_2) \subseteq \mathcal{M}_{\mathrm{TAI}}(T_1, T_2)$ and then $\tau_{\mathrm{TAI}}(T_1, T_2) \leq \tau_{\mathrm{ILST}}(T_1, T_2)$. In contrast to Theorem 2, the following theorem also holds.

Theorem 3 (cf., [10]): Let T_1 and T_2 be trees. Then, we can compute $\tau_{\mathrm{ILST}}(T_1, T_2)$ in $O(n^2 d)$ time, where $n = \max\{|T_1|, |T_2|\}$ and $d = \min\{d(T_1), d(T_2)\}$.

It is known that τ_{ILST} is the most general tractable variation of τ_{TAI} [11].

Example 1: Consider two caterpillars C_1 and C_2 illustrated in Figure 2. Then, M illustrated in Figure 2 is the optimum mapping between C_1 and C_2. Here, it holds that $M \notin \mathcal{M}_{\mathrm{ILST}}(C_1, C_2)$ and M is an alignable mapping corresponding to an alignment distance [7], [11]. Note that the problem of computing an alignment distance is MAX SNP-hard in general [5].

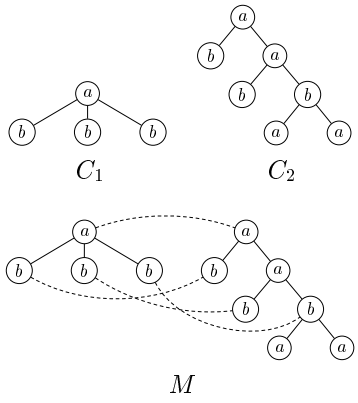


Fig. 2. Two caterpillars C_1 and C_2 and their optimum mapping M in Example 1.

Example 1 shows that there exists the minimum cost mapping between caterpillars which is not an isolated-subtree mapping. On the other hand, it remains open whether or not Theorem 2 holds for caterpillars.

Hence, in the next section, we discuss the problem of computing the edit distance between caterpillars and will solve this problem affirmatively.

III. Computing Edit Distance between Caterpillars

Example 1 also shows that we cannot apply the algorithm to compute τ_{ILST} and its variations [10], [11], [12], [14], which is based on the maximum cost maximum flow algorithm or the maximum weighted bipartite matching algorithm. On the other hand, by using the structural property of caterpillars, in this section, we design the algorithm to compute the edit distance between caterpillars.

Since every forest occurring in a caterpillar C is either obtained by deleting the path from the root to some internal node in $bb(C)$ or the complete subtree of $bb(C)$, a caterpillar C is one of the forms of $\{C\}$, $\{l_1, \ldots, l_k\}$ and $\{l_1, \ldots, l_k, C\}$, where l_i $(1 \leq i \leq k)$ is a leaf and C is a non-leaf caterpillar. By letting $L = \{l_1, \ldots, l_k\}$ and using a list representation of Prolog (*cf.*, [8]). we denote the above forests by $\langle \emptyset \,|\, C \rangle$, $\langle L \,|\, \emptyset \rangle$ and $\langle L \,|\, C \rangle$, respectively. In particular, we denote an empty forest $\langle \emptyset \,|\, \emptyset \rangle$ by Φ simply.

Let $C[v]$ be a caterpillar with the root v, where $L(v)$ denotes a (possibly empty) set of leaves as the children of v and $B(v)$ denotes at most one caterpillar of the child v. Then, $C[v]$ is one of the forms in Figure 3. Furthermore, by deleting v from $C[v]$, we obtain one of the forests of $\langle \emptyset \,|\, B(v) \rangle$, $\langle L(v) \,|\, \emptyset \rangle$ and $\langle L(v) \,|\, B(v) \rangle$, respectively.

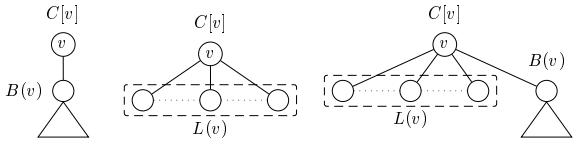


Fig. 3. The representation of a caterpillar $C[v]$.

Figure 4 illustrates the recurrences of computing the edit distance $\tau_{\mathrm{TAI}}(C_1[v], C_2[w])$ between two caterpillars $C_1[v]$ and $C_2[w]$, as $\delta_{\mathrm{TAI}}(\langle \emptyset \mid C_1[v] \rangle, \langle \emptyset \mid C_2[w] \rangle)$. Here, we denote the string representation of the set L of leaves under the alphabetical order on Σ by $s(L)$ and the string edit distance between two strings s_1 and s_2 [2] by $\sigma(s_1, s_2)$.

Theorem 4: The recurrences in Figure 4 are correct to compute the edit distance $\tau_{\mathrm{TAI}}(C_1[v], C_2[w])$ between $C_1[v]$ and $C_2[w]$ as $\delta_{\mathrm{TAI}}(\langle \emptyset \,|\, C_1[v] \rangle, \langle \emptyset \,|\, C_2[w] \rangle)$.

Proof: It is obvious that the recurrences in (A) compute the edit distance when at least one forest is empty and the recurrence in (B) computes the edit distance between two forests such that both of them consist of just leaves.

For the recurrence (C), let M be the minimum cost mapping between $\langle L_1 \mid \emptyset \rangle$ and $\langle L_2 \mid C_2[w] \rangle$. By focusing on w, M contains a pair of either (ε, w) or (v, w) for some $v \in L_1$. See Figure 5.

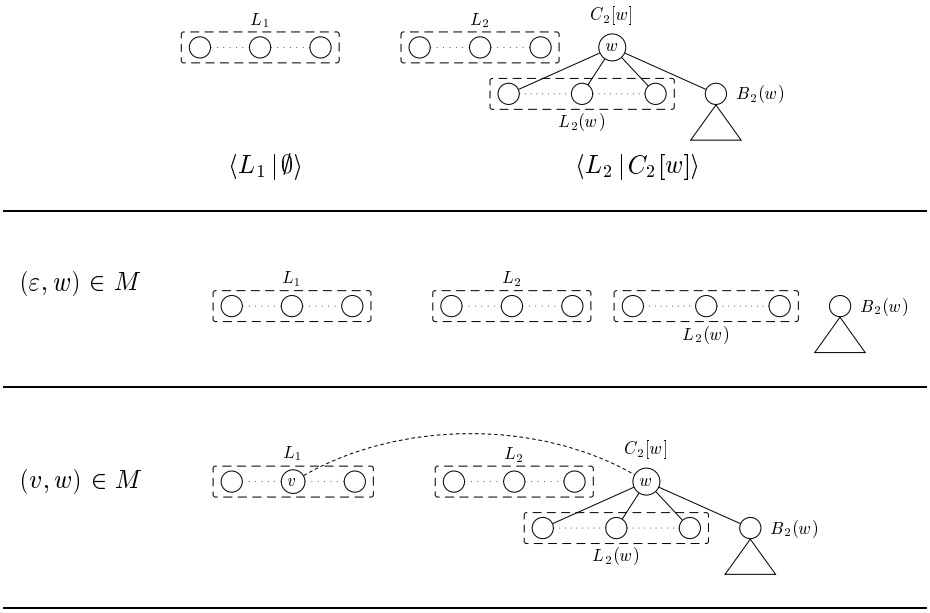


Fig. 5. $\langle L_1 \,|\, \emptyset \rangle$, $\langle L_2 \,|\, C_2[w] \rangle$ and the cases that $(\varepsilon, w) \in M$ and $(v, w) \in M$.

(A) $\delta_{\text{TAI}}(\langle L_1 \,|\, C_1 \rangle, \Phi) = \sum_{v \in L_1} \gamma(v, \varepsilon) + \sum_{v \in C_1} \gamma(v, \varepsilon)$.

$\delta_{\text{TAI}}(\Phi, \langle L_2 \,|\, C_2 \rangle) = \sum_{w \in L_2} \gamma(\varepsilon, w) + \sum_{w \in C_2} \gamma(\varepsilon, w)$.

(B) $\delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle) = \sigma(s(L_1), s(L_2))$.

(C) $\delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \,|\, C_2[w] \rangle)$

$$= \min \left\{ \begin{array}{ll} \gamma(\varepsilon, w) & \\ \quad + \delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \cup L_2(w) \,|\, B_2(w) \rangle), & (1) \\ \min\limits_{v \in L_1} \{ \gamma(v, w) + \delta_{\text{TAI}}(\langle L_1 \setminus \{v\} \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle) \} & \\ \quad + \delta_{\text{TAI}}(\Phi, \langle L_2(w) \,|\, B_2(w) \rangle) & (2) \end{array} \right\}.$$

(D) $\delta_{\text{TAI}}(\langle L_1 \,|\, C_1[v] \rangle, \langle L_2 \,|\, \emptyset \rangle)$

$$= \min \left\{ \begin{array}{ll} \gamma(v, \varepsilon) & \\ \quad + \delta_{\text{TAI}}(\langle L_1 \cup L_1(v) \,|\, B_1(v) \rangle, \langle L_2 \,|\, \emptyset \rangle), & (3) \\ \min\limits_{w \in L_2} \{ \gamma(v, w) + \delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \setminus \{w\} \,|\, \emptyset \rangle) \} & \\ \quad + \delta_{\text{TAI}}(\langle L_1(v) \,|\, B_1(v) \rangle, \Phi) & (4) \end{array} \right\}.$$

(E) $\delta_{\text{TAI}}(\langle L_1 \,|\, C_1[v] \rangle, \langle L_2 \,|\, C_2[w] \rangle)$

$$= \min \left\{ \begin{array}{ll} \gamma(v, w) + \delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle) & \\ \quad + \delta_{\text{TAI}}(\langle L_1(v) \,|\, B_1(v) \rangle, \langle L_2(w) \,|\, B_2(w) \rangle), & (5) \\ \gamma(v, \varepsilon) & \\ \quad + \delta_{\text{TAI}}(\langle L_1 \cup L_1(v) \,|\, B_1(v) \rangle, \langle L_2 \,|\, C_2[w] \rangle), & (6) \\ \gamma(\varepsilon, w) & \\ \quad + \delta_{\text{TAI}}(\langle L_1 \,|\, C_1[v] \rangle, \langle L_2 \cup L_2(w) \,|\, B_2(w) \rangle) & (7) \end{array} \right\}.$$

Fig. 4. The recurrences of computing the edit distance $\tau_{\text{TAI}}(C_1[v], C_2[w])$ between $C_1[v]$ and $C_2[w]$ as $\delta_{\text{TAI}}(\langle \emptyset \,|\, C_1[v] \rangle, \langle \emptyset \,|\, C_2[w] \rangle)$.

If $(\varepsilon, w) \in M$, then M maps nodes in $\langle L_1 \mid \emptyset \rangle$ to those in $\langle L_2 \cup L_2(w) \mid B_2(w) \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1 \mid \emptyset \rangle, \langle L_2 \cup L_2(w) \mid B_2(w) \rangle)$. Hence, the formula (1) computes the cost of M.

If $(v, w) \in M$ for some $v \in L_1$, then M maps nodes in $\langle L_1 \setminus \{v\} \mid \emptyset \rangle$ to those in $\langle L_2 \mid \emptyset \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1 \setminus \{v\} \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle)$. Since M is the minimum cost, it is necessary to minimize the value of $\gamma(v, \varepsilon) + \delta_{\text{TAI}}(\langle L_1 \setminus \{v\} \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle)$ for $v \in L_1$. Furthermore, once M contains (v, w) for some $v \in L_1$, M touches no descendants of w, that is, no nodes in $\langle L_2(w) \mid B_2(w) \rangle$, which is computed by $\delta_{\text{TAI}}(\Phi, \langle L_2(w) \mid B_2(w) \rangle)$. Hence, the formula (2) computes the cost of M.

The recurrence (D) is correct as same as the recurrence (C).

For the recurrence (E), let M be the minimum cost mapping between $\langle L_1 \,|\, C_1[v] \rangle$ and $\langle L_2 \,|\, C_2[w] \rangle$. By focusing on v and w, M contains one of the pairs of (v, w), (v, ε) and (ε, w). See Figure 6.

If $(v, w) \in M$, then M maps no nodes in $\langle L_1 \,|\, \emptyset \rangle$ to nodes in $\langle L_1(v) \,|\, B_1(v) \rangle$ and no nodes in $\langle L_2 \,|\, \emptyset \rangle$ to nodes in $\langle L_2(w) \,|\, B_2(w) \rangle$. Then, M maps nodes in $\langle L_1 \,|\, \emptyset \rangle$ to those in $\langle L_2 \,|\, \emptyset \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1 \,|\, \emptyset \rangle, \langle L_2 \,|\, \emptyset \rangle)$. Also M maps nodes in $\langle L_1(v) \,|\, B_1(v) \rangle$ to those in $\langle L_2(w) \,|\, B_2(w) \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1(v) \mid B_1(v) \rangle, \langle L_2(w) \mid B_2(w) \rangle)$. Hence, the formula (5) computes the cost of M.

If $(v, \varepsilon) \in M$, then M maps nodes in $\langle L_1 \cup L_1(v) \,|\, B_1(v) \rangle$ to those in $\langle L_2 \,|\, C_2[w] \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1 \cup L_1(v) \,|\, B_1(v) \rangle, \langle L_2 \mid C_2[w] \rangle)$. Hence, the formula (6) computes the cost of M.

If $(\varepsilon, w) \in M$, then M maps nodes in $\langle L_1 \mid C_1[v] \rangle$ to those in $\langle L_2 \cup L_2(w) \mid B_2(w) \rangle$, which is computed by $\delta_{\text{TAI}}(\langle L_1 \,|\, C_1[v] \rangle, \langle L_2 \cup L_2(w) \,|\, B_2(w) \rangle)$. Hence, the formula (7) computes the cost of M. ■

Example 2: Consider two caterpillars C_1 and C_2 in Figure 2 in Example 1. By applying the recurrences in Figure 4, we obtain that the edit distance $\tau_{\text{TAI}}(C_1, C_2)$ between C_1 and C_2 is 3 as follows. Here, we represent a caterpillar as a term-like representation, that is, $C_1 = a(b, b, b)$ and

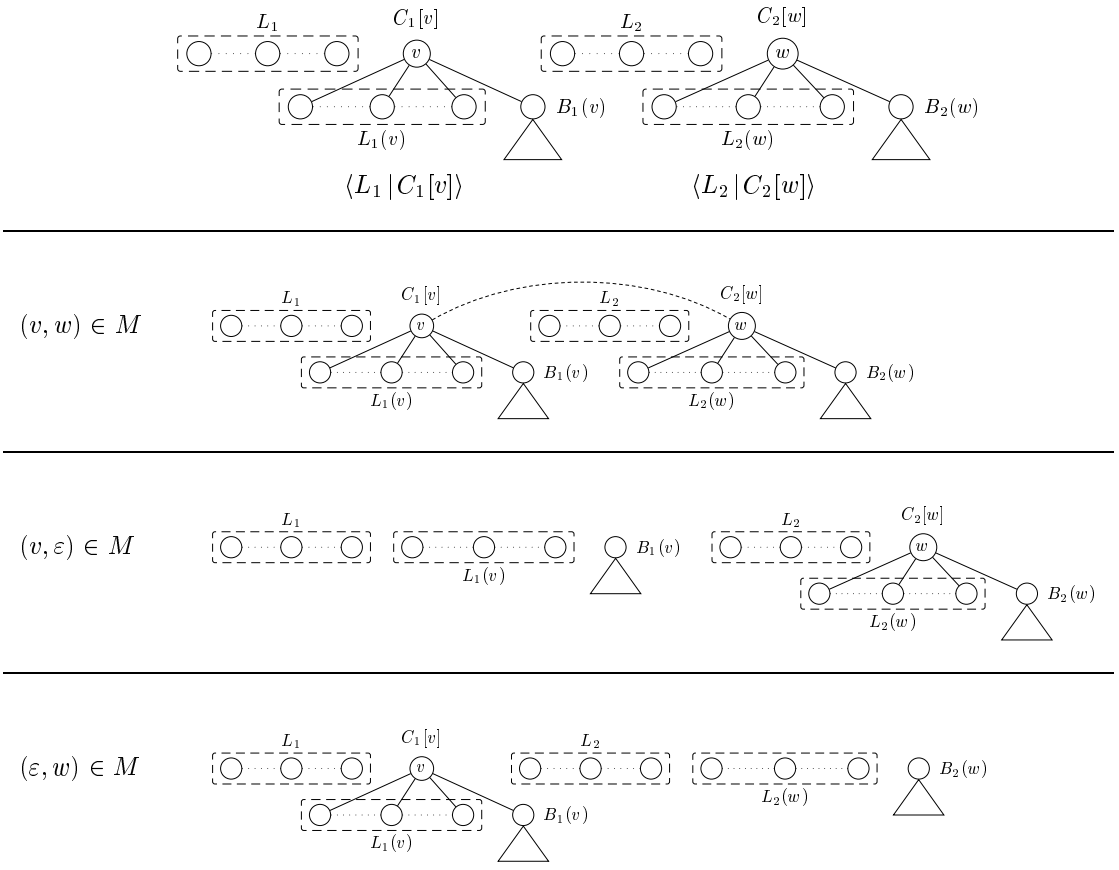


Fig. 6. $\langle L_1 \,|\, C_1[v] \rangle$, $\langle L_2 \,|\, C_2[w] \rangle$ and the cases that $(v, w) \in M$, $(v, \varepsilon) \in M$ and $(\varepsilon, w) \in M$.

$C_2 = a(b, a(b, b(a, a)))$.

$$
\begin{aligned}
& \tau_{\mathrm{TAI}}(C_1, C_2) \\
= \; & \delta_{\mathrm{TAI}}(\langle \emptyset \,|\, a(b,b,b) \rangle, \langle \emptyset \,|\, a(b, a(b, b(a, a))) \rangle) \\
= \; & \underbrace{\gamma(a,a)}_{=0} + \delta_{\mathrm{TAI}}(\langle \{b,b,b\} \,|\, \emptyset \rangle, \langle \{b\} \,|\, a(b, b(a,a)) \rangle) && \text{(E)(7)} \\
= \; & \underbrace{\gamma(\varepsilon,a)}_{=1} + \delta_{\mathrm{TAI}}(\langle \{b,b,b\} \,|\, \emptyset \rangle, \langle \{b,b\} \,|\, b(a,a) \rangle) && \text{(C)(2)} \\
= \; & 1 + \underbrace{\gamma(b,b)}_{=0} + \delta_{\mathrm{TAI}}(\langle \{b,b\} \,|\, \emptyset \rangle, \langle \{b,b\} \,|\, \emptyset \rangle) \\
& + \delta_{\mathrm{TAI}}(\Phi, \langle \{a,a\} \,|\, \emptyset \rangle\}) && \text{(C)(1)} \\
= \; & 1 + \underbrace{\sigma(bb,bb)}_{=0} + \underbrace{\gamma(\varepsilon,a)}_{=1} + \underbrace{\gamma(\varepsilon,a)}_{=1} && \text{(A),(B)} \\
= \; & 3.
\end{aligned}
$$

Hence, we can obtain the optimum mapping M between C_1 and C_2 illustrated in Figure 2, by collecting the pairs $(l_1, l_2) \in C_1 \times C_2$ such that $l_1 \neq \varepsilon$ and $l_2 \neq \varepsilon$ in $\gamma(l_1, l_2)$.

Let $C_1[v]$ and $C_2[w]$ be caterpillars. Then, we denote $bb(C_1[v])$ by a sequence $v_1, \ldots, v_n$ such that $v_n = v$ and $par(v_i) = v_{i+1}$ $(1 \leq i \leq n-1)$ and $bb(C_2[w])$ by a sequence $w_1, \ldots, w_m$ such that $w_m = w$ and $par(w_j) = w_{j+1}$ $(1 \leq j \leq m-1)$. In this case, we denote by $bb(C_1[v]) = [v_1, \ldots, v_n]$ and $bb(C_2[w]) = [w_1, \ldots, w_m]$. Also we use the same notations of $L_1(v_i)$ and $B_1(v_i)$ for $1 \leq i \leq n$ and $L_2(w_j)$ and $B_2(w_j)$ for $1 \leq j \leq m$.

Based on the recurrences in Figure 4, Algorithm 1 illustrates the algorithm to compute the edit distance $\tau_{\mathrm{TAI}}(C_1, C_2)$ between caterpillars C_1 and C_2. Here, the recurrence (A), (B), (C), (D) and (E) are corresponding to the lines 6 and 12, the line 3, the line 9, the line 15 and the line 19, respectively, in Algorithm 1.

Theorem 5: Let C_1 and C_2 be caterpillars. Then, we can compute the edit distance $\tau_{\mathrm{TAI}}(C_1, C_2)$ between C_1 and C_2 in $O(\lambda^2 h^2)$ time, where $\lambda = \max\{|lv(C_1)|, |lv(C_2)|\}$ and $h = \max\{h(C_1), h(C_2)\}$.

Proof: Let $bb(C_1) = [v_1, \ldots, v_n]$ and $bb(C_2) = [w_1, \ldots, w_m]$. Then, it is obvious that $h(C_1) = n + 1$ and $h(C_2) = m + 1$, so it holds that $m \leq h - 1$ and $n \leq h - 1$.

The algorithm $\tau_{\mathrm{TAI}}(C_1, C_2)$ in Algorithm 1 calls every pair $(v_i, w_j) \in bb(C_1) \times bb(C_2)$ just once. When computing $\delta_{\mathrm{TAI}}(\langle L_1(v_{i-1}) \,|\, C_1[v_i] \rangle, \langle L_2(w_{j-1}) \,|\, C_2[w_j] \rangle)$ for $2 \leq i \leq n$ and $2 \leq j \leq m$, it is necessary to construct the string representations $s_1 = s(L_1(v_1) \cup \cdots \cup L_1(v_{i-1}))$ and $s_2 = s(L_2(w_1) \cup \cdots \cup L_2(w_{j-1}))$ and compute the string edit

procedure $\tau_{\text{TAI}}(C_1, C_2)$
  /* C_1, C_2: caterpillars */
1 $\tau_{\text{TAI}}(C_1, C_2) \leftarrow \delta_{\text{TAI}}(\langle\emptyset\,|\,C_1\rangle, \langle\emptyset\,|\,C_2\rangle)$;
procedure $\delta_{\text{TAI}}(\langle L_1\,|\,C_1\rangle, \langle L_2\,|\,C_2\rangle)$
  /* L_1, L_2 : set of leaves, C_1, C_2: caterpillars */
2 **if** $C_1 = \emptyset$ **and** $C_2 = \emptyset$ **then**
3 $\delta_{\text{TAI}}(\langle L_1\,|\,\emptyset\rangle, \langle L_2\,|\,\emptyset\rangle) \leftarrow$ compute the recurrence (B);
4 **else if** $C_1 = \emptyset$ **and** $C_2 \neq \emptyset$ **then**
  /* $bb(C_2) = [w_1, \ldots, w_m]$, $w_m = r(C_2)$ */
5 **if** $L_1 = \emptyset$ **then**
6 $\delta_{\text{TAI}}(\Phi, \langle L_2\,|\,C_2\rangle) \leftarrow$ compute the recurrence (A);
  /* $\langle L_1\,|\,C_1\rangle = \Phi$ */
7 **else**
8 **for** $j = 1$ **to** m **do**
9 $\delta_{\text{TAI}}(\langle L_1\,|\,\emptyset\rangle, \langle L_2\,|\,C_2[w_j]\rangle) \leftarrow$ compute the recurrence (C);
10 **else if** $C_1 \neq \emptyset$ **and** $C_2 = \emptyset$ **then**
  /* $bb(C_1) = [v_1, \ldots, v_n]$, $v_n = r(C_1)$ */
11 **if** $L_2 = \emptyset$ **then**
12 $\delta_{\text{TAI}}(\langle L_1\,|\,C_1\rangle, \Phi) \leftarrow$ compute the recurrence (A);
  /* $\langle L_2\,|\,C_2\rangle = \Phi$ */
13 **else**
14 **for** $i = 1$ **to** n **do**
15 $\delta_{\text{TAI}}(\langle L_1\,|\,C_1[v_i]\rangle, \langle L_2\,|\,\emptyset\rangle) \leftarrow$ compute the recurrence (D);
16 **else**
  /* $bb(C_1) = [v_1, \ldots, v_n]$, $bb(C_2) = [w_1, \ldots, w_m]$, $v_n = r(C_1)$, $w_m = r(C_2)$ */
17 **for** $i = 1$ **to** n **do**
18 **for** $j = 1$ **to** m **do**
19 $\delta_{\text{TAI}}(\langle L_1\,|\,C_1[v_i]\rangle, \langle L_2\,|\,C_2[w_j]\rangle) \leftarrow$ compute the recurrence (E);

Algorithm 1: $\tau_{\text{TAI}}(C_1, C_2)$

distance $\sigma(s_1, s_2)$. The running time to construct the string representations is $O(\lambda \log \lambda)$ time (as same as that of sorting) and to compute the string edit distance is $O(\lambda^2)$ time [2].

Hence, the total running time of Algorithm 1 is described as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(2O(\lambda \log \lambda) + O(\lambda^2)) = O(\lambda^2)mn$$
$$\leq O(\lambda^2)(h-1)^2 = O(\lambda^2 h^2).$$

■

Theorem 5 also claims that the structural restriction of caterpillars provides the limitation of tractable computing the edit distance for unordered trees. We say that a tree is a *generalized caterpillar* if it is transformed to a caterpillar after removing all the leaves in it. Then, the following theorem also holds as corollaries in the proof of [1] or [4].

Theorem 6 (cf., [1], [4]): The problem of computing the edit distance between generalized caterpillars is MAX SNP-hard, even if the maximum height is at most 3.

Proof: It is straightforward from the proof of Corollary 4.3 in [1] or Theorem 1 in [4]. ■

Finally, Table I illustrates the number of pairs (C_1, C_2) of caterpillars such that $\tau_{\text{TAI}}(C_1, C_2) < \tau_{\text{ILST}}(C_1, C_2)$ and $\tau_{\text{TAI}}(C_1, C_2) = \tau_{\text{ILST}}(C_1, C_2)$, respectively, for all the pairs of 514 caterpillars in N-glycans (in Table II,

TABLE I
THE NUMBER OF PAIRS OF CATERPILLARS IN N-GLYCANS SUCH THAT $\tau_{\text{TAI}}(C_1, C_2) < \tau_{\text{ILST}}(C_1, C_2)$ AND $\tau_{\text{TAI}}(C_1, C_2) = \tau_{\text{ILST}}(C_1, C_2)$.

$\tau_{\text{TAI}} < \tau_{\text{ILST}}$		$\tau_{\text{TAI}} = \tau_{\text{ILST}}$	total
$\tau_{\text{ILST}} - \tau_{\text{TAI}} = 2$	$\tau_{\text{ILST}} - \tau_{\text{TAI}} = 1$		
5	1,218	130,618	131,841

Concerned with the 5 pairs in Table I, Figure 7 illustrates the caterpillars C_1 = G04187, C_2 = G00698, C_3 = G00933, C_4 = G01221, C_5 = G01454 and C_6 = G11051 in N-glycans such that $\tau_{\text{ILST}}(C_1, C_i) - \tau_{\text{TAI}}(C_1, C_i) = 2$ for $2 \leq i \leq 6$.

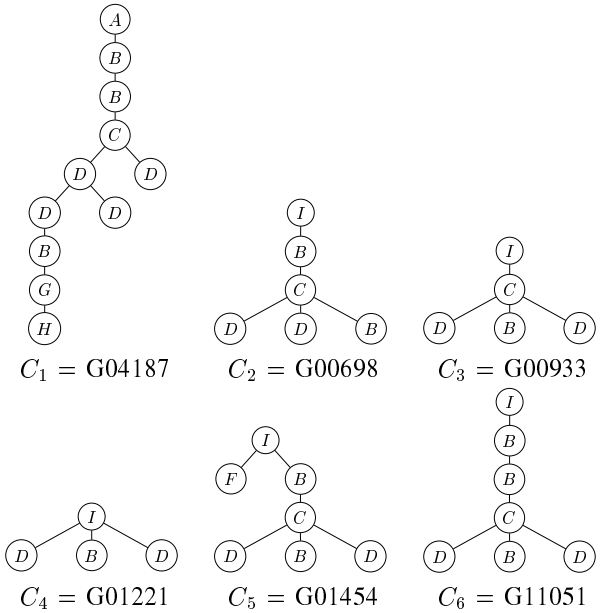


Fig. 7. The caterpillars C_1 = G04187, C_2 = G00698, C_3 = G00933, C_4 = G01221, C_5 = G01454 and C_6 = G11051.

Here, the following statements hold:

$$\tau_{\text{TAI}}(C_1, C_2) = 6,\ \tau_{\text{ILST}}(C_1, C_2) = 8,$$
$$\tau_{\text{TAI}}(C_1, C_3) = 7,\ \tau_{\text{ILST}}(C_1, C_3) = 9,$$
$$\tau_{\text{TAI}}(C_1, C_4) = 8,\ \tau_{\text{ILST}}(C_1, C_4) = 10,$$
$$\tau_{\text{TAI}}(C_1, C_5) = 7,\ \tau_{\text{ILST}}(C_1, C_5) = 9,$$
$$\tau_{\text{TAI}}(C_1, C_6) = 4,\ \tau_{\text{ILST}}(C_1, C_6) = 6.$$

Figure 8 illustrates the optimum mappings $M_1 \in \mathcal{M}_{\text{TAI}}(C_1, C_2)$ and $M_2 \in \mathcal{M}_{\text{ILST}}(C_1, C_2)$ for caterpillars C_1 and C_2 in Figure 7, which is the reason that $\tau_{\text{TAI}}(C_1, C_2) = 6$ (5 deleted nodes and 1 substituted node) and $\tau_{\text{ILST}}(C_1, C_2) = 8$ (6 deleted nodes, 1 inserted node and 1 substituted node).

IV. CONCLUSION AND FUTURE WORKS

In this paper, we have designed the algorithm to compute the edit distance between caterpillars in $O(\lambda^2 h^2)$ time, which

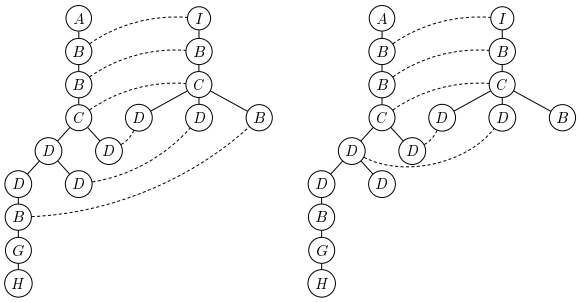


Fig. 8. The optimum mappings $M_1 \in \mathcal{M}_{\mathrm{TAI}}(C_1, C_2)$ and $M_2 \in \mathcal{M}_{\mathrm{ILST}}(C_1, C_2)$ for caterpillars C_1 and C_2 in Figure 7.

is the limitation of tractable computing the edit distance for unordered trees.

Whereas we have given a small experimental result in the last of Section III, it is necessary to implement Algorithm 1 more efficiently. Then, it is a future work to evaluate running time from all the data of caterpillars in Appendix by comparing that of the algorithm of computing τ_{ILST} [10], [11], [12] and to investigate the difference between τ_{TAI} and τ_{ILST}. Also, it is a future work to analyze the correlation for caterpillars in real data between the edit distance and the complete subtree histogram distance [1] or the path histogram distance [6].

Concerned with Theorem 6, it is a future work to give the strict limitation of tractable computing of the edit distance. In other words, it is a future work to investigate whether or not the problem of computing the edit distance between a caterpillar and a generalized caterpillar or a standard tree is tractable. In particular, concerned with D^- for $D \in \{\text{Auction, University, Protein, Nasa}\}$ in Table II in Appendix, it is a future work to investigate whether or not the problem of computing the edit distance between forests of caterpillars is tractable.

As the extension of the edit distance for rooted trees to that for unrooted trees, Zhang *et al.* [14] have extend the degree-2 distance for rooted trees to that for unrooted trees. In their algorithm, first we select a pair of nodes in unrooted trees, compute the degree-2 distance between the rooted trees whose pair of the roots is the selected pair and then select the minimum value of the distances as the degree-2 distance. It is a future work to investigate whether or not we can apply this idea to the problem of computing the edit distance between unrooted caterpillars and, if so, design the algorithm to compute it.

Appendix: Caterpillars in Real Data

In this appendix, we point out how large the number of caterpillars in real data. Table II illustrates the number of caterpillars in N-glycans and all glycans from KEGG[1], CSLOGS[2], dblp[3], and SwissProt, TPC-H, Auction, University, Protein and Nasa from UW XML Repository[4].

TABLE II
The number of caterpillars in N-glycans and all glycans from KEGG, CSLOGS, dblp, SwissProt, TPC-H, Auction, University, Protein and NASA.

dataset	#cat	#data	%
N-glycans	514	2,142	23.996
all glycans	8,005	10,704	74.785
CSLOGS	41,592	59,691	69.679
dblp	5,154,295	5,154,530	99.995
SwissProt	6,804	50,000	13.608
TPC-H	86,805	86,805	100.000
Auction	0	37	0
University	0	6,738	0
Protein	0	262,625	0
Nasa	0	2,430	0
Auction$^-$	259	259	100.000
University$^-$	74,638	79,213	94.224
Protein$^-$	1,874,703	2,204,068	85.057
Nasa$^-$	21,245	27,921	76.089

Here, #cat is the number of caterpillars and #data is the total number of data. Furthermore, for $D \in \{\text{Auction, University, Protein, Nasa}\}$, D^- denotes the trees obtained by deleting the root for every tree in D. Since one tree in D produces some trees in D^-, the total number of trees in D^- is greater than that of D.

References

[1] T. Akutsu, D. Fukagawa, M. M. Halldórsson, A. Takasu, K. Tanaka: *Approximation and parameterized algorithms for common subtrees and edit distance between unordered trees*, Theoret. Comput. Sci. **470**, 10–22 (2013).
[2] M. M. Deza, E. Deza: *Encyclopedia of distances* (4th ed.) Springer, 2016
[3] J. A. Gallian: *A dynamic survey of graph labeling*, Electorn. J. Combin. **14**, DS6 (2007).
[4] K. Hirata, Y. Yamamoto, T. Kuboyama: *Improved MAX SNP-hard results for finding an edit distance between unordered trees*, Proc. CPM'11, LNCS **6661**, 402–415 (2011).
[5] T. Jiang, L. Wang, K. Zhang: *Alignment of trees – an alternative to tree edit*, Theoret. Comput. Sci. **143**, 137–148 (1995).
[6] T. Kawaguchi, T. Yoshino, K. Hirata: *Path histogram distance for rooted labeled caterpillars*, Proc. ACIIDS'18, LNAI **10751**, 276–286 (2018).
[7] T. Kuboyama: *Matching and learning in trees*, Ph.D thesis, University of Tokyo (2007).
[8] L. S. Sterling, E. Y. Shapiro: *The art of Prolog* (2nd edition), The MIT Press (1994).
[9] K.-C. Tai: *The tree-to-tree correction problem*, J. ACM **26**, 422–433 (1979).
[10] Y. Yamamoto, K. Hirata, T. Kuboyama: *Tractable and intractable variations of unordered tree edit distance*, Internat. J. Found. Comput. Sci. **25**, 307–330 (2014).
[11] T. Yoshino, K. Hirata: *Tai mapping hierarchy for rooted labeled trees through common subforest*, Theory Comput. Sys. **60**, 759–783 (2017).
[12] K. Zhang: *A constrained edit distance between unordered labeled trees*, Algorithmica **15**, 205–222 (1996).
[13] K. Zhang, T. Jiang: *Some MAX SNP-hard results concerning unordered labeled trees*, Inform. Process. Lett. **49**, 249–254 (1994).
[14] K. Zhang, J. Wang, D. Shasha: *On the editing distance between undirected acyclic graphs*, Internat. J. Found. Comput. Sci. **7**, 45–58 (1996).

[1]Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/
[2]http://www.cs.rpi.edu/~zaki/www-new/pmwiki.php/Software/Software
[3]http://dblp.uni-trier.de/
[4]http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/www/repository.html

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 257–260
DOI: 10.15439/2018F121
ISSN 2300-5963 ACSIS, Vol. 15

A New Monte Carlo Algorithm for Linear Algebraic Systems Based on the "Walk on Equations" Algorithm

Venelin Todorov*†, Nikolay Ikonomov*, Ivan Dimov†, Rayna Georgieva†
*Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
8 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Email: vtodorov@math.bas.bg, nikonomov@math.bas.bg
†Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
Email: venelin@parallel.bas.bg, ivdimov@bas.bg, rayna@parallel.bas.bg

***Abstract*—A new Monte Carlo algorithm for solving systems of Linear Algebraic (LA) equations is presented and studied. The algorithm is based on the "Walk on Equations" Monte Carlo method recently developed by Ivan Dimov, Sylvain Maire and Jean Michel Sellier [4]. The algorithm is optimized by choosing the appropriate values for the relaxation parameters which leads to dramatic reduction in time and lower relative errors for a given number of iterations. Numerical tests are performed for examples with matrices of different size and on a system coming from a finite element approximation of a problem describing a beam structure in constructive mechanics.**

I. Introduction

MANY scientific and engineering applications are based on the problems of solving systems of LA equations. For some applications it is also important to compute directly the inner product of a given vector and the solution vector of a LA system. In Monte Carlo numerical algorithms we construct a Markov process and prove that the mathematical expectation of the process is equal to the unknown solution of the problem [3]. Iterative Monte Carlo algorithms can be defined as terminated Markov chains:

$$T = \{\alpha_{t_0} \to \alpha_{t_1} \to \alpha_{t_2} \dots \alpha_{t_k}\}, \tag{1}$$

where α_{t_q}, $q = 1, \dots, i$ is one of the absorbing states. By A and B we denote matrices of size $n \times n$, i.e., $A, B \in I\!R^{n\times n}$. We use the following presentation of matrices:

$$A = \{a_{ij}\}_{i,j=1}^{n} = (a_1, \dots, a_i, \dots, a_n)^t,$$

where $a_i = (a_{i1}, \dots, a_{in})$, $i = 1, \dots, n$ and the symbol t means *transposition*. The following norms of vectors:

$$\| b \| = \| b \|_1 = \sum_{i=1}^{n} |b_i|, \quad \| a_i \| = \| a_i \|_1 = \sum_{j=1}^{n} |a_{ij}|$$

The first author is supported by the Bulgarian National Science Fund under Projects DN 12/5-2017 and DN 12/4-2017 and by the Bulgarian Academy of Sciences through the Program for Career Development of Young Scientists, Grant DFNP-17-88/28.07.2017.

and matrices

$$\| A \|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|, \quad \| A \|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$$

are used, where $b \in \mathbb{R}^n$.

We consider a system of LA equations

$$Bx = f, \tag{2}$$

where $B = \{b_{ij}\}_{i,j=1}^{n} \in I\!R^{n\times n}$ is a given matrix; $f = (f_1, \dots, f_n)^t \in I\!R^{n\times 1}$ and $v = (v_1, \dots, v_n) \in I\!R^{1\times n}$ are given vectors.

We deal with the matrix $A = \{a_{ij}\}_{ij=1}^{n}$, such that $A = I - DB$, where D is a diagonal matrix $D = \mathrm{diag}(d_1, \dots, d_n)$ and $d_i = \frac{\gamma}{b_{ii}}$, $i = 1, \dots, n$, and $\gamma \in (0, 1]$ is a parameter that can be used to accelerate the convergence. The system (2) can be presented in the form of equation

$$x = Ax + b, \tag{3}$$

where $b = Df$. Let us suppose that the matrix B is diagonally dominant. It easily follows that if B is a diagonally dominant matrix, then the elements of the matrix A must satisfy the following condition: $\sum_{j=1}^{n} |a_{ij}| \leq 1$, $i = 1, \dots, n$.

A *stationary linear iterative algorithm* [3] can be used:

$$x_k = Ax_{k-1} + b, \quad k = 1, 2, \dots \tag{4}$$

and the solution x can be presented in a form of a Neumann series

$$x = \sum_{k=0}^{\infty} A^k b = b + Ab + A^2 b + A^3 b + \dots \tag{5}$$

The *stationary linear iterative Monte Carlo algorithm* is based on (5). As a result, the convergence of the Monte Carlo algorithm depends on the truncation error of the series (4) [3]. We are interested to evaluate the linear form $W(x)$ of the solution x of the system (3), i.e., $W(x) \equiv (w, x) = \sum_{i=1}^{n} w_i x_i$,

where $w \in I\!R^{n\times 1}$. We shall define a random variable $X[w]$, which expectation is equal to the above defined linear form, i.e., $EX[w] = W(x)$ using a discrete Markov process with a finite set of states. Then the problem is to determine repeated realizations of $X[w]$ and of connecting them into a proper statistical estimator of $W(x)$.

Consider an initial density vector $p = \{p_i\}_{i=1}^n \in I\!R^n$, such that $p_i \geq 0, i = 1, \ldots, n$ and $\sum_{i=1}^n p_i = 1$. Consider also a transition density matrix $P = \{p_{ij}\}_{i,j=1}^n \in I\!R^{n\times n}$, such that $p_{ij} \geq 0,\ i, j = 1, \ldots, n$ and $\sum_{j=1}^n p_{ij} = 1$, for any $i = 1, \ldots, n$.

We will be dealing with *permissible* densities [3] $\mathcal{P}_b$ and $\mathcal{P}_A$. It follows easily that in such a way the random trajectories constructed to solve the problems under consideration never visit zero elements of the matrix. Such an approach decreases the computational complexity of the algorithms [2]. It is also very convenient when large sparse matrices are used.

II. Probabilistic representation of the algorithm

Consider a real linear system of the form $x = Ax + b$ where the matrix A of size n is such that the convergence radius $\varrho(A) < 1$, its coefficients a_{ij} are real numbers and

$$\sum_{j=1}^{n} |a_{ij}| \leq 1, \forall 1 \leq i \leq n.$$

We now define a Markov chain T_k with $n + 1$ states $\alpha_1, \ldots, \alpha_n, \alpha_{n+1}$, such that

$$P(\alpha_{k+1} = j | \alpha_k = i) = |a_{ij}|$$

if $i \neq n + 1$ and

$$P(\alpha_{k+1} = n + 1 | \alpha_k = n + 1) = 1.$$

We also define a vector c such that $c(i) = b(i)$ if $1 \leq i \leq n$ and $c(n + 1) = 0$. Denote by $\tau = (\alpha_0, \alpha_1, \ldots, \alpha_k, \alpha_{n+1})$ a random trajectory that starts at the initial state $\alpha_0 < n + 1$ and passes through $(\alpha_1, \ldots, \alpha_k)$ until the absorbing state $\alpha_{k+1} = n + 1$. The probability to follow the trajectory τ is $P(\tau) = p_{\alpha_0} p_{\alpha_0\alpha_1} \ldots p_{\alpha_{k-1}\alpha_k} p_{\alpha_k}$. We use the MAO algorithm (see [1], [5]) for the initial density vector $p = \{p_\alpha\}_{\alpha=1}^n$ and for the transition density matrix $P = \{p_{\alpha\beta}\}_{\alpha,\beta=1}^n$, as well. The weights Q_α are defined:

$$Q_m = Q_{m-1} \frac{a_{\alpha_{m-1},\alpha_m}}{p_{\alpha_{m-1},\alpha_m}}, \quad m = 1, \ldots, k, \quad Q_0 = \frac{c_{\alpha_0}}{p_{\alpha_0}}. \quad (6)$$

The estimator $X_\alpha(\tau)$ can be presented as $X_\alpha(\tau) = c_\alpha + Q_k \frac{a_{\alpha_k\alpha}}{p_{\alpha_k}}, \quad \alpha = 1, \ldots n$ taken with a probability $P(\tau) = p_{\alpha_0} p_{\alpha_0\alpha_1}, \ldots p_{\alpha_{k-1,k}\alpha_k} p_{\alpha_k}$.

For the convergence of the process we use that the random variable $X_\alpha(\tau)$ is an unbiased estimator of x_α [4], i.e.

$$E\{X_\alpha(\tau)\} = x_\alpha. \quad (7)$$

Consider the variance of the random variable $X_\alpha(\tau)$ for evaluation the linear form for the solution $W(x)$. We use the following notations: $\overline{A} = \{|a_{ij}|\}_{i,j=1}^n$, $\widehat{c} = \{c_i^2\}_{i=1}^{n+1}$. The special choice of the probability densities leads to the Markov chain:

$$c_{\alpha_0} \to a_{\alpha_0\alpha_1} \to \ldots \to a_{\alpha_{k-1}\alpha_k}. \quad (8)$$

For this finite chain we have that

$$A_c^k = c_{\alpha_0} \prod_{s=1}^{k} a_{\alpha_{s-1}\alpha_s}, \quad (9)$$

where $c \in I\!R^{n\times 1}$ and $c(i) = b(i)$ if $1 \leq i \leq n$ and $c(n+1) = 0$. The variance of the random variable $X_\alpha^k(\tau)$ is defined as [4]

$$X_\alpha^k(\tau) = \frac{c_{\alpha_0}}{p_{\alpha_0}} \frac{a_{\alpha_0\alpha_1}}{p_{\alpha_0\alpha_1}} \frac{a_{\alpha_1\alpha_2}}{p_{\alpha_1\alpha_2}} \ldots \frac{a_{\alpha_{k-1}\alpha_k}}{p_{\alpha_{k-1}\alpha_k}} \frac{c_{\alpha_k}}{p_{\alpha_k}} = \frac{A_c^k c_{\alpha_k}}{P^k(\tau)}. \quad (10)$$

The variance of the random variable $X_\alpha^k(\tau)$ is very important for the quality of the algorithm. Smaller variance $Var\{X_\alpha^k(\tau)\}$ leads to better convergence of the stochastic algorithm. It is proven that [4]:

$$Var\{X_\alpha^k(\tau)\} = \frac{c_{\alpha_0}}{p_{\alpha_0} p_\alpha} (\bar{A}_c^k \hat{c})_\alpha - (A_c^k c)_\alpha^2. \quad (11)$$

III. An improved Monte Carlo algorithm for LA systems

We use the Sequential Monte Carlo (SMC) method for linear systems introduced by John Halton [6]. We introduce the new improved Monte Carlo algorithm for the computation a linear functional form $W(x)$ of the solution of a linear system with real coefficients. The matrices B and the right-hand side f are normalized to accelerate the convergence rate of the stochastic process. Special values of the relaxation parameter $\gamma_i = b_{ii},\ i = 1, \ldots, n$, have been chosen, compared to the constant $\gamma \in (0, 1]$ in [4]. Numerical experiments show that it leads to balancing of the iteration matrix A.

The initial equation is picked uniformly at random among the n equations. After that for each state i we define the total score $S(i)$ and the total number of visits $V(i)$ that are modified as soon as state i is visited during a walk.

The following algorithm describes the above:

Now we give a description of the improved Monte Carlo algorithm for computing all the components of the solution. We compute scores for all the states (seen as new starting states) that are visited during a given trajectory. The initialization and preprocessing are the same as in the previous algorithm.

IV. Numerical examples and results

In order to check the accuracy of a computed solution $\hat{x}$, we compute the residual $r := B\hat{x} - f$ and "weighted residual" [8]:

$$\rho := \frac{||r||}{||B||\ ||\hat{x}||}. \quad (12)$$

The number of SMC iteration is N and the computational time t is measured in seconds. In the Tables and Figures below we present the values of the weighted residual. We perform a comparison with the refined iterative Monte Carlo (RIMC) [3] and the original "walk on equations" (WE) method which is completely described in [4]. For our method we use the

Algorithm 1 Computing one component x_{i_0} of the solution $x_i,\ i = 1, \dots n$.

Require: Initialization with initial data: the matrix B, the vector f, the constants $\gamma_i = b_{ii},\ i = 1, \dots, n$, and the number of random trajectories M.

Ensure: Preliminary calculations (preprocessing): compute the matrix A using the parameter $\gamma \in (0, 1]$:

$$\{a_{ij}\}_{i,j=1}^{n} = \begin{cases} 1 - b_{ii} \text{ when} & i = j \\ -b_{ij} \text{ when} & i \neq j \,. \end{cases}$$

Set $S := 0$.
for k=1 **to** M **do**
  set $m := i_0$
  set $S := S + f(m)$
  $test = 0,\ sign = 1$
  update $S := S + sign * f_m$;
end for
return $x_{i_0} = \frac{S}{M}$.

Algorithm 2 Computing all components $x_i,\ i = 1, \dots n$ of the solution.

Require: Initialization.
Ensure: Preprocessing.
Set $S(i) := 0$; $V(i) := 0$.
for $k = 1$ **to** M **do**
  set $m := rand(1 : n)$
  set $test := 0$; $m_1 := 0$
  $V(m) := V(m) + 1$; $m_1 = m_1 + 1$; $l(m_1) = m$
  for $q = 1$ **to** m_1 **do**
    $S(l(q)) := S(l(q)) + f_{l(q)}$
  end for
end for
return
for $j = 1$ **to** n **do**
  $V(j) := \max\{1, V(j)\}$
  $x_j = \frac{S(j)}{V(j)}$
end for

notation improved "walk on equations" algorithm (IWE). We have done numerical experiments with different matrices with dimensions $n = 7, 100, 5000$, where the number of equations in the linear system is n and $B \in \mathbb{R}^{n \times n}$. The number of random trajectories for lower dimensions $n = 7$ is $10n$. The number of random trajectories for $n = 100$ is $5n$ and for $n = 5000$ is n. A few sequential steps for the improved IWE algorithm combined with SMC are necessary.

In the examples below we try to find the solution x defined by the linear systems of algebraic equations $Bx = f$, where the matrix B and the vector b are preliminary given.

Example 4.1: In the first example we deal with two solutions x_1 and x_2 of $Bx = f$, where the matrix B and the vectors b_1 and b_2 are given below. The matrix is:

$$B = \begin{pmatrix} 5 & -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 5 & -1 & -1 & 0 & 0 & -1 \\ -1 & -1 & 5 & -1 & -1 & 0 & 0 \\ 0 & -1 & -1 & 5 & -1 & -1 & 0 \\ 0 & 0 & -1 & -1 & 5 & -1 & -1 \\ -1 & 0 & 0 & -1 & -1 & 5 & -1 \\ -1 & -1 & 0 & 0 & -1 & -1 & 5 \end{pmatrix}. \quad (13)$$

The vectors f_1 and f_2 are:

$$f_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad f_2 = \begin{pmatrix} 4 \\ -2 \\ -1 \\ 0 \\ -1 \\ -2 \\ 4 \end{pmatrix}. \quad (14)$$

The solutions are

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \quad (15)$$

Example 4.2: Let B is the matrix NOS4 from the Harwell-Boeing Collection [7], and $b \in \mathbb{R}^{100}, b_i = 1,\ i = 1, \dots, 100$. This particular matrix is taken from an application connected to finite element approximation of a problem describing a beam structure in constructive mechanics [7].

Example 4.3: Let B is a dense matrix 5000×5000 with elements in [0,1], and $f \in \mathbb{R}^{5000}, f_i = 1,\ i = 1, \dots, 5000$.

TABLE I
WEIGHTED RESIDUAL FOR THE MATRIX $B \in \mathbb{R}^{7\times7}$ AND x_1

N	RIMC	t	WE	t	IWE	t
2	2.28e-15	0.02	4.15e-02	0.11	1.00e-01	0.007
5	7.89e-16	0.07	1.39e-02	0.23	2.29e-03	0.026
10	8.07e-16	0.21	3.18e-06	0.68	1.24e-06	0.04
15	7.45e-16	0.35	3.94e-08	1.11	2.55e-10	0.1
20	6.66e-16	0.69	1.71e-10	2.53	7.78e-14	0.23
30	5.79e-16	1.14	8.12e-15	3.69	8.32e-17	0.49

TABLE II
WEIGHTED RESIDUAL FOR THE MATRIX $B \in \mathbb{R}^{7\times7}$ AND x_2

N	RIMC	t	WE	t	IWE	t
2	1.54e-01	0.003	4.63e-01	0.11	8.21e-02	0.003
5	1.01e-01	0.01	9.93e-03	0.23	2.48e-03	0.008
10	3.55e-02	0.04	1.43e-06	0.68	1.42e-06	0.05
15	4.04e-02	0.08	6.17e-09	1.11	2.66e-10	0.09
20	5.00e-02	0.14	1.40e-09	2.53	6.56e-14	0.16
30	4.72e-02	0.24	1.53e-14	3.69	5.03e-17	0.29

TABLE III
WEIGHTED RESIDUAL FOR THE MATRIX $NOS4 \in \mathbb{R}^{100\times100}$.

N	RIMC	t,s	WE	t,s	IWE	t,s
2	7.253e-02	0.05	4.178e-01	0.84	3.028e-03	0.08
5	5.449e-02	0.22	4.148e-01	2.37	3.071e-05	0.24
10	4.319e-02	0.56	5.943e-03	5.31	7.461e-08	0.61
15	3.520e-02	0.78	2.419e-06	9.1	1.217e-10	0.89
20	3.197e-02	1.11	3.336e-09	13.5	1.022e-13	1.13
30	1.835e-02	2.15	3.660e-12	18.6	1.109e-16	1.92

TABLE IV
WEIGHTED RESIDUAL FOR THE MATRIX $B \in \mathbb{R}^{5000\times5000}$.

N	RIMC	t	WE	t	IWE	t
2	5.438e-03	10.05	4.304e-02	3.95	2.931e-02	0.15
5	3.875e-03	60.2	1.217e-01	13.3	1.816e-04	0.9
10	2.866e-03	130.5	2.301e-05	32.3	1.235e-07	2.4
15	2.367e-03	310.7	6.486e-09	67.8	1.833e-10	5.1
20	1.941e-03	811	3.205e-09	171.5	1.054e-14	11.1
30	1.701e-03	2135	1.126e-07	418.6	2.481e-16	25.2

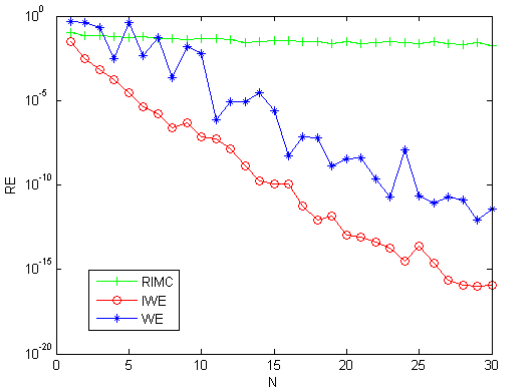


Fig. 1. Weighted residual for the matrix $B \in \mathbb{R}^{100\times100}$.

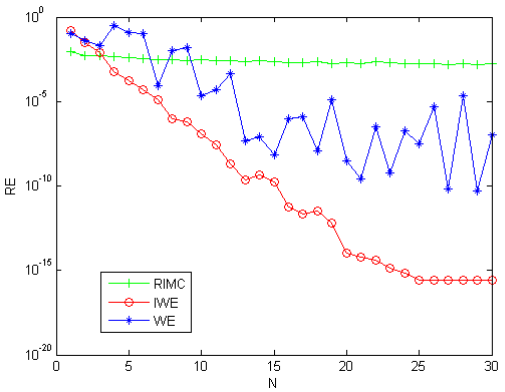


Fig. 2. Weighted residual for the matrix $B \in \mathbb{R}^{5000\times5000}$.

It can be seen that for the 7 dimensional case the difference in the accuracy between the WE and IWE for a given number of iterations is 2-3 order for $N > 15$ – see Table I and Table II. It is worth mentioning that refined iterative Monte Carlo convergence is very slow except for the trivial solution x_1 of Example 1. For the 100 dimensional case of the matrix NOS4 IWE produces much better results than WE and it is nearly 5-6 times faster – see Table III. The prior behavior of the proposed MC algorithm does not depend on the matrix density. The matrix NOS4 has only 5.9 average non-zeros per row and per column. The advantages of the algorithm hold for dense matrices. Also the difference in the accuracy for a fixed number of iterations is 3-5 order – see Figure 1. For larger dimensions the advantage of IWE over WE is even more pronounced. For the last example after 30 iterations WE has accuracy of 10^{-9} while IWE gives accuracy of 10^{-16} – see Figure 2. Also the time for the IWE is 15 times better than WE – see Table IV. The advantages of the proposed MC algorithm can be observed especially for larger matrix size. The special choice of relaxation parameters leads to the balancing of the iteration matrix and the experiments show that for larger dimensions the improvements leads to lower relative errors for small number of SMC iterations for IWE.

V. CONCLUSION

A new improved Monte Carlo algorithm for solving linear algebra problems is presented and studied. It is used for evaluating all the components of the solution of real valued systems. Due to the optimization techniques it gives superior results to the standard "walk on equations" method and it is established as one of the fastest and accurate Monte Carlo algorithm for solving systems of LA equations.

REFERENCES

[1] Curtiss, J.H., *Monte Carlo methods for the iteration of linear operators*, J. Math Phys., Vol. 32(4), 209–232, 1954, DOI: 10.1002/sapm1953321209.
[2] Dimov, I., Optimal Monte Carlo Algorithms, *Proceedings IEEE John Vincent Atanasoff 2006 International Symposium on Modern Computing*, October 2006, Sofia, Bulgaria, IEEE, Los Alamitos, California, 125–131, 2006, DOI: 10.1109/JVA.2006.37.
[3] Dimov, I., *Monte Carlo Methods for Applied Scientists*, New Jersey, London, Singapore, World Scientific, 291p, 2008.
[4] Dimov, I.T., S. Maire, J.M. Sellier, A New Walk on Equations Monte Carlo Method for Linear Algebraic Problems, *Applied Mathematical Modelling*, Vol. 39(15), 4494–4510, 2015, DOI: 10.1016/j.apm.2014.12.018.
[5] Golub, G.H., Van Loon C.F., *Matrix computations*, Third Edition, Johns Hopkins Univ. Press, Baltimore, 1996.
[6] Halton, J., Sequential Monte Carlo, *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 58(1), 57–78, 1962, DOI: 10.1017/S0305004100036227.
[7] NOS4: Lanczos with partial reorthogonalization. Finite element approximation to a beam structure. http://math-.nist.gov/MatrixMarket/data/Harwell-Boeing/lanpro/nos4.html
[8] Errors for Linear Systems: http://www.math.umd.edu/ petersd/466/linsysterrn.pdf

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 261–268
DOI: 10.15439/2018F213
ISSN 2300-5963 ACSIS, Vol. 15

Autonomous Graph Partitioning for Multi-Agent Patrolling Problems

Bernát Wiandt and Vilmos Simon
Budapest University of Technology and Economics
in Budapest
Magyar tudósok krt 2., Hungary
Email: [bwiandt,svilmos]@hit.bme.hu

***Abstract*—Patrolling algorithms are coordinating multiple agents with the goal of visiting points of interest in a timely manner. These algorithms play a major role in efficient use of UAVs or other autonomous vehicles for precision agriculture, large area monitoring or security use cases. These algorithms are either centralized and in need of constant connection with the agents or solve NP-hard problems to plan the routes of individual agents on the graph. These requirements become unfeasible when the number of agents or points of interest grow or become dynamic. In this article we further elaborate the performance characteristics of the Partition Based Patrolling Strategy (PBPS) algorithm. The partitioning requires only local interactions between agents, therefore it is scalable to a large number of nodes and agents. On these subgraphs agents patrol independently from each other, therefore the approach eliminates interference between agents.**

I. Introduction

PATROLLING is the task of visiting points of interest in a timely manner. These points of interest can be selected for security purposes (monitoring entry points of an area) or points in space one would like to monitor[1] by making photographs or measurements. Every problem that involves points of interests distributed in space and limited possibilities to move between them (modeled by a graph) is a patrolling problem. When patrolling inside a building, points of interest can be intersections or doors needed to be inspected in a timely manner. When dealing with a large open area, like an agricultural crop, the points of interest are locations where one wants to make measurements, i.e. water level, infections, fruit maturity, etc.

Points of interest are usually modeled as nodes of a graph, where the edges represent the option to travel between them directly and the length of the edges represent the distance or cost of moving between them. Patrolling is usually performed by physical agents, such as Unmanned Autonomous Vehicles (UAVs) or robots[2], either on ground or in the air. In this article we refer to these entities, without loss of generality, as agents. Patrolling with a single agent is a matter of calculating a sequence of points of interest for the agent to visit. Instead, our focus is on multi-agent patrolling, where the task is to efficiently coordinate an arbitrary amount of agents in order to minimize some performance metric associated with the patrolling task.

Coordinating multiple agents to perform patrolling can be reduced to a multiple traveling salesman problem, therefore it is an NP-hard problem[3]. Patrolling algorithms either plan the routes for each agent ahead of time and feed those routes to the agents as a priori information or try to coordinate the agents in real time by allowing them to communicate through some wireless medium and employing various heuristics in the agents to decide on the next graph node to visit.

Patrolling on graphs is either implemented by allowing every agent to visit all nodes in the graph or the graph is partitioned into subgraphs[4], [5] and the agents work only on their own subgraph without interfering with each other. The first case involves interference between the agents that can result in agents blocking each other on the way from one node to another. When agents try to move in the same direction or meet at an intersection, they inhibit each other from carrying out their tasks as fast as possible. These interference events can have a negative impact on the performance and scalability of the patrolling algorithm. Since patrolling is usually a long running task, these interference events will have a continuous, most of the time unpredictable effect on the performance.

To counter this effect, researchers started investigating partition-based patrolling. Currently partitioning is usually done before patrolling starts, on a central entity, and the resulting partitions are assigned to the agents as a priori information. This method is not always feasible as environments can be dynamic: appearing or disappearing graph nodes and failing or newly introduced agents can trigger the recalculation of partitions. Therefore agents need the central entity throughout the patrolling task and need synchronization at times of partition changes. A self-organizing partition based strategy is more desirable, because agents need less synchronization, require only local information and it makes the whole system more robust against agent failures.

II. Related works

Patrolling strategies can be really simple heuristic based approaches or complex learning algorithms that try to find the optimal route for a multi-agent team. It is not always clear whether a complex approach or a simple heuristic can offer better performance. In [6] patrolling is cast into a multi-agent Markov Decision Process and optimization techniques are applied to efficiently find the optimal paths for agents on the graph. However in the same article this solution is compared to a simple reactive algorithm and performance

differences with regards to the instantaneous idleness metric were negligible. This experiment and the conclusion of [7] suggest that computationally complex approaches are not always justified and simple reactive algorithms can match the performance of their more elaborate counterparts.

Agents can be restricted to their own subgraph or partition and perform patrolling alone in that region. An early theoretical study [3] gives some insight into the performance characteristics of the partition-based and cyclic approaches (for example the previously mentioned SingleCycle). Authors of this article claim that the optimal patrolling strategy for a single agent can be obtained by finding a shortest closed walk on the graph. This can be reduced to the Traveling Salesman Problem for which good approximations exist (for example the Christofides algorithm with $O(n^3)$ for a $\frac{3}{2}$ approximation). When multiple agents are considered, one strategy can be to traverse the closed path with evenly distributed agents along the path. The performance of such a strategy is bound by the edge length distribution of the graph [3]. For example on graphs with “long corridors" (edges connecting two distant nodes) a partition-based strategy could be a better fit, because those “long corridors" can be avoided by leaving them out of the partitions assigned to the agents.

So far the solutions enumerated were dealing with the multi-agent patrolling problem based on the assumption that agents cannot interfere with each other. However, interference can occur when two agents are too close to each other and cannot proceed in the direction of their targets. Interference is important when the algorithm is designed to be deployed on physical agents and can render theoretically better algorithms inferior in the real world. Accounting for interference results in a much harder problem, because the theoretically best results are no longer optimal if agents have to stop before colliding with each other or possibly choose different targets. Several new heuristic approaches are taking into account this phenomena, for example in [8] authors implement Greedy Bayesian Strategy (GBS) and State Exchange Bayesian Strategy (SEBS). Both strategies require global communication meaning that agents can communicate with all other agents involved in the patrolling task. They employ reactive agents (so the agents do not form plans ahead of time) and the decision making algorithm that chooses the next graph node to visit is based on Bayes' rule. The latter strategy (State Exchange Bayesian Strategy (SEBS)) aims to minimize the interference between agents while patrolling, by communicating the agent's next node it is intending to visit. SEBS is able to achieve better performance than any other heuristic studied in the article. Interference minimalization and graphs with non uniform edge length distributions contributed to the interest in partition based strategies.

Numerous other approaches have been tried to address the multi-agent patrolling problem, such as spanning tree[9], [10] or graph partitioning based approaches[11], [12]. Others involve task allocation based strategies[13] and evolutionary algorithms[14] or linear programming[15]. Auction based strategies involve agents placing bids on nodes to patrol on the graph. These kind of approaches implement a decentralized task allocation, where the agents decide on the outcome and able to reassign nodes or partitions of poorly performing agents to others[16], [17].

III. The Multi-Agent Patrolling Problem

In this article our goal is to further elaborate the performance characteristics of our proposed PBPS algorithm [18]. In our model the only global knowledge the agents have is the map they are patrolling on. They do not have the capability to communicate globally with each other. Agents' behavior will result in solving the patrolling problem efficiently together, without third party coordination.

In the model, the connected graph $G(V, E)$ consists of nodes $v_1, v_2, ..., v_n \in V$ and undirected edges $e_{i,j} \in E$. Nodes are the points of interest, they are visited by the agents periodically. Each node has a counter that measures time between the last visit to that node and the current time. This counter is usually referred to as the *instantaneous idleness* of the node at time t and is defined as $I_{v_i}(t) = t - I'_{v_i}$, where I'_{v_i} is the time at the last visit to the node and $v_i \in V$. In some use cases patrolling is done to measure some physical quantity at the point of interest periodically and as frequently as possible. Let us suppose that the measurement at each point of interest takes the same amount of time for each agent.

Agents' movement is restricted to the edges between connected nodes. Agents have a priori knowledge of the graph nodes and edges, and are capable of moving between nodes along the edges with constant and uniform speed. Agents are able to communicate with each other, therefore influence each other's decisions by broadcasting messages via a wireless medium.

Patrolling algorithms have very different characteristics with respect to the routes they take, the order in which they visit certain nodes, etc. To be able to compare their performances, we have to define metrics that represent the goodness of an algorithm from a certain point of view. The metric used for this task is usually the *average idleness* or the *worst idleness*. Both of these metrics are based on the *instantaneous idleness* defined before as $I_{v_i}(t)$. This basic metric can be averaged in a window or throughout the simulation to obtain the *average idleness*, defined as:

$$\overline{I_{v_i}}(t) = \frac{\overline{I_{v_i}}(t_{v_i}) \cdot C_i + I_{v_i}(t)}{C_i + 1} \tag{1}$$

where C_i is the number of visits to v_i. The *average graph idleness* is the average of *average idleness* values, such that:

$$\overline{I_G}(t) = \frac{1}{|V|} \sum_{i=1}^{|V|} \overline{I_{v_i}}(t) \tag{2}$$

A problem with the *average graph idleness* metric is that there can be two patrolling algorithms with the same *average graph idleness* but significantly different behaviors. *Average graph idleness* masks important characteristics of patrolling algorithms and as a consequence, its value can be misleading

when comparing different algorithms. The idleness times between visits are samples from an unknown distribution and to reason about that unknown distribution based on the mean of a limited amount of its samples can be misleading. Therefore in this article we use the distribution of the *worst idleness* values to compare performances. The *worst idleness* associated with a node between times t_a and t_b is defined as the maximum *instantaneous idleness* value:

$$WI_{v_i} = max\{I_{v_i}(t_a), ..., I_{v_i}(t_b)\} \quad (3)$$

The worst idleness times are the worst case scenario, therefore by gathering enough samples, one can be reasonably sure of an upper bound of the idleness times that can be observed during patrolling. Patrolling use cases like reading measurements from sensors or taking photographs of certain physical locations usually benefit from a known maximum time between measurements. The *worst idleness* distribution makes it possible to evaluate patrolling algorithms based on this criteria. It also shows the fairness of the algorithm: the difference between the minimum and maximum *worst idleness* values on the graph over time. Moreover the distribution of *worst idleness* can be easily visualized using box plots.

IV. Partition Based Patrolling Strategy (PBPS)

The patrolling strategy we use in this article follows the partition based approach. Based on [3], [19], the performance of such a patrolling algorithm is in theory inferior compared to a cycle-based strategy on graphs without “long corridors". However, if the model accounts for the effects of interference between agents, this may no longer be true. Partitioning the graph between the agents is not a new idea, authors in [8] and [5] improved their patrolling algorithms by limiting the amount of interference between agents. Partition based strategies take this idea to the extreme as there is no interference between agents, they patrol their own partitions independently.

PBPS was published in [18] along with the first results obtained on well-known graphs. Here we give only a short introduction to the algorithm. PBPS has two important assumptions about the environment: agents can localize themselves and they know the structure of the graph a priori. It should be noted, that these assumptions are not typical in self-organized systems, because they require agents with greater knowledge than usual. However the information requirement of PBPS is the same as other patrolling algorithms, such as SEBS or Greedy Bayesian Strategy (GBS). The problem of discovering the environment and forming consensus about it among the agents is not discussed in this article.

The goal of the algorithm is to partition the graph into non-overlapping connected subgraphs and in the process create a global partitioning that enables the efficient patrolling of the graph. All agent’s partitions are empty in the beginning. A partitioning of G is defined as $P = \{P_1...P_k\}$, such that $P_1 \cup ... \cup P_k = V$ and $P_i \cap P_j = \emptyset$. $\{G_1...G_k\}$ refer to subgraphs induced by the partitioning, thus $G_i = (V \cap P_i, E \cap (P_i \times P_i))$. The partition growing process is self-organized, such that there is no global coordination between agents and no global state shared among them. Agents grow their own partitions by claiming nodes from each other’s partitions and broadcasting their new partition after claiming a new node. In the process (see for example Figure 1) of growing partitions, there can be local violations to the strict non-overlapping partitioning introduced above, but these are temporary and will resolve over time. The algorithm always reaches a state, where the partitions are stable, i.e. no more changes made by the agents. When agents arrive at a stable configuration, each partition will belong to one and only one agent and each partition will be a connected subgraph with no overlapping nodes with other partitions. The resulting partitioning assigns a subgraph to each agent to patrol with roughly equal number of nodes.

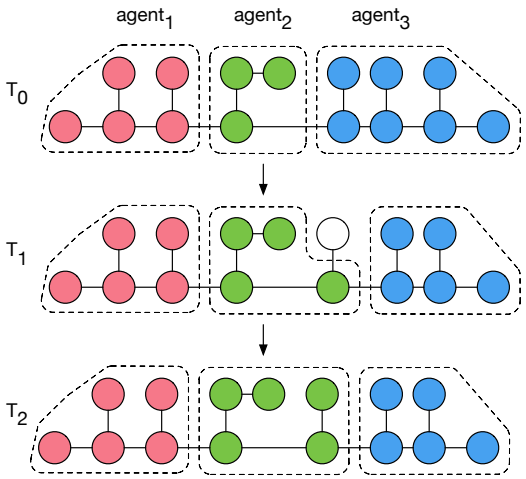


Fig. 1: Claiming of nodes between agents in order to achieve a balanced partitioning.

V. Simulation environment

We have conducted numerous simulations to investigate the performance characteristics of PBPS and compared patrolling on the formed partitions to well-known patrolling algorithms. First we describe the environment in which the experiments were carried out as it was designed to resemble possible real-world usage scenarios. We used the osmnx python library[20] to download maps of two Hungarian cities: Budapest and Komárom. We chose Budapest, since it is the capital of Hungary and like large cities it has a lot of intersections and shorter edges. Komárom on the other hand tends to have differently distributed edge lengths and node degrees (as can be seen on Fig. 2 and 3) as it is a smaller city on the countryside. Node degrees tend to be lower for Komárom but edge lengths are significantly longer than the ones for Budapest. We chose two different real maps to model possible usage scenarios and also to have environments with different characteristics. Our partitioning strategy is based on the assumption that the edge length distribution of the graph is more or less even. The map excerpt from Komárom has a longer tail in its edge length distribution, therefore more variety than the map excerpt from Budapest. The downloaded maps were simplified as the

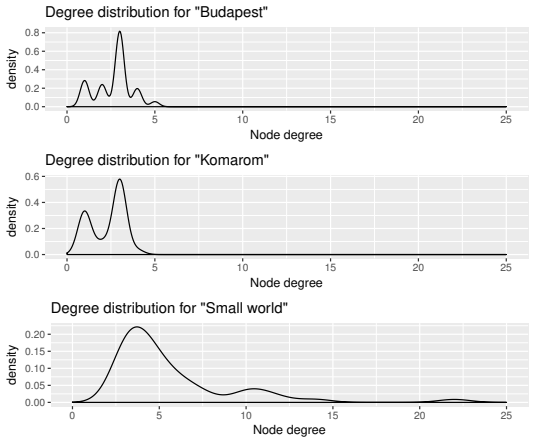


Fig. 2: Degree distributions of the maps used for the experiments.

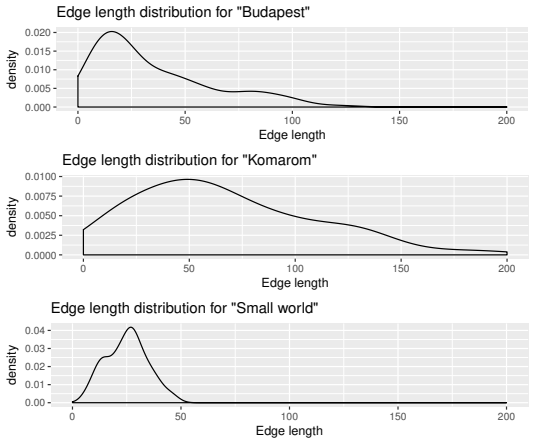


Fig. 3: Edge length distributions of the maps used for the experiments.

original Open Street Map data contains a lot of short edges and unnecessary nodes. The simplification step was done using a built-in functionality in osmnx. A third map was introduced, called Small World, which is a generated small world type graph (Barabási-Albert graph) with a completely different node degree distribution than the other two maps. This graph has 50 nodes and was grown with 3 edges preferentially attached to existing nodes using the networkx python library's built in graph generator. This third environment has some really high degree nodes and claiming those nodes in the partitioning phase can effect other partitions, as these high degree nodes act as access points to the rest of the graph.

We carried out experiments in these graphs with three different patrolling algorithms with parameters given in Table I. First, Conscientious Reactive[21] was used as a baseline as it is a greedy heuristic with no communication requirements among the agents. The second algorithm tested was State Exchange Bayesian Strategy[8], which is based on the idea that agents can perform better if they exchange their intentions and

Parameter	**Value**
Number of agents	$1, 2, 4, \ldots, 32$
Double number of agents every	$200000s$
Number of iterations	10
Agent speed	$1m/s$

TABLE I: Simulation parameters for the local vs. global metric experiment.

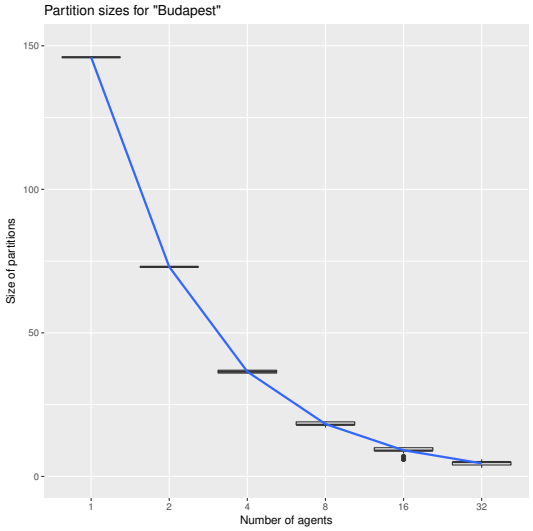


Fig. 4: Partition sizes for Budapest

factor the intentions of other agents in when making a decision which node they visit next. The third was PBPS, which forms partitions on the map in a self-organized manner and agents perform patrolling on their own partition only.

VI. Partitioning algorithm results

The goal of this experiment was to gain insight into the performance characteristics of the partition forming part of the PBPS algorithm. To this end we have disabled the mobility of the agents and tracked the partitioning events throughout the simulation. Partitioning events are either node free or nonfree node claims, or partition resets. In the beginning of every simulation run, only one agent is on the map at a randomly chosen node and it starts forming its partition. After this initial agent claims the whole map as its partition there are no more partitioning events occurring. The simulator waits until there are no more partitioning events for 5000 time steps and then doubles the number of agents on the map. The agents already on the map retain their partitions and the new agents are assigned a random starting node as their initial partition. This mechanism for adding agents sometimes cause overlaps in the partitioning, because the previous generation of agents claimed the whole map together (every node belongs to one and only one agent). However this is not a problem as the first time a new agent broadcasts its initial partition, the overlaps will be resolved by the old agent releasing that node from its partition.

The number of agents simulated on each map are exponentially increasing $(1, 2, \ldots, 16)$. The exponential increment in

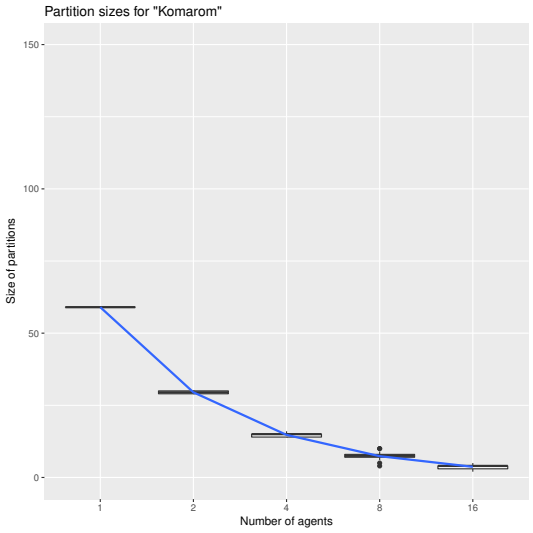


Fig. 5: Partition sizes for Komárom

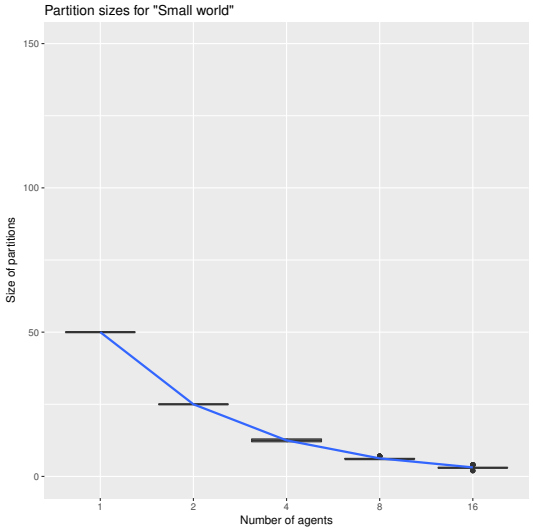


Fig. 6: Partition sizes for Small World

the number of agents can reveal the scalability characteristics of the partition forming part of PBPS with respect to the different maps. As a performance metric, the time to arrive at a stable partitioning was measured for different number of agents $(1, 2, 4, \ldots, 16)$ and at the same time the number of nodes in each partition were tracked.

This was done in order to confirm whether a stable partitioning is also a balanced one. A balanced partitioning in this terminology is where the number of nodes in each agent's partition is close to the $\frac{number\_of\_nodes}{number\_of\_agents}$ number. In other words, a balanced partitioning is where each partition has a roughly equal amount of nodes. Depending on the structure of the map this balanced property might not be achievable, since the tested maps are not full graphs with an edge connecting every node. Figures 4, 5 and 6 report the distribution of

Num. of agents	1	2	4	8	16	32
Budapest belváros						
Mean	146.0	73.0	36.5	18.25	9.125	4.5625
Var	0.0	0.0	0.2564	0.3670	0.6383	0.3534
IDI	0.0	0.0	0.0070	0.0201	0.0699	0.0774
Komárom						
Mean	59.0	29.5	14.75	7.3750	3.6875	N/A
Var	0.0	0.2631	0.5000	1.1487	0.3671	N/A
IDI	0.0	0.0089	0.0338	0.1557	0.0995	N/A
Small world						
Mean	50	25	12.5	6.25	3.125	N/A
Var	0.0	0.0	0.2564	0.1898	0.1729	N/A
IDI	0.0	0.0	0.0205	0.0303	0.0553	N/A

TABLE II: Mean, variance and index of dispersion values of resulting partition sizes for all three tested maps and different number of agents.

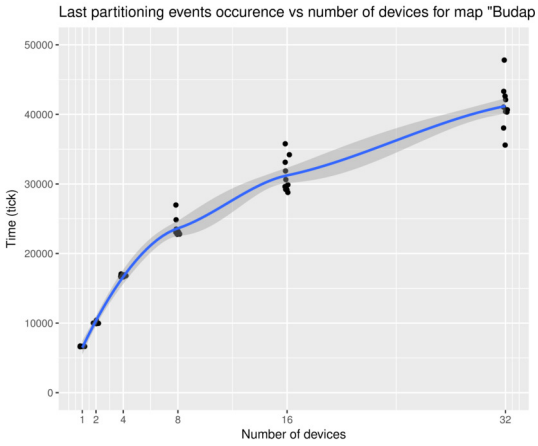


Fig. 7: Times of last partitioning events for versus number of agents for Budapest

partition sizes versus the number of agents. These are box plots[22] where the box contains 75% of the observed values. The blue curve shows the balanced partition sizes for different number of agents, its value is $\frac{number\_of\_nodes}{number\_of\_agents}$, where $number\_of\_nodes$ is constant and a property of the map. The results here show that in all simulated experiments the partitioning algorithm was able to arrive at solutions close to the optimal. Some exceptions show up as outliers on the box plots and indicate that the initial conditions have some influence on the final partitioning. All simulations were done 10 times with different starting positions for the agents, therefore each box represents the distribution of $10 * number\_of\_agents$ different resulting partition sizes.

Additionally the mean, variance and index of dispersion[23] metrics were calculated for the three different maps and for the different agent numbers. These results are summarized in Table II.

The next set of results discuss the time requirements of the partitioning algorithm, paying attention to its scaling properties. The question here is how does adding more agents to a map influence the time needed to arrive at a stable configuration? A stable configuration is one, where no more

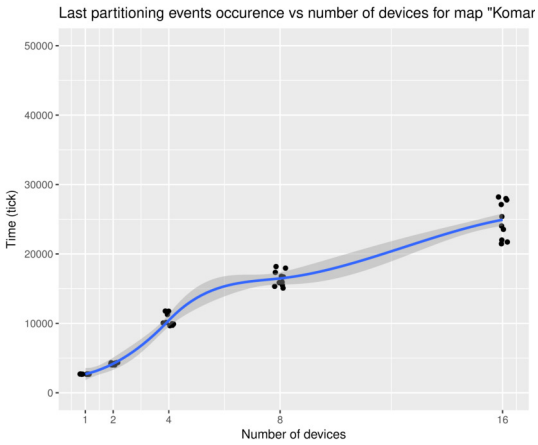


Fig. 8: Times of last partitioning events for versus number of agents for Komárom

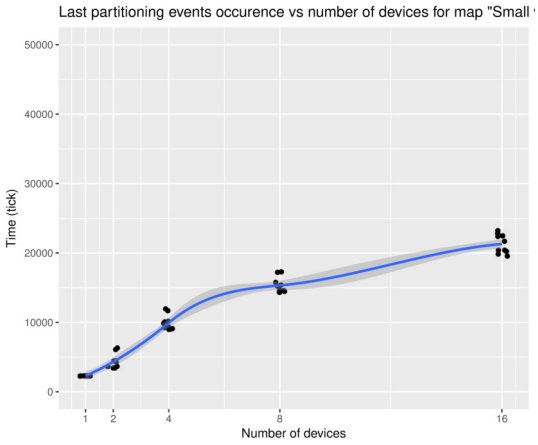


Fig. 9: Times of last partitioning events for versus number of agents for Small World

partitioning events occur. One of the advantages of using a self-organized approach is that in theory it is scalable and can enable multi-agent patrolling on larger graphs for a larger number of agents, than centralized algorithms. Figures 7, 8 and 9 show the distribution of time of the last partitioning event for each different simulated agent population versus the number of agents across all 10 iterations. Each group of values consists of 10 measurements and a curve is fitted to show the observed general trend of the values. It is important to note that the 5000 time step period between agent additions is subtracted from the values on the figures to show an ideal situation where one could determine the exact moment when the partitioning stops. The figures show that although the number of agents increase exponentially, the time needed to partition the graph increases only linearly. This indicates, that for the tested maps, the algorithm behaves well with a large number of agents and might be suitable to handle very large number of agents partitioning the same graph simultaneously. The variance of the values increase for larger number of devices. This indicates that the intermediate resulting partitionings and randomly assigned starting nodes for added agents (initial condition for each addition step) has an increasingly large effect when the number of agents are high. It is possible that by carefully choosing the insertion point for new agents, the observed variance might be decreased. However in this analysis we were interested in the possible outcomes and was aiming of getting a full picture of the dynamic behavior PBPS, rather than optimizing its parameters for ideal performance.

VII. Patrolling results

After investigating the different aspects of the partitioning, we compared the PBPS algorithm with two other algorithms: Conscientious Reactive (CR) and State Exchange Bayesian Strategy (SEBS). Conscientious Reactive (CR) is a simple greedy reactive algorithm, where the agent makes decisions based only on the visits made by itself. Other agents are not taken into account, therefore no communication is needed by this algorithm. Upon arriving to a node it resets the node's idle timer in its own local data structure and evaluates the node's neighbors in the graph. For the next node to visit the agent chooses the one with the largest instantaneous idle time value. CR is a trivial algorithm that requires only an agent with the capability to move and store the graph and timers in its internal memory, therefore the algorithm is an ideal baseline to compare other, more sophisticated algorithms to.

The State Exchange Bayesian Strategy (SEBS) strategy takes into account the interference caused by other agents in the same area. Agents can inhibit each others' movements by standing in the way or forcing other agents to move slower to avoid collision. This effect can cause strategies like CR and GBS to perform poorly in situations, where the number of agents in an area becomes high. GBS and SEBS builds on the idea of Bayesian decisions, where the a priori probability to visit a neighbor node is offset by the instantaneous idleness of the node known to the agent. This way GBS can take into account priorities input by the designer and the actual state of the system (instantaneous idleness of the nodes). SEBS extends GBS by having agents broadcast their intentions (the node they are intend to visit next) for their immediate neighbors. This information is used in all agents when evaluating the agent's next node to visit. If one or more agents intend to visit the same node, it is not beneficial to let another one go in that same direction, even if the instantaneous idleness value of the node is high. SEBS aims to lower the amount of time an agent loses by executing collision avoidance behaviors, causing them to slow down or chose a different target node altogether.

In these simulation runs, each group size (number of agents) ran for 200000 time steps, and after the number of agents were doubled. This process was repeated until there were 16 agents for Komárom and Small World, and 32 agents for Budapest on the map. We ran 10 iterations for each algorithm-map pair with random starting positions for every agent for every iteration. The structure of the graphs and the starting position of the agents can influence their measured performance, therefore in

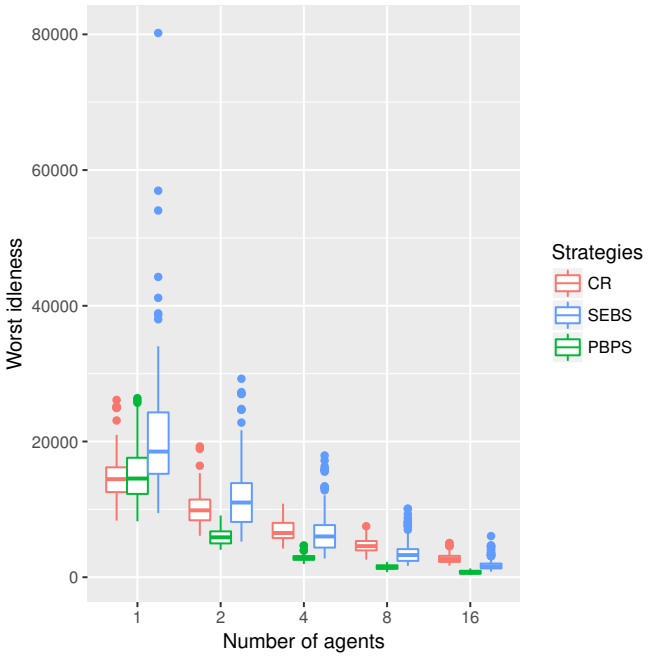


Fig. 10: Worst idleness time distributions for different strategies versus number of agents for Budapest

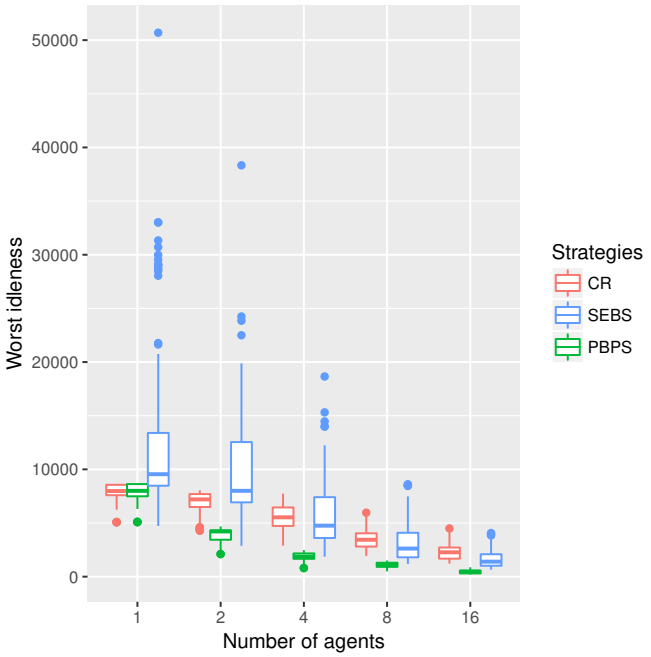


Fig. 11: Worst idleness time distributions for different strategies versus number of agents for Komárom

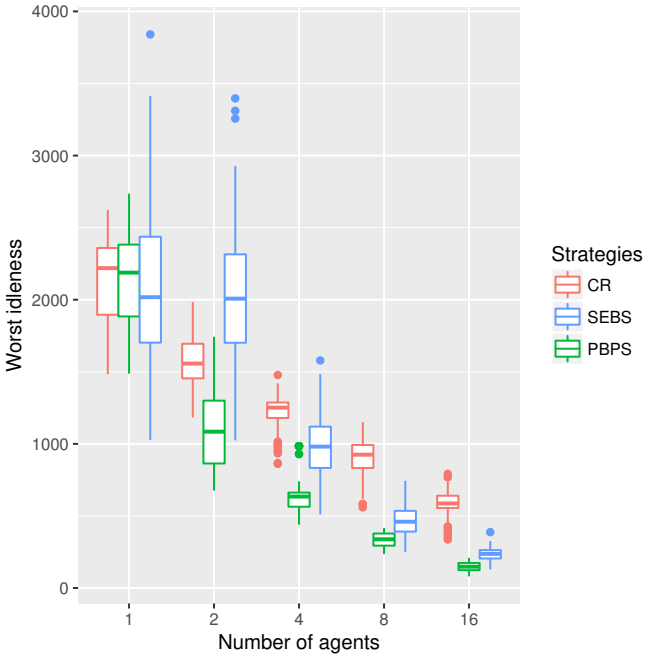


Fig. 12: Worst idleness time distributions for different strategies versus number of agents for Small World

order to gain as much information about the performances of different algorithms as possible, random non-overlapping starting points were chosen for the agents when they were added to the team. In the case of CR and SEBS strategies this means a random node from the graph and in the case of PBPS, this results in a random node from the agent's partition.

Partitioning strategies have a tendency to perform worse when they start on a graph, because they base their decisions on idleness times (and other agents' intentions in the case of SEBS), but in the beginning this data is not available for them. Therefore the first 100000 time steps for each different configuration were dropped in order to let the agents "warm up" and the worst idleness times for each node were collected for the last 100000 of each configuration. Figures 10, 11 and 12 report the worst idleness time distribution measured on the whole map for different strategies and different number of agents.

For one agent, CR and PBPS are almost identical, this is caused by the fact that PBPS employs CR as the patrolling strategy in the agents' partitions. In all tested configurations the performance of PBPS were better than the other tested algorithms. Comparing worst idleness distributions between SEBS and PBPS reveals that PBPS requires on average half the amount of agents to achieve the same performance as SEBS. In the case of map Komárom this effect is even more pronounced as the required number of agents for PBPS is one fourth of that for SEBS.

Another interesting facet is that SEBS tends to have a distribution with a longer tail than CR and PBPS. This means that the majority of values are larger than the mean for maps

Budapest and Komárom (see in Figures 10 and 11), which indicates that judging the performance of SEBS by the mean of its worst idleness times can be misleading, by indicating that the average time between visits to nodes in the map is smaller than what can be interpreted from the distribution of the values. This property means that performance metrics reported in articles like [8] can be only part of the whole picture as one cannot judge every aspect of the performance of the algorithms accurately from the average idleness values.

VIII. Conclusion

We have elaborated on the performance characteristics of the PBPS algorithm first published in [18]. The result of the partitioning algorithm is a set of subgraphs formed by the individual agents cooperating with each other that is used by patrolling agents as their patrolling tasks. These subgraphs are formed cooperatively by the agents on the graph without any central help or control. Using subgraphs of the original graph allows to completely avoid interference between agents, therefore have them performing in a parallel and deterministic manner. The cost function used in this article was the difference in number of nodes between partitions, but changing this cost function opens up possibilities for designers of patrolling systems to tailor the resulting partitions to their use case.

Simulations were performed on three different maps, representing different environments: the inner city of Budapest with shorter edges and more neighbors, part of the small city Komárom with longer edges and fewer neighbors and a random graph, which is a generated small world type graph (Barabási Albert graph). PBPS arrived at a balanced partitioning in every simulation run with low variance in the number of nodes assigned to an individual agent. Moreover PBPS scales well with the number of agents, meaning that the time needed to finish partitioning increased sublinearly with the number of agents tested. This property is in contrast with planning approaches, that usually have worse scaling properties, making it viable to employ this strategy on larger graphs and with a larger amount of agents.

We compared the performance of agents using different patrolling strategies, such as CR, SEBS and PBPS. Patrolling performance was judged by the distribution of worse idleness times for all nodes in the graph, represented as box plots. In every tested case PBPS came out ahead, and on two maps the necessary number of agents for PBPS were half of what was required by CR and SEBS for the same performance. This results in a tradeoff decided by the designer or implementor later: either employ the same amount of agents and get lower worst idleness times throughout the graph, or in a resource constrained scenario, change the patrolling strategy and get the same performance level with only half the agents required by other strategies.

References

[1] Gonçalo Cabrita, Pedro Sousa, Lino Marques, and A de Almeida. Infrastructure monitoring with multi-robot teams. In *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 18–22, 2010.

[2] David Portugal and Rui P Rocha. Cooperative multi-robot patrol in an indoor infrastructure. In *Human Behavior Understanding in Networked Sensing*, pages 339–358. Springer, 2014.

[3] Yann Chevaleyre. Theoretical analysis of the multi-agent patrolling problem. In *Intelligent Agent Technology, 2004.(IAT 2004). Proceedings. IEEE/WIC/ACM International Conference on*, pages 302–308. IEEE, 2004.

[4] Talita Menezes, Patrícia Tedesco, and Geber Ramalho. Negotiator agents for the patrolling task. In *Advances in Artificial Intelligence-IBERAMIA-SBIA 2006*, pages 48–57. Springer, 2006.

[5] David Portugal and Rui Rocha. Msp algorithm: multi-robot patrolling based on territory allocation using balanced graph partitioning. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1271–1276. ACM, 2010.

[6] Jean-Samuel Marier, Camille Besse, and Brahim Chaib-Draa. Solving the continuous time multiagent patrol problem. In *ICRA*, pages 941–946. Citeseer, 2010.

[7] Alessandro Almeida, Geber Ramalho, Hugo Santana, Patrícia Tedesco, Talita Menezes, Vincent Corruble, and Yann Chevaleyre. Recent advances on multi-agent patrolling. In *Advances in Artificial Intelligence–SBIA 2004*, pages 474–483. Springer, 2004.

[8] David Portugal and Rui P Rocha. Distributed multi-robot patrol: A scalable and fault-tolerant framework. *Robotics and Autonomous Systems*, 61(12):1572–1587, 2013.

[9] Pooyan Fazli, Alireza Davoodi, and Alan K Mackworth. Multi-robot repeated area coverage. *Autonomous robots*, 34(4):251–276, 2013.

[10] Yoav Gabriely and Elon Rimon. Spanning-tree based coverage of continuous areas by a mobile robot. *Annals of mathematics and artificial intelligence*, 31(1-4):77–98, 2001.

[11] Tiago Sak, Jacques Wainer, and Siome Klein Goldenstein. Probabilistic multiagent patrolling. In *Brazilian Symposium on Artificial Intelligence*, pages 124–133. Springer, 2008.

[12] Ruben Stranders, E Munoz De Cote, Alex Rogers, and Nicholas R Jennings. Near-optimal continuous patrolling with teams of mobile information gathering agents. *Artificial intelligence*, 195:63–105, 2013.

[13] François Sempé and Alexis Drogoul. Adaptive patrol for a group of robots. In *Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on*, volume 3, pages 2865–2869. IEEE, 2003.

[14] Oswaldo Aguirre and Heidi Taboada. An evolutionary game theory approach for intelligent patrolling. *Procedia computer science*, 12:140–145, 2012.

[15] Burcu B Keskin, Shirley Rong Li, Dana Steil, and Sarah Spiller. Analysis of an integrated maximum covering and patrol routing problem. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):215–232, 2012.

[16] Charles Pippin, Henrik Christensen, and Lora Weiss. Performance based task assignment in multi-robot patrolling. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 70–76. ACM, 2013.

[17] Cyril Poulet, Vincent Corruble, and Amal El Fallah Seghrouchni. Working as a team: using social criteria in the timed patrolling problem. In *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, volume 1, pages 933–938. IEEE, 2012.

[18] Bernát Wiandt, Vilmos Simon, and András Kőkuti. Self-organized graph partitioning approach for multi-agent patrolling in generic graphs. In *Smart Technologies, IEEE EUROCON 2017-17th International Conference on*, pages 605–610. IEEE, 2017.

[19] David Portugal, Charles Pippin, Rui P Rocha, and Helen Christensen. Finding optimal routes for multi-robot patrolling in generic graphs. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 363–369. IEEE, 2014.

[20] Geoff Boeing. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Browser Download This Paper*, 2016.

[21] Aydano Machado, Geber Ramalho, Jean-Daniel Zucker, and Alexis Drogoul. Multi-agent patrolling: An empirical analysis of alternative architectures. In *Multi-Agent-Based Simulation II*, pages 155–170. Springer, 2002.

[22] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[23] DR Cox and PAWL Lewis. The statistical analysis of series of events. 1966.

Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- BEDA'18—1st International Workshop on Biomedical & Health Engineering and Data Analysis
- CANA'18—11th Computer Aspects of Numerical Algorithms
- C&SS'18 - 5th International Conference on Cryptography and Security Systems
- CPORA'18—3rd Workshop on Constraint Programming and Operation Research Applications
- LTA'18—3rd International Workshop on Language Technologies and Applications
- MMAP'18—11th International Symposium on Multimedia Applications and Processing

Area Supervisory Committee

- Burdescu, Dumitru Dan, MMAP'18
- Damasevicius, Robertas, LTA'18
- Ge, Mouzhi, DaSCA'18
- Janicki, Artur, BigDAISy'18
- Królak, Aleksandra, BEDA'18
- Ksiezopolski, Bogdan, C&SS'18
- Paprzycki, Marcin, 4A'18
- Ristov, Sashko, WSC'18
- Sitek, Pawel, CPORA'18
- Stpiczyński, Przemysław, CANA'18

1st International Workshop on Biomedical & Health Engineering and Data Analysis

IN the recent years, technology has had an accelerating impact on the field of medicine and healthcare. We could observe, in particular, a proliferation of personal health monitoring devices and wearables, such as activity bracelets. These technologies supported by newest developments in data analysis, such as big data and deep learning, have a substantial impact on daily life of many people. They increase awareness of daily activities, help improve sport performance, and can lead to early detection of certain diseases.

The workshop on Biomedical & Health Engineering and Data Analysis—BEDA'2018—provides an open forum for researchers in domains of Biomedical Engineering, Health Technologies, Personal Monitoring Devices, and Data Analysis to communicate high-quality and timely research results. The primary focus is on practical applications, but highly relevant theoretical papers are also of interest.

TOPICS

- Biomedical signal processing;
- Biomedical imaging and image processing;
- Biosensors and bioinstrumentation;
- Neural engineering, neuromuscular systems and rehabilitation engineering;
- Wearable biomedical sensors and systems;
- Health informatics and technology, e-health and telemedicine;
- Biomedical systems management;
- Personal health monitoring devices and wearables;
- Application studies of biomedical and health technologies.

EVENT CHAIRS

- **Królak, Aleksandra,** Łódź University of Technology, Division of Medical Electronics, Poland
- **Wiktorski, Tomasz,** University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering

PROGRAM COMMITTEE

- **Agrawal, Bikash,** DNV GL
- **Augustyniak, Piotr,** AGH University of Science and Technology
- **Bajcsy, Peter,** National Institute of Standards and Technology
- **Byambajargal, Byambayav**
- **Caraiman, Simona,** University of Iasi
- **Chakravorty, Antorweep,** University of Stavanger
- **Eftestøl, Trygve,** University of Stavanger, Faculty of Science and Technology, Department of Electrical and Computer Engineering
- **Lundervold, Arvid,** University of Bergen
- **Materka, Andrzej,** Lodz University of Technlogy
- **Moldoveanu, Alin,** University POLITEHNICA of Bucharest
- **Pissaloux, Edwige,** Université Pierre et Marie CURIE
- **Strumiłło, Paweł,** Lodz University of Technology
- **Strzelecki, Michal,** Lodz University of Technology, Poland
- **Szczypiński, Piotr,** Lodz University of Technlogy
- **Tadeusiewicz, Ryszard,** AGH University of Science and Technology, Poland
- **Torbicz, Władysław,** Polish Academy of Sciences
- **Unnþórsson, Rúnar,** University of Reykjavik
- **Velázquez, Ramiro,** Universidad Panamericana
- **Vinhais, Carlos,** Instituto Superior de Engenharia do Porto, Portugal

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 273–277
DOI: 10.15439/2018F57
ISSN 2300-5963 ACSIS, Vol. 15

Prediction of Alzheimer's Disease in Patients using Features of Pupil Light Reflex to Chromatic Stimuli

Wioletta Nowak*, Minoru Nakayama†, Tomasz Kręcicki‡ and Andrzej Hachoł*
* Institute of Biomedical Engineering and Instrumentation
Wrocław University of Science and Technology, Wrocław, Poland 50–370
Email: wioletta.nowak@pwr.edu.pl
† Department of Information and Communications Engineering
Tokyo Institute of Technology, Tokyo, Japan 152-8552
Email: nakayama@ict.e.titech.ac.jp
‡ Wrocław Medical University Rektorat, wybrzeze Ludwika Pasteura 1, 50-367 Wrocław, Poland

***Abstract*—A diagnostic procedure to predict the probability of diagnosing a patient with Alzheimer's Disease (AD) was developed using features of pupil light reflex (PLR) waveforms. 15 features of PLRs for three colours of light pulses at two levels of brightness were measured. Participants were 12 AD patients and 7 control group subjects. A logistic regression analysis was introduced to identify AD patients using two factor scores of features of PLR. The prediction performance of combinations of factor scores for features of PLRs were then evaluated using a test of fitness. An MCMC technique was introduced to estimate the parameters of the regression functions. The model provides a distribution of the probability of diagnosis of AD patients and control group subjects.**

I. Introduction

The pupil light reflex (PLR), which produces changes in pupil diameter in response to a light pulse of white or red, has been introduced to diagnose Alzheimer's Disease (AD) [1], [2]. In addition to this, the recent discovery of intrinsically photosensitive retinal ganglion cells (ipRGCs) [3] reveals the possibility of using various diagnostic procedures that involve a shorter light wavelength, such as blue light [4]. For example, PLRs related to ipRGCs can be used to detect symptoms of Age-Related Macular Degeneration (AMD) [5]. Some critical studies have suggested that common sources may be the origin of both AD and AMD diseases [6], [7], [8], [9]. Also, because most AD patients are elderly, the influence of aging on PLRs should be evaluated carefully.

The authors have been studying a diagnostic procedure for detecting AD symptoms using PLRs of various types of light pulses and observing the conditions the light pulses produce [10]. Though these results show the possibility of aiding the diagnosis of the disease, a more flexible procedure is required. In particular, the number of AD patients used in the experimental survey was restricted, and the assessment of the prediction of accuracy was insufficient. Therefore, significant features of PLRs, which can be used in the diagnostic procedure, should be extracted as necessary. If the distribution of features can be estimated, the possibility of diagnosing AD patients may be predicted using the Bayesian process.

In this paper, features of PLRs are extracted and the differences between AD patients and control group subjects are discussed. Some procedures for predicting the probability of diagnosing AD patients are compared. The following topics are addressed:

1) Features of PLRs are analysed and their factors are extracted. The differences in features and factor scores between AD patients and control group subjects are compared. Also, the influence of aging is evaluated.
2) Logistic regression is introduced to calculate the probability of diagnosing the disease in AD patients and control group subjects. The models of fitness of the groups are then compared.
3) The MCMC technique is introduced to estimate the parameters of the models, and the performance of the models is discussed.

II. Method

A. Participants

A conventional PLR experiment was performed on 19 participants (42∼89 years old, mean age:70.6), 12 of which were healthy individuals with normal vision (Control group: 62∼89 years old, mean age:72.1) and 7 of which were patients with Alzheimer's Disease (AD Patients: 42∼84 years old, mean age:68.1) who had already been diagnosed by medical doctors. It was not easy to invite volunteers who were aged over 80. The age levels are summarised in Table I.

Informed consent was obtained from all participants prior to the experiment.

B. PLR measurement

The stimuli consisted of three chromatic lights, red (635nm), blue (470nm) and white (CIE x:0.28, y:0.31), at two levels of

TABLE I
Number of subjects by age

Label	Age	Control	Patient
L	≤ 70	5	3
M	71∼80	5	3
H	$81\leq$	2	1

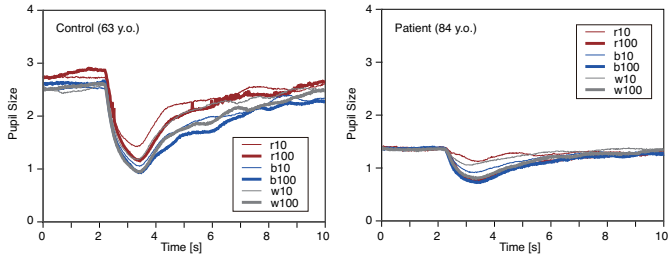


Fig. 1. An example of PLRs for a control group subject and an AD patient

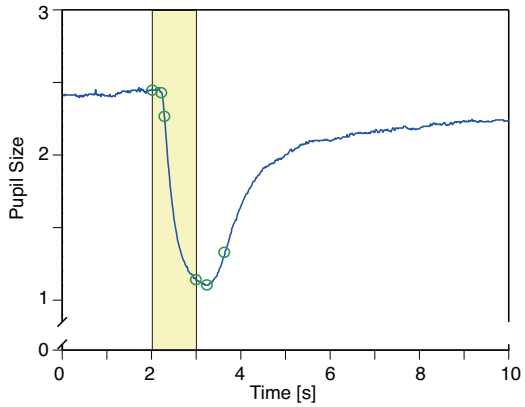


Fig. 2. Feature extraction from PLR responses

brightness (10 and 100 cd/m^2). These stimuli were labelled as r10, r100, b10, b100, w10 and w100.

The duration of observations was 10 seconds, with the first 2s being a pre-stimulus phase as a rest period, followed by a 1s light pulse and 7s as a restoration phase. Pupil diameters were measured in mm at 60Hz using a system developed by some of the authors [11]. PLRs for each stimuli were observed in single trials using a repeated-measure design.

Examples of measurements for a healthy individual and for an AD patient are shown in Figure 1. In these figures, PLRs are illustrated in response to 6 stimuli, namely the 3 colours and two levels of brightness.

III. Results

1) Feature definitions: A typical PLR waveform shape is illustrated in Figure 2. In the figure, the light pulse overlaps for a period of 2∼3 seconds. As the figure illustrates, there are pupillary response delays due to the shrinking of pupil and its restoration to normal size.

Some features are extracted to specify the PLR response, and these variable features are summarised in Table II. They are pupil size, velocity of pupillary change, duration of change, and integration of the waveform. These features are calculated for each PLR response.

A. Comparison of features

The extracted features for each stimulus were compared between two groups. The results are summarised in Table III.

TABLE II
Definitions of PLR features

Variable	Definition & notes
ps_base	Mean of the pupil size for time before light pulse
ps_lon	Pupil size where light pulse on
ps_loff	Pupil size where light pulse off
ps_min	Minimum pupil size
RA	Range of pupil size (ps_lon - ps_min)
v_con	Max amplitude of pupil construction velocity
v_rest	Max amplitude of pupil re-constriction velocity
ac_max	Min amplitude of pupil acceleration
t_delay	Pupil response delay
t_min	Time when “ps_min” appears
t_v_con	Time when “v_con” appears
t_v_rest	Time when “v_rest” appears
int_con	Integration of constriction phase
int_rest	Integration of restoration phase
int	Overall integration (int_con + int_rest)

When there is a significant difference between pairs of values, the values are displayed in bold face. As the table shows, there are many significant pairs for the b100 and r10 conditions, but few pairs for the white stimulus. In regards to the significant differences for the b100 condition, such as pupil size, velocity and acceleration of pupillary change, the pupil size for AD patients is relatively small, and responds slowly.

All participants are elderly, and in addition to being AD patients, their ages may affect pupil responses. Therefore, the effect of two factors (participant group and age level) on pupillary changes is examined using two-way ANOVA. The variables with deviations which contribute most significantly are selected and summarised in Table IV using means across age levels. These means change along with age levels. Most variables selected are related to velocity and time delay. Since there are few significant interactions between the two factors, they may be independent of each other in regards to PLR features. Additionally, most significant differences between age levels appeared for white stimulus and most differences between two groups occurred when w100 light was used. These results may be related to the mechanism of the PLR, and thus a detailed analysis of this will be a topic of our further study.

B. Factor analysis

There are significant differences in some of the PLR features of the AD and control groups. Though their variables exhibit the qualitative tendencies of a change, their sources could not be determined, as the physical variables and measurement units are completely different; some are expressed using sizes and others are expressed as velocities or accelerations.

Factor analysis was used to extract latent sources of the variables which were measured repeatedly. Since overall integration (int) is a summation of two parts such as int_con and int_rest, it has been omitted during the analysis, thus 14 variables were measured. A two factor structure was estimated using a principal component solution and a screw plot. A factor loading matrix using Promax rotation was produced, as shown in Table V.

TABLE III
MEANS OF PLR FEATURES

Feature	b10(N=19)		b100(N=17)		r10(N=18)		r100(N=19)		w10(N=19)		w100(N=19)	
Variable	Control	Patient	Control	Patient	Control	Patient	Control	Patient	Control	Patient	Control	Patient
ps_base	19.88	16.69	**20.51**	**14.21**	**20.16**	**14.37**	19.59	16.41	19.2	16.40	19.20	16.46
ps_lon	20.24	16.72	**20.81**	**14.18**	**20.29**	**14.22**	20.00	16.26	19.56	16.54	19.45	16.46
ps_loff	10.90	8.02	**9.76**	**6.76**	13.02	9.29	10.15	7.68	11.61	8.80	9.14	7.03
ps_min	10.68	7.59	**9.20**	**5.88**	12.75	8.96	9.76	7.21	11.31	8.56	8.77	6.29
RA	9.56	9.13	**11.60**	**8.30**	7.53	5.27	10.24	9.05	8.26	8.08	10.69	10.17
v_con	-0.46	-0.47	**-0.51**	**-0.31**	**-0.36**	**-0.22**	-0.44	-0.37	-0.40	-0.40	-0.48	-0.42
v_rest	0.13	0.10	**0.18**	**0.08**	**0.13**	**0.08**	0.16	0.10	0.14	0.10	0.14	0.11
ac_max	-0.06	-0.07	**-0.07**	**-0.04**	**-0.05**	**-0.03**	-0.06	-0.05	-0.06	-0.05	-0.07	-0.06
t_delay	0.24	0.26	0.23	0.24	0.25	0.26	0.24	0.25	0.26	0.28	0.23	0.23
t_min	**1.13**	**1.31**	**1.30**	**1.43**	1.09	1.26	1.22	1.36	1.10	1.22	**1.26**	**1.41**
t_v_con	0.34	0.35	**0.33**	**0.36**	0.35	0.39	0.37	0.37	0.37	0.38	0.32	0.33
t_v_rest	1.84	1.78	1.91	2.14	1.65	2.06	1.88	2.11	1.67	1.96	1.74	1.98
int_con	293.1	277.0	**345.0**	**211.0**	**230.0**	**144.5**	301.5	260	249.4	241.6	330.5	300.3
int_rest	748.5	823.0	1015.9	802.7	532.1	444.1	849.8	796.9	612.6	652.3	906.9	941.7
int	1041.6	1100	1361	1014	762.1	588.7	1151.3	1056.9	862.0	893.9	1237.3	1242.0

pairs of **bold means** show significant differences

TABLE IV
AGE AFFECTED FEATURES (LIST OF SIGNIFICANT VARIABLES)

		age levels		
Stimulus	Variable	L	M	H
b10	ac_max	-.11	-.05	-.04
b100	ps_min	8.0	6.9	10.9
r10	t_v_con	0.37	0.36	0.43
r100	t_delay	0.23	0.24	0.31
	int_con	381.4	249.5	176.3
w10	RA	11.2	7.9	4.7
	v_con	-.64	-.33	-.25
	ac_max	-.09	-.04	-.03
w100	RA	14.3	9.6	7.6
	v_con	-.64	-.39	-.33
	t_v_con	0.28	0.34	0.36
	int_con	459.6	280.8	217.9
	int_rest	1279.1	853.7	659.8
	int	1738.8	1134.5	877.7

Age factor is significant ($p < 0.05$)

TABLE V
FACTOR LOADING MATRIX FOR PLR FEATURES WITH PROMAX ROTATION

Variables	Factor1	Factor2
ps_base	**0.596**	**0.586**
ps_lon	**0.596**	**0.586**
int_rest	**1.007**	-0.271
RA	**0.997**	-0.064
int_con	**0.988**	0.006
v_con	**-0.906**	-0.067
ac_max	**-0.902**	-0.021
t_v_con	**-0.737**	0.110
t_delay	**-0.712**	0.062
v_rest	**0.643**	0.099
ps_min	-0.111	**1.025**
ps_loff	-0.079	**0.999**
t_min	0.086	**-0.643**
t_v_rest	0.066	**-0.438**
Contribution ratio(1)	0.42	0.21
Contribution ratio(2)	0.52	0.32
Correlation between factors		r=0.38

(1): Each factor with other factors eliminated
(2): Each factor with other factors ignored

The fundamental variables of each participant, such as pupil sizes commonly contribute to both factors. The second factor contains variables which are concerned with features of the progress of restoration of the pupil after a pulse of light, and the first factor contains the remaining variables of the features of PLR. As mentioned above, both factors contain two variables, and there is a significant correlation between these two factors ($r = 0.38$). Since even the contribution ratio of each factor when the other factors are eliminated is over 60%, the two factors can account for the deviation.

The factor scores ($factor1$,$factor2$) were calculated using the factor loading matrix, and their means for each stimuli are summarised in Figure 3 according to group. The horizontal axis indicates the first factor, and the vertical axis indicates the second factor. The error bars show standard errors. The two groups are indicated using suffixes, such as "p" for patients, or "c" for the control group.

When means between the two groups are compared for red (r10 and r100) or white (w10 and w100) stimulus lights, they are located in proximity to each other, but the means of the two groups are relatively far apart from each other for blue stimulus lights (b10 and b100). Pupil reactions in response to light colours are represented in Figure 3. In the results of applying t-tests to pairs of the means for the two subject groups, there are significant differences in the first factor scores for b100 ($t(15) = 2.64$, $p < 0.05$), in the second factor scores for b100($t(15) = 2.88$, $p < 0.05$) and for w100($t(17) = 2.15$, $p < 0.05$). Also, differences without much significance were observed for the second factor scores for b10($t(17) = 2.07$, $p < 0.10$) and for r10($t(17) = 1.77$, $p < 0.10$).

The results show that between the two subject groups there are significant differences in both factor scores for the b100 condition. As Table III shows many significant differences, the factor scores also represent these differences. Also, the second factor seems to reflect the difference in features of PLR between the two groups.

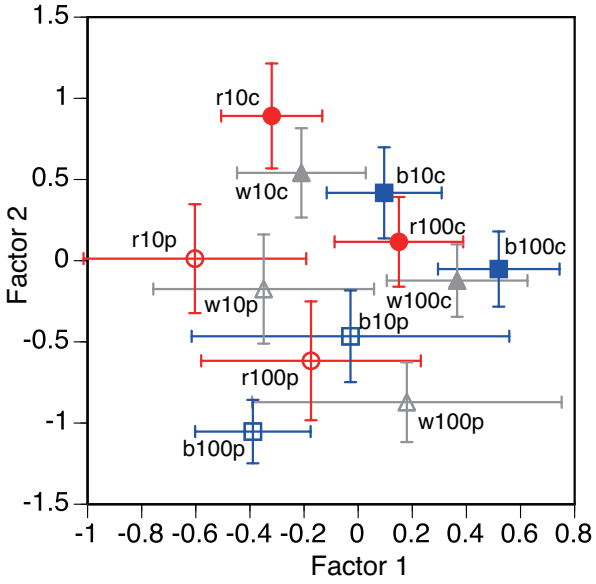


Fig. 3. Distributions of factor scores with error bars showing standard errors

TABLE VI
PERFORMANCE OF MODELS

Stimulus	AIC	R^2	Accuracy	AUC
b10	25.16	0.26	82.1	0.82
b100	17.24	0.47	89.4	0.89
r10	27.33	0.17	73.8	0.74
r100	28.05	0.14	72.6	0.73
w10	27.35	0.18	69.0	0.69
w100	24.70	0.28	79.8	0.80
b(10+100)	21.07	0.48	89.4	0.89
r(10+100)	30.26	0.22	71.4	0.71
w(10+100)	22.94	0.47	90.5	0.91
b+r	18.50	0.73	100	1.00
r+w	29.70	0.50	95.2	0.95
b+w	18.00	0.73	100	1.00
b+r+w	26.00	0.73	100	1.00

Though the influence of age level on factor scores was analysed, it did not affect either factor.

C. Introducing logistic regression

As mentioned in the introduction, this paper introduces logistic regression analysis to estimate the probability of diagnosing AD patients using a binary response variable (p) and PLR features. Here, $p = 1$ for the control subject and $p = 0$ for the AD patient, then p is given by the following equation with logit function.

$$\hat{y}_i = a + b_1 factor1_{i,j} + b_2 factor2_{i,j}$$
$$p_i = \text{logit}^{-1}(\hat{y}_i) = \frac{1}{1 + \exp(-\hat{y}_i)}$$

Suffix i represents the subject, and suffix j represents the stimulus light condition.

Logistic regression analysis was applied to the above factor scores for several conditions, such as the two factor scores for a specific stimulus light condition (light intensity: 10 or 100),

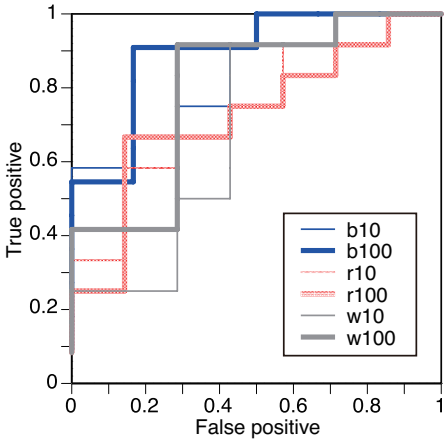


Fig. 4. Comparison of ROC curves

four factor scores for a colour condition (blue, red and white including light intensity 10 and 100) and further combinations such as two colours or all conditions (6 conditions and 2 factor scores). Every model was evaluated for fitness of model using AIC (Akaike Information Criteria), and prediction accuracy using R^2. The accuracy was measured using an appropriate threshold, and performance was summarised using two dimensional metrics such as true positive and false positive. The relationships are then illustrated as ROC (Receiver Operating Characteristics) curves. Figure 4 shows ROC curves of every stimulus condition. The surface area of the curve is also a measure of the AUC (the area under the ROC curve). Their indices are summarised in Table VI.

Results of analyses suggest that discriminant performance is higher for blue light stimuli, in particular for the b100 condition. The ROC curves show step-wise changes, since the number of participants influenced the results. However, the performance of AUC for b100 produced the highest reaction of any single stimulus condition.

D. Model parameter estimation

As Table III shows the possibility of discriminating between AD patients and control group subjects using a logistic regression function. However, the parameters of these functions can not be estimated sufficiently because the amount of data is too limited. Here, the Markov chain Monte Carlo (MCMC) method was introduced for more accurate estimation of the parameters. In regards to the data generation procedure using the MCMC technique, the burn-in period was 2000 and the number of samples was 10000 [12]. Using this procedure for the b100 condition, the parameters are estimated as follows.

$$\hat{y}_i = 5.4229 + 0.4722 * factor1_i + 7.3801 * factor2_i$$

The magnitude of coefficient for the second factor ($factor2$) is 15 times that of the coefficient for the first factor ($factor1$).

As an additional example, the parameters for blue light stimuli were estimated as follows:

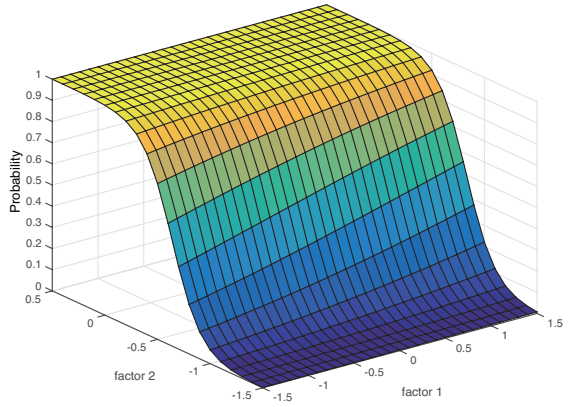


Fig. 5. Probability distribution for b100

$$\hat{y}_i = 9.1664 + 2.5320 * factor1_{i,b10} - 1.1366 * factor2_{i,b10} \\ -1.4011 * factor1_{i,b100} + 12.0891 * factor2_{i,b100}$$

The magnitude of these parameters suggests that the coefficient for the second factor at a high level of brightness (b100) is relatively larger than for others (b10 and the first factor at a low level of brightness). As the patterns of coefficients depend on the light stimulus, the wavelengths of the stimuli may affect these reactions.

In regards to the discussion in the previous section, the probability of control group subjects (1) or AD patients (0) may be illustrated using two dimensional information ($factor1$ and $factor2$), as shown in Figure 5. Figure 5 shows that the probability distribution against patients with AD depends mainly on the scores of the second factor. Also, the score of the first factor helps to more finely adjust the probability during the period where the curve is steep.

In regards to the experimental procedure, the features of PLR can be measured best during a one second high brightness level pulse of blue light (b100). Also, factor scores were calculated using the factor loading matrix shown in Table V. Finally, the probability of diagnosing AD patients can be predicted using the function above.

However, the possibility of developing a more flexible procedure for use in future experiments involving additional new participants will be a subject of our further study.

IV. Summary

This paper presents a procedure for predicting the probability of diagnosing AD patients using features of PLR, which respond to the activities of ipRGCs.

Three colour lights at two levels of brightness were illuminated for 1 second, and pupil light responses were observed. 15 features were extracted from each PLR, and two factor scores were calculated using a factor loading matrix. The following results were produced.

1) There are significant differences in some features between AD patients and control group subjects, in particular for the b100 condition. Also, for a few features for white light there are significant differences between age levels.
2) Logistic regression analysis was introduced to discriminate AD patients from the control group using two factor scores in response to chromatic stimuli. The performance was evaluated using the indices of the fitness of equations. As a result, the performance for b100 was the highest.
3) The MCMC technique was introduced to estimate the parameters of the regression functions. The model provides a distribution of probability for AD patients and the control group.

The validity of the probability estimations should be confirmed using the PLR data of patients. This will be a subject of our further study.

Acknowledgment

Polish Ministry of Science and Higher Education research grant NN518 405338 partially supported this research.

References

[1] D. F. Fotiou, V. Setergiou, D. Tsiptsios, C. Lithari, M. Nakou, and A. Karlovasitou, "Cholinergic deficiency in Alzheimer's and Parkinson's disease: Evaluation with pupillometry," *International Journal of Psychophysiology*, vol. 73, pp. 143–149, 2009.

[2] D. M. Bittner, I. Wieseler, H. Wilhelm, M. W. Riepe, and N. G. Müller, "Repetitive pupil light reflex: Potential marker in Alzheimer's disease?" *Journal of Alzheimer's Disease*, vol. 42, pp. 1469–1477, 2014.

[3] P. D. Gamlin, D. H. McDougal, and J. Pokorny, "Human and macaque pupil responses driven by melanopisn-containing retinal ganglion cells," *Vision Research*, vol. 47, pp. 946–954, 2007.

[4] A. Kawasaki and R. H. Kardon, "Intrinsically photosensitive retinal ganglion cells," *Journal of Neuro-Ophthalmology*, vol. 27, pp. 195–204, 2007.

[5] M. Nakayama, W. Nowak, H. Ishikawa, K. Asakawa, and Y. Ichibe, "Discovering irregular pupil light responses to chromatic stimuli using waveform shapes of pupillograms," *EURASIP J. in Bioinformatics and System Biology*, no. #18, pp. 1–14, 2014.

[6] T. Yoshida, K. Ohno-Matsui, S. Ichinose, T. Sato, N. Iwasa, T. C. Saido, T. Hisatomi, M. Mochizuki, and I. Morita, "The potential role of amyloid β in the pathogenesis of age-related macular degeneration," *The Journal of Clinical Investigation*, vol. 115, no. 10, pp. 2793–2800, 2005.

[7] J.-D. Ding, J. Lin, B. Mace, R. Herrmann, P. Sullivin, and C. Rickman, "Targeting age-related macular degeneration with alzheimer's disease based immunotherapies: Anti-amyloid-β antibody attenuates pathologies in an age-related macular degeneration mouse model," *Vision Research*, vol. 48, pp. 339–345, 2008.

[8] K. Ohno-Matsui, "Parallel findings in age-related macular degeneration and alzheimer's disease," *Progress in Retinal and Eye Research*, vol. 30, pp. 217–238, 2011.

[9] J. M. Sivak, "The aging eye: Common degenerative mechanisms between the alzheimer's brain and retinal disease," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 1, pp. 871–880, 2013.

[10] W. Nowak, M. Nakayama, M. Pieniążek, and A. Hachoł, "Feature analyses of pupil light reflex to chromatic stimuli in alzheimer's patients," in *Proceedings of 2nd International Conference on Frontiers of Signal Processing*, 2016, pp. 58–62.

[11] W. Nowak, A. Żarowska, E. Szul-Pietrzak, and M. Misiuk-Hojło, "System and measurement method for binocular pupillometry to study pupil size variability," *BioMedical Engineering Online*, vol. 13, no. #69, pp. 1–16, 2014.

[12] Sas/stat 13.1 user's guide, the mcmc procedure. [Online]. Available: http://support.sas.com/documentation/onlinedoc/stat/131/mcmc.pdf

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 279–282
DOI: 10.15439/2018F335
ISSN 2300-5963 ACSIS, Vol. 15

Towards Amblyopia Therapy Using Mixed Reality Technology

Adam Nowak, Mikołaj Woźniak, Michał Pieprzowski, Andrzej Romanowski
Institute of Applied Computer Science
Lodz University of Technology
90-924 Lodz, Poland
email: 203151@edu.p.lodz.pl, mikolaj@pawelwozniak.eu, 210893@edu.p.lodz.pl, androm@iis.p.lodz.pl

***Abstract*—This paper presents an approach towards aiding the rehabilitation exercises in amblyopia care using mixed reality technology. The Lazy Eye Syndrome is tackled here through an interactive holographic application implemented on Microsoft HoloLens device. It provides an entertaining way for the handicapped eye workout as it is based on a simple game of skill. The game is designed in a way that the majority of aware-requiring objects and events are displayed for the cured eye only, remaining the other eye responsible for background and additional information perception. Such disproportion forces an increased activity of the lazy eye, which is to perform more movements and impose the brain to process the sight more extensively. The proposed prototype is an extension of a novel approach towards treating amblyopia, employing software-based stimulation techniques, which could be easily adapted to various age and ability correlated needs of the user, with minimal requirements regarding the exercise setting and preparation.**

I. Introduction

THE amblyopia, often referred as the lazy eye syndrome is one of the most frequently diagnosed vision disorders among children. The cause of this affliction lays in disrupted cooperation between the eye and the brain. The brain tends to prefer the other eye for visual perception, therefore one of the eyes performs significantly less movements and activity, which causes further decrease of vision in the handicapped eye [1].

The lazy eye syndrome, if not properly treated during childhood, is likely to maintain an issue during adulthood. Such disease prevents an individual from obtaining professional driving licenses, which significantly limits the job opportunities. Moreover, the improper balance of eye activity might increase the risk of other eye injuries and diseases for the healthy eye [2].

The traditional approach of treating amblyopia is to address children only, as the therapy provides best efficiency during the so-called window of visual cortex development. The most common method prescribes putting a patch over the properly working eye and performing high eye-awareness requiring activities, forcing the brain to perform with the disabled eye more intensely. Such approach brings satisfactory results when applied to children in the age between 5 and 8 years old. However, the method is being widely discussed, as it employs highly artificial conditions of eye operation, with no regard of the stereoscopic vision.

Furthermore, noncompliance to wearing the patch (full-time or part-time, accordingly to the diagnosis) drives the treatment to failure. Therefore, the commitment required from the children, as well as their parents - as the proper supervision is necessary for that method, is significantly affecting daily life. Moreover, wearing a patch may cause social anxiety in the peer group and pose a huge challenge in performing some of school and educational tasks.

Thompson et. al. concerned about adult treatment, where the main problem seemed to be the lack of brain plasticity. Experiment with repetitive Transcranial Magnetic Stimulation proved that contrast sensitivity may be improved in the amblyopic eye by regular 10 minutes sessions with rTMS. This is the evidence that neuronal plasticity is not fully attenuated in adulthood. [3]. Further, it is suggested that amblyopia has a binocular nature, being a disrupt of analysing the signals perceived by both eyes. Having provided extended time of viewing and different contrast for each eye, the improvement of vision is noticed [4]. Therefore, the binocular manner of exercising may be perceived as most favorable and beneficial.

In this paper, we would like to propose low-commitment approach towards exercising the amblyopic eye, suitable for both children and adults (yet dedicated mainly to youngster users) concerning workout sessions in home setting.

II. Related Work

Multiple attempts have been made towards alleviating the constraints of traditional amblyopia treatment. One of the most promising approaches is to employ video games for stimulating the amblyopic eye through display design, with no constraint of losing the stereoscopic vision. Li et al. [5] prove that playing video games utilizing the lazy eye significantly develop the fundamental vision functions of the patient [5]. The improvements to a different extent were found for both visual and positional acuity, spatial attention, and the most significant for stereopsis. Cross-over study shown that occlusion approach cannot provide improvements in all those areas.

Those conclusions inspired further development of a tool which could serve as the exercise equipment for game-based treatment. Eastgate et. al. [6] proposed an integrated system, employing VR technology, enabling to practise through

playing various 2D and 3D games, stimulating the vision in different manners, as well as implemented a control view for the clinician. The binocular system provided an extensive setup for game-based lazy-eye treatment, while its high-complexity and ambient requirements made it available only for clinical use, through organised sessions. One of recent reports related to use of Oculus VR technology for adult patients is [7]. Such manner of operation does not solve the high-commitment constraint, as forcing daily visits to the clinic pose additional challenge to the patient's everyday routine.

Bringing the therapy to patient's home seemed to be an ultimate goal in a way to reducing the commitment and obstacles present in the process. Birch et. al.[8] tried to answer this need by proposing a binocular system based on Apple iPad applications. This approach enabled users to exercise in home setting, while being constrained with limited support of optic equipment available for iPad. The study has shown that playing the games using binoculars brought favorable results while being combined with traditional patch-based treatment. Gargantini et al. [9] took another endeavour towards providing affordable solution of home-exercising of the lazy eye syndrome, using Google Cardboard and the smartphone application. This solution differs the level of details displayed in each eye, forcing the amblyopic one to process more information than the healthy eye. The application also supported additional features for doctors. The main advantage of this solution is low-cost implementation for commonly used devices.

A commercial approach to bringing the VR treatment experience to patient's home has been made by Vivid Vision [10]. This solution employed commercially accessible technologies such as Oculus Rift, HTC Vive, Samsung Gear VR. However, even the end-user VR solutions pose high requirements towards the setting of playing, as well as disabling the awareness of any outer factors. In terms of children usage, the setup requires parental supervision, as the risk of undesired action is decent (eg. walking into furniture); see Fig 1. Therefore, we can establish the need for a solution which possess the advantages of VR projection while remaining ambient awareness of the performed activity. Mezad-Koursh et. al. [11] prove the necessity of daily exercising in successful treatment through pilot study of home use of a similar system, based on watching animated television using binoculars. The regular training enables improvements of the amblyopic eye activity even for children older than the assumed visual cortex development window.

Our attempt aims to bring the convenience of mobile solution [8], supported with high eye-involvement and immersive experience of the [6].

III. HoloLens Implementation

The proposed solution employs the Microsoft HoloLens, which is the leading commercial equipment for augmented reality projections. Due to AR approach, the displayed projection can be easily adapted to the ambient conditions, both in terms of visuals and their color contrasts, but also through real-object recognition and automated space-aware algorithm, so the artifacts will not be displayed on the physical obstacles, which becomes a major issue concerning the solution might be used by children.

Unlike the traditional, full-covering VR technology, the HoloLens offers undisturbed perception of the surrounding. As a consequence, the risk of undesired physiological reactions is significantly lower than for full-covering VR [12]. What is more, the experience of full immersion into VR activities is an additional challenge to the brain, which is naturally opposing the perception through display-equipped eyes and other senses. This effect is especially not recommended to the children [13]. Such effects are not observed for augmented reality displays, as the frame of reference is not altered. Therefore, it is more advantageous not to distract the brain with the senses' counteractions and maintain focus on extended activity of the amblyopic eye. The lack of harmful side effects of using the system is especially important due to the specific of the amblyopia curing procedure - the eye shall be forced to operate with enhanced activity for longer periods of time.

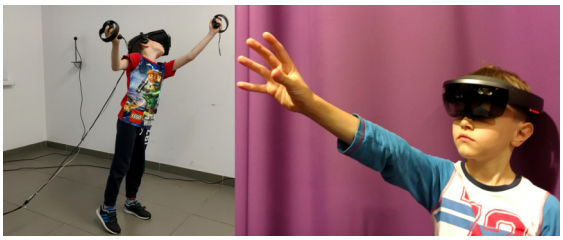


Fig. 1. 8yo user operating VR app (left) and the HoloLens app (right). The AR game is being displayed while maintaining the user fully space-aware.

Another advantage of HoloLens implementation is lack of problems with calibrating the display for children. Fully-covering VR setups are likely to present improper behavior when calibrated for children [14]. The device may lose the focus point, so the sight does not precisely follow the movements of user's head. Another advantageous feature of the HoloLens approach is the independent control of the contents displayed to each eye. Therefore, full control on the balance of the object projected for each eye is maintained and may be adjusted accordingly to the clinical diagnosis, to stimulate the lazy eye with different difficulty/intensity, while supporting the skill of stereo-vision.

Microsoft HoloLens supports streaming of the video displayed for user to any web browser. Thereby other persons may see the real-time cast of the holographic projection. This feature might be advantageous in supervision of the treatment process, as the doctor can dynamically monitor the patients activity. Patient can be therefore guided on specific actions performed within the solution, which is supportive for more complex exercise. Moreover, the clinician is able to assess the performance of the user in completion of the holographic tasks. Therefore the level of difficulty can be easily adapted to skills and abilities of the particular user.

Furthermore, the ability of the HoloLens to record the displays throughout the holographic session might be an asset for improving the exercise scenarios for future patents and enable extensive and precise analysis of users' performance and the results of the therapy applied.

IV. The exercise application

A simple game has been implemented, in which the user moves a spacecraft trying to omit falling asteroids. The space ship reacts to the movements of user's head. Asteroids are falling down one set after another with constant speed. One can adjust number of asteroids in a single set and the time interval between them. The artifacts displayed have been divided accordingly to the importance and awareness requirements to the user.

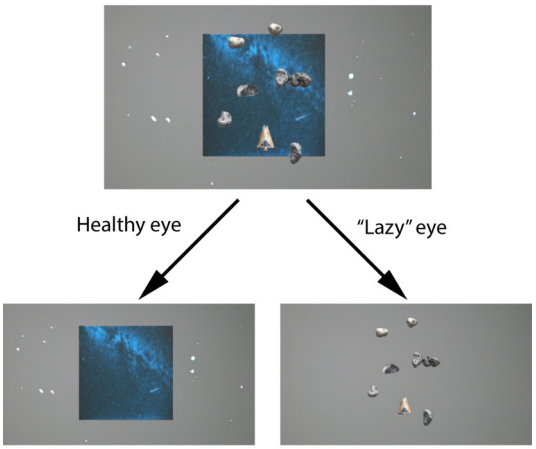


Fig. 2. Exemplary image displayed in the app. Left: the healthy eye part, containing the bacground. Right: the lazy eye part, containing roaming objects in the display.

All the important objects, such as spacecraft controlled and falling asteroids are displayed by single eye projection only - the one set to operate for the amblyopic eye, whereas supplementary objects, such as background and navigation panel are displayed in the healthy-eye part. Therefore, the user's brain is forced to operate the lazy eye to participate in the game. The character of the game imposes high activity of the eye treated. Simultaneously, the healthy eye is not excluded from the perception, unlike in the traditional treatment approach. Therefore, the game acts as an exercise for restoring the proper stereo vision.

Moreover, the exercise is presented as a form of entertainment, which makes the treatment process more pleasant, encouraging to greater compliance to the prescribed training routine. Regarding that the lazy-eye syndrome therapy concerns mostly children, it is even more desirable to provide an enjoyable manner of exercising. Working version of the system was tested on 4 adults (38yo on average) and 4 children (6.5yo on average), including two patients with amblyopia; one in each group. All of the participants were rather new to mixed reality technology. After initial training (c.a. 5 min for adults and c.a. 2 min for children) participants were able to play the game successfully. Children reported more initial problems with starting as well as during the normal operation in the game, yet did not reported problems within the game itself. Adult users reported less problems with starting the game but assessed the play as difficult or requiring too quick reactions for them to follow. Amblyopia child assessed game to be more demanding in terms of both effort as well as initial training in comparison to his experience (very limited, yet wider) with VR Ocullus Rift version of Vivid Vision app. All of the adults as well as all of the children reported more comfort when using HoloLens comparing to VR experience. However, we cannot present any structured nor longitudinal study results so far neither in terms of app assessment nor in terms of therapeutic matters.

V. Discussion

The proposed solution employs recent findings in the lazy-eye syndrome treatment in terms of using modern visual interfaces towards stimulating the disabled vision apparatus. Similar solutions have been offered before, while being aimed mostly for clinical or supervised use. Our system enables the participant to reach more self-reliance during the treatment and make the exercise procedure highly more convenient. The supervised mode is optional and requires additional device to which the sight from HoloLens could be casted.

The efficiency of the exercise is expected to be comparable to that obtained using traditional, full-covering VR approach. However, reduced number and intensity of possible side effects, like digestive discomforts, vertigo and labyrinth disorders is expected [15]. Due to limited projection range and the holographic technology applied, there is less risk of high eye-exhaustion or eye strain syndromes to occur [16]. The proposed solution is the next step towards providing the accessible and convenient tool for home-based amblyopia treatment-supporting procedures. The systems offers an entertaining manner of exercising the disabled eye, with additional activity monitoring features for enhanced supervision. The suggested approach lacks of common disadvantages of the arrangements used so far - those are the undesired side effects, inability to use the system for a long time, complex operation and/or difficult setting requirements.

VI. Future Work

This paper presents a prototype software for the lazy-eye stimulation. Further improvements of the game interface; i.e. more entertainment/gamification features are planned in order to maintain interest of the user for long-period treatments. Additional context-aware sensing may be advisable, for precise assessment of the physical objects in sight and the distance between them and the player, so to obtain the system which could revise itself in order to provide as safety playing conditions as possible, which is a crucial issue concerning the children being a decent group within the target users. Furthermore, the device might be supplied with the gaze tracking system and help in controlling user interfaces [17] [18]. Regarding the efficiency

of using the system in terms of clinical improvements in eye operation, it is necessary to provide a long-term study with a satisfactorily large sample and the control group to assess system's efficiency as a treatment tool. Its' operation is expected to be similarly effective as the VR-based solutions, with the extended support for stereo-vision development. Yet it is expected to overperform VR versions since AR technology is superior in terms of environment awareness and therefore lack of nausea effect and better suitability to young brains development. It would be interesting to employ this technology and similar app design to other settings such as the industrial environments, where similar features (distinct eye display mode) could be utilized for more efficient data analysis and flow instalations control parameters visualisation [19] [20] [21] [22]. The approach may also become useful considering modern visualisation systems for control purposes [23]. [24]. Perhaps most interesting and challenging would be to use the AR technology to implement in hybrid systems such as context-aware data processing or crowdsourcing applications [25] [26] [27].

VII. Conclusions

This work shows a new approach to using computer vision technology for amblyopia therapy. Namely, mixed reality paradigm is applied in order to overcome disadvantages of virtual reality displays that already proven their potential for lazy eye treatment. Therefore, the proposed approach using Hololens device enables to exercise the weaker eye without full immersing into the virtual world. This is of a key advantage over the VR technology especially for children under 14 due to immature nervous system, middle ear bone labyrinth and eye fundus development.

References

[1] J. M. Holmes and M. P. Clarke, "Amblyopia," *The Lancet*, vol. 367, no. 9519, pp. 1343 – 1351, 2006.

[2] P. Waddingham, S. Cobb, R. Eastgate, and R. Gregson, "Virtual reality for interactive bi nocular treatment of amblyopia," in *Proc. 6th Intl Conf. Disability, Virtual Reality & Assoc. Tech., Esbjerg, Denmark,*, 2006.

[3] B. Thompson, B. Mansouri, L. Koski, and R. Hess, "Brain plasticity in the adult: Modulation of function in amblyopia with rtms," *Current Biology*, vol. 18, Issue 14, 2008.

[4] R. F. Hess, B. Mansouri, and B. Thompson, "A new binocular approach to the treatment of amblyopia in adults well beyond the critical period of visual development," *Restorative Neurology and Neuroscience 28 1–10*, 2010.

[5] R. Li, C. Ngo, and D. L. J. Nguyen, "Video-game play induces plasticity in the visual system of adults with amblyopia," *PLoS Biol 9(8)*, 2011.

[6] R. Eastgate, "Modified virtual reality technology for treatment of amblyopia," *Eye volume 20, pages 370–374*, 2006.

[7] P. Žiak, A. Holm, J. Halička, P. Mojžiš, and D. P. Piñero, "Amblyopia treatment of adults with dichoptic training using the virtual reality oculus rift head mounted display: preliminary results," *BMC Ophthalmology*, vol. 17, p. 105, Jun 2017.

[10] J. Blaha, "Initial study results indicate vr game is effective in improving vision in people with lazy eye," 2015.

[8] E. Birch, "Binocular ipad treatment for amblyopia in preschool children," *Journal of American Association for Pediatric Ophthalmology and Strabismus JAAPOS , Volume 19 , Issue 1 , 6 - 11*, 2014.

[9] A. Gargantini, "A low-cost virtual reality game for amblyopia rehabilitation," in *REHAB '15 Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques*, 2015.

[11] D. Mezad-Koursh, "Home use of binocular dichoptic video content device for treatment of amblyopia: a pilot study," *Journal of American Association for Pediatric Ophthalmology and Strabismus JAAPOS*, 2016.

[12] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks, Jr., "Physiological measures of presence in stressful virtual environments," *ACM Trans. Graph.*, vol. 21, pp. 645–652, July 2002.

[13] M. S. Maria V. Sanchez-Vives, "From presence to consciousness through virtual reality," *Nature Reviews Neuroscience volume 6, pages 332–339*, 2005.

[14] Z. Pan, A. D. Cheok, H. Yang, J. Zhu, and J. Shi, "Virtual reality and mixed reality for virtual learning environments," *Computers & Graphics*, vol. 30, no. 1, pp. 20 – 28, 2006.

[15] P. C. Grigore C. Burdea, *Virtual Reality Techology*. 2003.

[16] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3d displays," *Displays*, vol. 25, no. 2, pp. 99 – 108, 2004.

[17] A. Wojciechowski and K. Fornalczyk, "Exponentially smoothed interactive gaze tracking method.," *Springer, Cham, In International Conference on Computer Vision and Graphics (pp. 645-652)*, 2014.

[18] G. Glonek and A.Wojciechowski, "Hybrid orientation based human limbs motion tracking method," *Sensors*, vol. 17(12), 2017.

[19] A. Romanowski, "Big data-driven contextual processing methods for electrical capacitance tomography," *IEEE Transactions on Industrial Informatics*, vol. doi:10.1109/TII.2018.2855200, p. in press, 2018.

[20] K.Grudzien, A.Romanowski, D.Sankowski, and R.A.Williams, "Gravitational granular flow dynamics study based on tomographic data processing," *Particulate Science and Technology*, vol. 26, no. 1, pp. 67–82, 2008.

[21] V. Mosorov and D. Sankowski, "Estimation of the rotation angle of gas/solid swirl flow by subpixel image resizing," *Asia-Pacific Journal of Chemical Engineering*, vol. 13, no. 2, p. e2177, 2018.

[22] K. Grudzien, "Visualization system for large-scale silo flow monitoring based on ect technique," *IEEE Sensors Journal*, vol. 17, pp. 8242–8250, December 2017.

[23] M. Woźniak, A. Polak-Sopińska, A. Romanowski, K. Grudzień, Z. Chaniecki, A. Kowalska, and M. Wróbel-Lachowska, "Beyond imaging - interactive tabletop system for tomographic data visualization and analysis," in *Advances in Manufacturing, Production Management and Process Control* (W. Karwowski, S. Trzcielinski, B. Mrugalska, M. Di Nicolantonio, and E. Rossi, eds.), (Cham), pp. 90–100, Springer International Publishing, 2018.

[24] A. Romanowski, K. Grudzien, Z. Chaniecki, and P. Wozniak, "Contextual processing of ECT measurement information towards detection of process emergency states," in *Hybrid Intelligent Systems (HIS), 2013 13th International Conference on*, pp. 291–297, 2013.

[25] C. Chen, P. W. Woźniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, and M. Fjeld, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 52:1–52:25, 2016.

[26] I. Jelliti, A. Romanowski, and K. Grudzien, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *FedCSIS'16, ACSIS, vol. 8. IEEE*, p. 1613–1619.

[27] A. Romanowski, "Contextual processing of electrical capacitance tomography measurement data for temporal modeling of pneumatic conveying process," in *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, FedCSIS'18, ACSIS, vol. IEEE*, p. in press.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 283–286
DOI: 10.15439/2018F171
ISSN 2300-5963 ACSIS, Vol. 15

Contextual processing of electrical capacitance tomography measurement data for temporal modeling of pneumatic conveying process.

Andrzej Romanowski
Institute of Applied Computer Science
Lodz University of Technology
Lodz, 90-924 Poland
Email: androm@iis.p.lodz.pl

***Abstract*—This work covers deployment of contextual processing of measurement data in application to temporal modeling of pneumatic conveying industrial process. Electrical capacitance tomography (ECT) used as a non-invasive process monitoring tool is supported by data mining for regularization of non-linear inverse problem solution. Processing of a larger number of archived experimental datasets enables extracting additional constraints for inference. Contextual data processing model (CDPM) extracts demanded information from the data in order to incorporate it as an expert knowledge about the process temporal behavior. Then it is incorporated into the Bayesian inference framework. Comparative analysis with previous work and domain expert prepared baseline to the proposed approach is demonstrated. Additionally, simplified parameterization is tested and verified by the quantitative experimental analysis.**

I. Introduction and related work

A. Pneumatic conveying and ECT

Bulk solids, powders and particulates cover about 2/3 of all solid materials used in industry at various stages of manufacturing. However, proper monitoring of processes that involve bulk solids is difficult because of their volumetric and opaque nature. The most promising techniques involve non-invasive and non-intrusive tools such process tomography methods, while electric capacitance tomography is one of the most popular modalities [1] [2]. However, there are some issues related to nonlinear nature of electrical field associated with extracting required process-related information from ECT data [3] [4] [5]. Therefore some methods aiming at improving the inverse problem conditions were developed over last 20 years [6] [7] [8] [9] [10] [11] [12] [13] [14]. Here a contextual tomography-based measurement data processing approach is proposed. It is based on the same contextual data processing model (CDPM) as described in [15] that was valided for big data driven aspects there. In contrast, current work validates CDPM within the scope of inverse problem regularization support. The main contribution of this work is therefore the proposed theoretical model for contextual data processing applied to temporal inference about the process behaviour. It is validated here with binary classification technique comparing to the baseline Bayesian inference framework.

B. Contextual Model for Measurement Data Processing

Contextual methods are derived from a concept of context-aware services or context-awareness in general. These concepts are extensively used in human computer interaction (HCI) discipline. Here it is postulated to outspread these notions to the field of measurement data processing for industrial processes monitoring [16] [17] [18]. Though some modification is required but the core of the concept remains to be based on a simple idea of using the additionally available information describing the object of interest in order to broaden the set of data to be incorporated as the input to the system.

I propose to expand this typical understanding of context

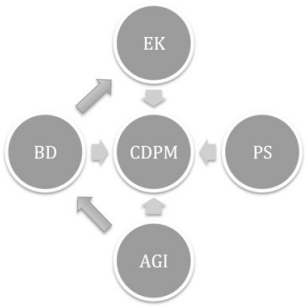


Fig. 1: General diagram of contextual data processing model (CDPM). EK, BD AGI, PS refer to as Expert Knowledge, Big Data, Artificially Generated Input, Peripheral Sensors accordingly.

related to the information coming from peripheral sensors (PS) to the four inter-related categories that form the CDPM model as shown on Fig.1. While PS still stays as substantial pillar of the model, the extension goes towards incorporating the following factors: Expert knowledge (EK), Big Data analysis (BD) and Artificially generated inputs (AGI) into the model. EK refers to any prior knowledge that can be defined independently to the current situation (experiment). BD stands for a broader base of previously conducted experiments or measurement datasets that can be analyzed in order to search for similarities, patterns or knowledge that can be extracted

out of it. AGI works as a complementary tool to supplement knowledge base, especially in case of sporadic phenomena for events that are rarely captured. While AGI can be either a distinct pillar of the CDPM it can also contribute to BD component. While exploring the BD analysis for CDPM is postulated in [15] [19] [20], this paper focuses at showing the method for incorporating EK into the process of inverse problem solving for ECT application to the monitoring of pneumatic conveying flow of bulk solids.

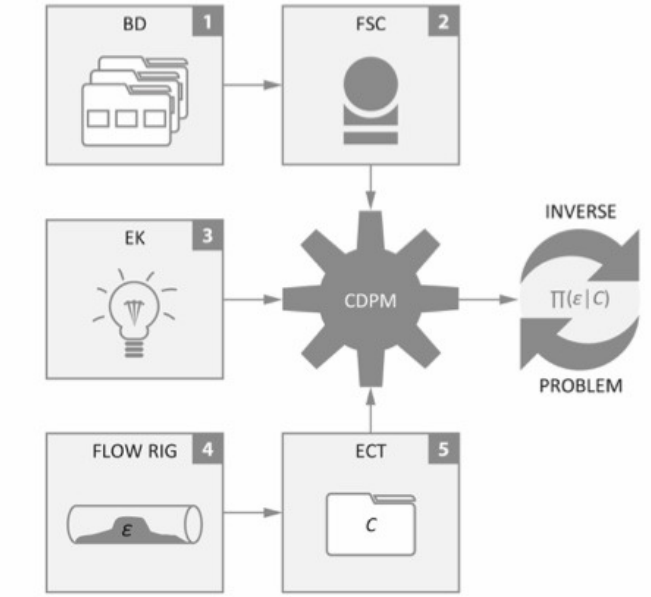


Fig. 2: CDPM for temporal modeling of pneumatic conveying. EK, BD, PS, FSC, ECT, C refer to as Expert Knowledge, Big Data, Artificially Generated Input, Peripheral Sensors, Flow State Classifiers, Electrical Capacitance Tomography, Capacitance data accordingly.

C. Pneumatic Conveying Experimental Setup

Experimental part of this research was conducted at the Tom Dyakowski Process Tomography Laboratory at the Lodz University of Technology. The ECT dual plane, 8-electrode sensors were fixed on a 65mm horizontal section of pneumatic conveying test rig as shown on Fig. 2. More details about the equipment can be found here [18]. Measurement campaign spanned over a range of settings preserving regular slug flow for different combinations of material feed rate, and air pressure (10.0 - 16 Hz inverter, 60 - 100 Hz of the rotary valve) for approx. 3 cubic mm polyamide pellets.

II. Investigation procedure

A. Baseline requirements

This work aims at proving that using available previous experimental data one can derive information useful for current measurement-related computational problem [10]. Spatio-temporal modeling of pneumatic conveying based on Bayesian inference and statistical methods demonstrated possibility of

Fig. 3: Experimental setup: ECT sensor equipped measurement section at horizontal pneumatic conveying rig and a corresponding ECT data acquisition device.

omitting the image reconstruction stage on the way to estimate characteristic flow parameters [8]. Current step is to simplify parametric modeling within the temporal modeling concurrently preserving or increasing the accuracy with the aid of CDPM. Computational environment with use of Hadoop is similar to desribed in [21]. The investigation is based upon the following postulations:

1) Mean concentration of bulk solid is taken as a main parameter describing flow state at any time point. ECT is the main measurement tool to supply estimated electric permittivity distribution (related to bulk concentration) based on capacitance measurement vectors.
2) There is additional information available in form of archived experimental datasets coupled with supplementary information such as estimated flow rate and weighed quantity of total material being transported, material geometry, properties, valves states, other metadata. Fig. 2 illustrates the basic workflow for the inverse problem using CPDM support.
3) Fragments of archive datasets are taken especially for slug rise (ECT recorded slug build up) and fall (ECT recorded slug tail) in order to regularize inverse problem for temporal analysis. Previously geometrical modeling was proposed for temporal smoothing varied in time that resulted in high uncertainty [22].
4) CPDM takes fragments, full-lenght experimental datasets and optional classifiers built on top of BD employment as well as any other general knowledge in order to incorporate it into the EK inverse problem solution as described in [8].
5) Correspondence to baseline, expert-annotated datasets in terms of mean electrical permittivity change in time was proposed as a principal measure to assess the proposed approach accuracy.
6) Calculated total transported material weight (relevant to flow rate) was chosen as an extra measure to verify if the simplification relying on substitution of geometrical modeling with mean concentration change is reasonable.

B. Inference Framework

Eq. (1) shows the approximation of the posterior probability density function related to the unknown distribution of the electric permittivity of the transported bulk solids mixture within the ECT sensor space. The critical factor from the CDPM is the kE which stands for the prior knowledge with relation to the expected constraints on the electric permittivity ϵ distribution in relation to obtained electric capacitance records C and in fact kE(ϵ) is EK equivalent.

$$p(\varepsilon \mid C) \propto p(C \mid \varepsilon) * k_E(\varepsilon) \tag{1}$$

where p(ϵ) is the inverse problem for ECT and bulk solids flow and the p(C—ϵ) is the forward problem that can be numerically approximated using FEM method. Hence the EK can be defined here as the regularization prior, and for an electric permittivity case can be denoted as kE(ϵ) Eq (2):

$$k_E(\varepsilon) \propto \exp\left\{-\beta_s \|\varepsilon\|_l^l\right\}; \beta_s > 0, 0 \le l \le 2. \tag{2}$$

Such stated kE(ϵ) leads to Laplacian distribution with a 1 (l=1) norm and leads to Gaussian distribution for norm 2 (l=2). Now extending this approach to a temporal dependence analysis the following relation expresses how consecutive frames dependence can be described (Eq. 3):

$$\Pi\left(\kappa^t \mid \kappa^{t+1}, \kappa^{t-1}\right) \propto \exp\left\{-\beta_t \left(\kappa^t - \bar{\kappa}^t\right)^2\right\} \tag{3}$$

where Kt is a set of estimated parameters at a given time point t, and βt decides on the level of correlation between these values in consecutive time points (in contrast to βs in a spatial distribution case in Eq. 2). The procedure of tackling the is shown on right hand-side of Fig. 1. There are several possible options for solving the Bayesian-based approach for ECT inverse problem solving. Related work referred to here is based on highly iterative MCMC scheme that is both computationally and time demanding [8] [10]. Current work was decided to use the same option yet thanks to GPU computing and reduced number of parameters the calculations are far less demanding and time consuming as shown in [11].

III. Results

A. Comparative Analysis

The results are given for the modeling of the pneumatic conveying slug flow for several different flow configurations as discussed in experimental setup section. Results are divided into 3 classes with respect to flow rate, i.e. average transported medium rate over time. Table 1 provides the comparison between the compliance of estimated mean concentration of solids (corresponding to amount of medium transported) on the basis of comparison between the raw data, reconstructed images analysis, previous approach [18] and the CDPM model. The datasets are cut to the 100 frames series that always include both the slug and stationary layer portion (i.e. consecutive periods of frames with lower or higher mean concentration values). Each of the 3 categories: low flow rate (Table 1, row 1), medium flow rate (Table 1, row 2) and high flow rate (Table 1, row 3) are arbitrarily divided into 3 consecutively rising classes based on the results and parameters of the performed experiments. Results in rows 1-3 indicate percentage compliance averaged over 10 different calculations and standard deviation both rounded to first decimal digit. Percentage is calculated based on binary classification of a consecutive measurement frame as either belonging or not belonging to a slug as shown in [9]. Row 4 reveals average error. Row 5 shows supplementary measure of total material transported weight comparing to the scale-recorded values for whole experimental datasets.

TABLE I: Pneumatic conveying CDPM results

	Slug flow rates averaged over time	**Results: average classification compliance (accuracy) comparing to the expert ground truth baseline**							
		Raw data		*Reconstructed images*		*Previous approach*		*CDPM*	
		[%]	SD	[%]	SD	[%]	SD	[%]	SD
1	*Low flow rate*	78.6	15.8	91.3	6.9	89.6	7.1	90.4	6.3
2	*Medium flow rate*	81.4	13.6	78.4	16.2	86.9	8.0	92.3	3.8
3	*High flow rate*	92.3	6.1	86.5	11.6	88.8	5.5	94.6	2.3
4	*Average error*	15.9		14.6		11.6		7.6	
5	*Total weight*	89.2	7.7	83.7	10.3	88.6	4.1	91.5	3.2

B. Discussion and directions for future work

The proposed CDPM model using temporal information as an input for expert knowledge (EK) extension to the normal prior knowledge applied in Bayesian inference framework performs well comparing to the baseline, i.e. to the expert provided ground truth in form of marked test datasets. CDPM results for all three classes reached 90%+ accuracy. CDPM performs better than other methods especially for medium and high flow rates for which both reconstructed images post-processing as well as previous work based on geometrical parameterization for temporal modeling obtained the weakest scores. Interesting feature is that the reconstructed images based analysis outperform both compared methods for low flow rates while is giving worse results for more dynamic flow regimes while the other two gain more accuracy for higher flow rates. It will be interesting to verify performance of the proposed model for a truly large data sample, especially of mixed origins of different experimental installations [23] [24]. On the other hand, it is noticeable that CDPM beats direct estimation based on raw data records by a small difference yet with much lower variance as well.

Total weight of material transported by the pneumatic conveying flow rig shown in row 5 of Tab. 1 showed superior

performance of CDPM model over the others however accuracy on 91.2% is not yet sufficient to treat ECT-based systems as a reliable, stand-alone, online monitoring tool for pneumatic conveying process. Nevertheless as stated in the beginning this measure proved that simplification of modeling process meant by the reducing the parameters number is feasible. Model that assumed mean material concentration value related to mean electric permittivity performed better than geometrical modeling of assumed cross-sectional areas of homogeneous material distribution.

More extensive computational study is required to derive more definitive conclusion about the performance of the CDPM model. Especially, further research work on the larger number of datasets in order to verify the range of applicability of this method to more general classes of applications for ECT monitoring of powder flows in vertical and inclined sections is needed. Next step is to construct a distributed computational environment for big experimental measurement data employing map-reduced paradigm in order to cope and test CDPM performance extensively. It would be interesting to see the AR-based study and track users what and how these professionals perceive the industrial environments to obtain a baseline for further development [15][25][26].

IV. Summary

This work illustrates experimental verification of the proposed CDPM model for temporal modeling of industrial pneumatic conveying process. The method is based on the incorporation of the extra expert knowledge as the regularization factor into the inverse problem solving for electrical capacitance tomography. As the initial results show, the methodology is suitable for temporal modeling of ECT-based monitored pneumatic conveying of bulk solids flow since it helps to identify the flow states (regimes) with similar or better accuracy than the state of the art methods. Calculated total quantity of material transported based on the proposed approach is of approximately 5-10% more accurate than the previous research shown. Hence the proposed simplified model based on mean material concentration value seems to be sufficient and in at least some aspects superior over previously reported geometrical parameterization for temporal bulk solid flow modeling.

References

[1] M. S. Beck and R. A. Williams, "Process tomography: a european innovation and its applications," *Measurement Science and Technology*, vol. 7, no. 3, p. 215, 1996.

[2] W. Q. Yang, A. L. Stott, M. S. Beck, and C. G. Xie, "Development of capacitance tomographic imaging systems for oil pipeline measurements," *Review of Scientific Instruments*, vol. 66, no. 8, pp. 4326–4332, 1995.

[3] W. Fang, "Reconstruction of permittivity profile from boundary capacitance data," *Appl. Math. Comput.*, vol. 177, pp. 178–188, June 2006.

[4] K. Grudzien, A. Romanowski, D. Sankowski, and R. A. Williams, "Gravitational granular flow dynamics study based on tomographic data processing," *Particulate Science and Technology*, vol. 26, no. 1, pp. 67–82, 2007.

[5] D. Wanta, J. Kryszyn, J. Buraczyk, and W. T. Smolik, "Www interface for an electrical capacitance tomography system," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 344–347, May 2018.

[6] M. Soleimani and W. R. B. Lionheart, "Nonlinear image reconstruction for electrical capacitance tomography using experimental data," *Measurement Science and Technology*, vol. 16, no. 10, p. 1987, 2005.

[7] K. Grudzien, "Visualization system for large-scale silo flow monitoring based on ect technique," *IEEE Sensors Journal*, vol. 17, pp. 8242–8250, December 2017.

[8] A. Romanowski, K. Grudzien, R. Aykroyd, and R. Williams, "Advanced statistical analysis as a novel tool to pneumatic conveying monitoring and control strategy development," *Particle & Particle Systems Characterization*, vol. 23, no. 34, pp. 289–296, 2006.

[9] T. Rymarczyk, P. Tchorzewski, P. Adamkiewicz, K. Duda, and J. Szumowski, "Practical implementation of electrical tomography in a distributed system to examine the condition of objects," *IEEE Sensors Journal*, vol. 7, no. 1, pp. 11–16, 2017.

[10] K. Grudzien, A. Romanowski, and R. Williams, "Application of a bayesian approach to the tomographic analysis of hopper flow," *Particle & Particle Systems Characterization*, vol. 22, no. 4, pp. 246–253, 2006.

[11] A. Kowalska, R. Banasiak, R. Wajman, A. Romanowski, and D. Sankowski, "Towards high precision electrical capacitance tomography multilayer sensor structure using 3d modelling and 3d printing method," in *2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE*, pp. 238–243, 2018.

[12] R. Banasiak, R. Wajman, T. Jaworski, P. Fiderek, P. Kapusta, and D. Sankowski, "Two-phase flow regime three-dimensonal visualization using electrical capacitance tomography algorithms and software," *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie rodowiska*, vol. T. 7, nr 1, pp. 11–16, 2017.

[13] M. Panczyk, T. Rymarczyk, and J. Sikora, "Comparison of the inverse problem solutions for a 2d damp wall multilayer and nonhomogenous models," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 81–84, May 2018.

[14] V. Mosorov and D. Sankowski, "Estimation of the rotation angle of gas/solid swirl flow by subpixel image resizing," *Asia-Pacific Journal of Chemical Engineering*, vol. 13, no. 2, p. e2177, 2018.

[15] A. Romanowski, "Big data-driven contextual processing methods for electrical capacitance tomography," *IEEE Transactions on Industrial Informatics*, vol. doi:10.1109/TII.2018.2855200, p. in press, 2018.

[16] A. Schmidt, "Implicit human computer interaction through context," *Personal Technologies*, vol. 4, no. 2, pp. 191–199, 2000.

[17] E. Pascalau, G. Nalepa, and K. Kluza, "Towards a better understanding of context-aware applications," in *FedCSIS13, ACSIS, IEEE*, p. 959962, 2013.

[18] A. Romanowski, K. Grudzien, Z. Chaniecki, and P. Wozniak, "Contextual processing of ECT measurement information towards detection of process emergency states," in *Hybrid Intelligent Systems (HIS), 2013 13th International Conference on*, pp. 291–297, 2013.

[19] I. Jelliti, A. Romanowski, , and K. Grudzien, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *FedCSIS16, ACSIS, vol. 8. IEEE*, p. 16131619, 2016.

[20] C. Chen, P. W. Woźniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, and M. Fjeld, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 52:1–52:25, 2016.

[21] M. Skuza and A. Romanowski, "Sentiment analysis of twitter data within big data distributed environment for stock prediction," in *2015 FedCSIS'15*, pp. 1349–1354, Sept 2015.

[22] H. Garbaa, L. Jackowska-Strumillo, K. Grudzien, and A. Romanowski, "Neural network approach to ect inverse problem solving for estimation of gravitational solids flow," in *2014 Federated Conference on Computer Science and Information Systems*, pp. 19–26, Sept 2014.

[23] R. A. Darwich and L. Babout, "Investigating local orientation methods to segment microstructure with 3d solid texture," *IET Image Processing*, vol. 12, no. 7, pp. 1265–1272, 2018.

[24] S. Waktola, K. Grudzien, L. Babout, and J. Adrien, "Local concentration changes in eccentric and concentric silo discharging modes using x-ray tomography," in *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 377–380, May 2018.

[25] A. Wojciechowski and K. Fornalczyk, "Exponentially smoothed interactive gaze tracking method.," *Springer, Cham, In International Conference on Computer Vision and Graphics (pp. 645-652)*, 2014.

[26] A. Wojciechowski and R. Staniucha, "Mouth features extraction for emotion classification," in *FedCSIS16, ACSIS, vol. 8. IEEE*, p. 16851692, IEEE, 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 287–294
DOI: 10.15439/2018F36
ISSN 2300-5963 ACSIS, Vol. 15

Supporting gastroesophageal reflux disease diagnostics by using wavelet analysis in esophageal pH-metry

Piotr M. Tojza, Grzegorz Redlarski
Gdansk University of Technology
in Gdansk
ul. G.Narutowicza 11/12 80-233 Gdansk, Poland
Email: {poitr.tojza, grzegorz.redlarski}@pg.edu.pl

Maria Janiak
Gdansk Medical University
in Gdansk
ul. Marii Sklodowskiej-Curie 3A, 80-210 Gdansk, Poland
Email: marj@gumed.edu.pl

***Abstract*—This paper presents a new approach to computer supported esophageal pH-metry measurement analysis performed in order to diagnose gastroesophageal reflux disease. In this approach wavelet analysis was used to analyse the esophageal pH-metry course. The research was performed on three groups of pH-metry courses: whole 24-hour pH-metry course, sleep only pH-metry course and 20 minutes after the end of a meal pH-metry course. After performing a 128 level decomposition of the pH-metry course, the W_x was defined as a parameter of extreme differential. This parameter was used to distinguish patients esophageal pH-metry results and on that basis classify patients as healthy or sick. Using this method the sensitivity of 77% was achieved.**

I. Introduction

GASTROESOPHAGEAL reflux disease (GERD) is one of the most commonly diagnosed diseases of the upper gastrointestinal tract, especially among the inhabitants of developed countries [1], [2], [3], [4]. It is estimated that the symptoms occur at least once a month in 44% of American adults, about 20% of Europeans, 6.6% of Japanese and Singaporeans or 3.5% of Koreans. However, among people in Africa and some Asian countries the disease is diagnosed very rarely [2]. The impact on the occurrence and development of the disease is largely influenced by the lifestyle of inhabitants of developed countries including the type of diet, the use of stimulants (alcohol, coffee, smoking cigarettes) or stress. In addition, the symptoms of GERD may increase as a result of misalignment during sleep or during increased physical effort (eg. during exercise in the gym) [5]. Studies suggest that many people are not fully aware of having the disease, though being effected by it's developing symptoms, and reporting to the doctor only when the disease has developed [5].

Among the numerous methods of diagnosing GERD 24hour esophageal pH-metry and 24-hour pH-metry with impedance measurements are the most popular invasive diagnostic techniques. These methods are characterized by a very high percentage of correctly diagnosed patients, but their main drawback is the time needed by a specialist gastroenterologist to evaluate the results of the measurement. Analysis of 24-hour format pH and impedance is a tedious and time-consuming task, that reduces the time the physician can spend on treating other patients or performing studies [6], [7], [8], [9].

Due to the impact of GERD on the condition of the upper gastrointestinal tract, which reflects on the health and lifestyle of patients, special attention should be paid to the process of early GERD diagnosis. This is particularly important in the context of the increase in the amount of positively diagnosed inhabitants of developed countries. In view of the increasing trend in the incidence of patients affected by GERD, taking into account the time-consuming diagnostic test, it is advisable to take steps to automate the assessment process of pH and pH with impedance measurements. Automating the process of pH-impedance courses evaluation will shorten the laborious analysis allowing the physician to quickly assess the results using their knowledge and experience. In order to implement the automation assessment process of pH courses discrete wavelet analysis was used.

II. Upper gastrointestinal tract reflux diseases

The mechanism of regurgitation of the stomach to the esophagus is a physiological process that occurs naturally in the human circadian cycle [2], [10]. Excessive exposure of tissues of the esophagus to the material alleged to reflux - mainly of hydrochloric acid and pepsin is prevented by the antireflux barrier, consisting of four components: the gastro-esophageal connection (lower esophageal sphincter), a mechanism for cleaning the esophagus of hydrochloric acid - the so called acid clearance, the upper esophageal sphincter and the resistance of esophageal mucosa [9]. These mechanisms help to protect esophagus tissue against chemical reactions destroying the tissues, since the esophagus is not adapted, as is the case of the stomach or duodenum, to longer exposure to the harmful effects of gastric acid.

The pathological situation will occur when, for various reasons, the physiological mechanisms protecting the esophagus from the gastric acid fail. Given that the hydrochloric acid and pepsin are the most harmful secretions of the upper gastrointestinal tract, often such a situation leads to occurrence and development of upper gastrointestinal tract reflux

TABLE I
PARAMETERS USED TO CALCULATE THE *Total DeMeester Parameter*

Lp.	Required parameter
1	Number of reflux episodes
2	Number of long reflux episodes (longer than 5 minutes)
3	Time of the longest reflux episode
4	Time during pH < 4 in horizontal position [%]
5	Time during pH < 4 in suspine position [%]
6	Total time during which pH < 4

diseases, mainly gastroesophageal reflux disease - GERD and reflux laryngo-pharyngeal - LPR. If left untreated, pathological changes result in deterioration of the patients quality of life, and in the extreme case lead to tumour lesions that are more complicated to treat and can lead to death [2], [5], [11], [12].

III. GERD DIAGNOSTICS

GERD diagnosis is possible with the use of a number of invasive and non-invasive tests. Apart from 24-hour pH-metry and 24-hour pH-metry impedance other methods are used, such as: gastrointestinal endoscopy, radiography with a double contrast or esophageal manometry [2], [12]. Each of these methods has a number of advantages, however, their limitations result in a lower GERD detection efficiency in comparison with pH-metry with impedance measurements. Therefore esophageal pH-impedance measurement is considered to be the best and the most common invasive GERD diagnostic method. Multichannel intraluminal Impedance-pH metry (MII-pH) is now the "gold standard" in reflux disease diagnostics [2], [5], [12], [13], [14], [15], [16], [17], [18], [19].

Esophageal pH-metry can determine the pH of the contents in the esophagus. It is accepted that the exposure of the esophagus to content which pH is less than 4 is harmful. Analysis of pH courses involves determining and defining certain parameters, characteristic to GERD, based on the method described by *DeMeester* [8]. These characteristic parameters are shown in Table I.

After characteristic parameters are calculated, the Total DeMeester Count can be calculated and compared with a reference value of 14.71 [8]. If the value of the Total DeMeester Count is higher than the reference value the patient is diagnosed as sick. Unfortunately, pH-metry alone has some limitations, as it allows track only pH changes in the esophagus, but it does not allow to determine the physical state of content passing from the stomach into the esophagus. This is especially problematic, since non-acid reflux episodes cannot be detected [12]. An extensions of diagnostic capabilities can be provided by studying esophageal impedance. Esophageal impedance measurement was first described in 1991 [12] and has since gained considerable popularity in the diagnosis of diseases of the upper gastrointestinal tract. This method, however, was not the subject of this research work.

IV. WAVELET ANALYSIS

Wavelet transform is currently one of the most used signal processing technique [20]. It allows for the use of other than the sine basis function (Fourier analysis can decompose a signal into components of a sinusoidal). As a result, it is possible to decompose the analysed signal components based other shapes, which often is highly useful in the identification of the signal's characteristics. In addition, the Fourier transform allows to obtain the data only in the frequency domain, while wavelet transform can give information in both time and frequency domains. Wavelet transform of a signal is calculated according to the formula (1)

$$S_\psi(a,b) = \frac{1}{\sqrt{a}} \int_{inf}^{\inf} s(t)\psi(\frac{t-b}{a})dt \qquad (1)$$

where: a: scale parameter, b: ptranslation parameter, $s(t)$: examined signal, ψ: chosen wavelet, $S_\psi(a,b)$: wavelet coefficient and $\psi(\frac{t-b}{a})$: kernel [20].

The coefficient a in the formula (1) is responsible for the scale representation of the selected wavelets. Values between 0 and 1 cause the wavelet to be shortened, and for a value above 1 is extended. The b in the formula (1) is responsible for moving the wavelet in the time domain (for b greater than 0 the wavelet is moved to the right on the timeline). This transformation should be viewed in the context of the five most popular families of wavelets, containing the ranks of their representatives: orthogonal (*Haar, Daubechies, Symlets*, etc.), the biorthogonal (*BiorSplines, ReversBiors*, etc.), function scaling (eg. *Meyer*) without scaling function (*Morlet, Mexican hat, Gaussian*, etc.) and the type Complex (*Shanon, Gaussian Complex, Complex Morlet*, etc.). The most common representatives of each of these families usually include wavelet type *Daubechies, Taking, Meyr, Morlet, Shannon*. The set of wavelet functions used to transform the signal consists of a basic waveform and the features that are scaled and time-shifted copies of the output signal.

A. pH-metry wavelet analysis

The study aimed to develop new mechanisms to accelerate the evaluation of esophageal pH measurements by the gastroenterologist. Currently used methods of pH-metry analysis are based on strictly defined coefficients *DeMeester*. In this paper an attempt to develop an alternative evaluation method is shown, which is based on wavelet digital signal processing.

To determine the effectiveness of diagnostic method the sensitivity parameter is used. Sensitivity is one of the most commonly used parameters to assess diagnostic tests [21]. The sensitivity of the test means the number of detected ill patients compared to the total number of patients in the study group of patients.

The esophageal pH courses were subjected to wavelet analysis. The aim of the decomposition was to find and determine clear criteria to distinguish and classify different pH courses into two groups: healty and sick. Three different approaches were adopted in relation to the types of analysed pH data:

- 21-hour test results (total registration period) divided into 4-hour intervals,
- courses of 20 minutes measured from the end of meals,

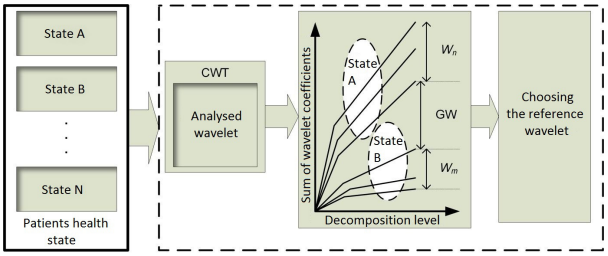


Fig. 1. The idea of wavelet decomposition of pH-metry courses.

- waveforms representing the 6 hour and 25 minute periods of sleep.

Wavelet decomposition was carried out using the five most common representatives of wavelet families *Daubechies, Taking, Meyer, Morlet, Shannon*.

B. Wavelet selection

The method of selecting wavelets for the analysis of non-stationary signals (biomedical signals) was preceded by a comprehensive analysis of the literature [22], [23], [24], [25], [26]. This analysis shows that different states of patients health (registered on pH courses) can be presented in the form of graphs, representing the sum of wavelet coefficients as a function of the level of the signals decomposition. The resulting graph shows the sum of wavelet coefficients in the form of ribbons. It was suspected that different health states will cause those ribbons to be located close to each other in a specific part of the coordinate system. In another approach it was suspected that health states could be distinguished when calculating and comparing the length between value of extreme wavelet sum coefficient, distinguished in the coordinate system. Such parameters were named W_m and W_n. Also a GW coefficient was defined as width of the gap between W_m and W_n regions. This approach is illustrated in Fig. 1, where the defined health states are: A - sick, and B - healthy.

In each case of the carried out analysis a 128 level, continuous decomposition was performed. For each level of decomposition, a series of wavelet coefficients were obtained, which then were summed. As a result a vector of dimension 1 x 128 containing sums of wavelet coefficients was obtained. This approach was used directly in the analysis of 20-minute pH courses after the end of a meal or sleep time. In the case of 21 hours pH courses, the whole course was divided into seven 4-hour episodes and subjected to decomposition separately. Eventually vectors were combined formed in one graph.

In the next step, W_x was defined and named - parameter of extreme differential, calculated using the (2) equation, that is related to maximal differential between extreme wavelet ribbons coefficients values.

$$W_x = L_{max128} - L_{min128} \tag{2}$$

where: L_{max128}: maximal value for the 128 level of decomposition and L_{min128}: minimal maximal value for the 128 level of decomposition.

To determine which of the wavelets from selected wavelet families is the best, an experiment was conducted, using a model pH course of a healthy person. In this experiment a 128 level continuous wavelet decomposition was carried out, followed by an examination - for which of the checked wavelets the difference between the extreme values W_n will be the lowest. The results of the experiment are shown in Table II. The analysis showed that in two cases - for wavelets type *Shannon Haar*, the result of the decomposition virtually precludes their further use in these studies. Very good, as expected, results were obtained for wavelets *db3, Bior* and *Morlet*. However, in the case of wavelet type *Meyer* when plotting the ribbons representing the sum of the wavelet coefficients, it turned out that the value of the designated indicators W_1, W_2 and GW are highly unsatisfactory. This issue is caused by an overlap in some of the waveforms representing health states: A and B (which in practice means the inability to distinguish them). To sum up, taking into account the previously made assumptions about the selection process, it has been shown that the most effective wavelet to decompose pH waveforms is *Morlet* wavelet type.

V. 21-HOUR PH-METRY COURSE ANALYSIS

Conducting research in the relevant field, wavelet decomposition of each registered 21-hour pH course was performed. For this purpose, each 24-hour pH course was divided into seven 4-hour intervals, and then - after the process of their decomposition - the wavelet coefficients were summed and plotted on a single graph, representing a patient exam results.

Patients results were divided into two groups: a test and a validation group. The test group consisted of patients whose Total DeMeester Count was below 50, and therefore diagnosed as healthy and/or mildly sick. The validation group included patients whose Total DeMeester Count was above 50.

In the first place the W_x coefficient was calculated for patients of the test group. The calculated values, as well as the *DeMeester* coefficients are shown in the Table III.

As a result, it was observed that in the case of healthy individuals the extreme differential coefficient W_x is equal to 336 (first case) and 557 (in second) units. The above observation can be justified by analyzing the pH courses of healthy patients, in whose case - because of the a small or marginal amount of reflux episodes - there is no frequent changes of pH during 24 hours. Therefore, the amount of wavelet components in the pH courses remain relatively low, which in turn leads to low sum of wavelet coefficients values.

For sick patients the extreme differential coefficient W_x vary from 789 to 2.327 units. High W_x values are associated with a large dynamics in pH signal changes, resulting from various kinds of components, causing significant variations in pH over a short of period of time. High values of W_x however are not directly proportional to the *DeMeester* count, which can be seen by comparing patients with D and G. The *DeMeester* count for the G patient is lower than that of patient D, while the value of the extreme differential takes a value equal to 2371, which is 250% higher than for the patient D. High *DeMeester*

TABLE II
VALUES OF MAXIMUM WIDTH OF WAVELET COEFFICIENTS W_n

Wavelet decomposition level	32			64			96			128		
Wavelet type	W1	GW	W2	W1	GW	W2	W1	GW	W2	W1	GW	W2
db3	90	191	45	392	400	118	1416	815	337	1416	815	337
bior4.4	74	163	40	325	341	100	787	434	177	1359	523	278
Meyr	-47	-29	-91	-213	-11	-31	-452	32	-68	-629	-27	-91
Morlet	46	93	27	169	235	61	379	343	102	832	283	155
Shannon	solutions ambiguous											
Haar	solutions ambiguous											

TABLE III
VALUES OF W_x PARAMETER FOR PATIENTS FROM THE TEST GROUP: WHOLE 21 HOUR PH COURSES

Patients ID	W_x parameter	Pateints diagnosis	*DeMeester* count
C	789	sick	68,0
D	924	sick	94,0
E	1417	sick	107,8
F	1477	sick	102,0
G	2327	sick	80,4
A	336	healthy	1,0
B	557	healthy	11,7

TABLE IV
VALUES OF W_x PARAMETER FOR PATIENTS FROM THE VALIDATION GROUP: WHOLE 21 HOUR PH COURSES

Patients ID	W_x parameter	Pateints diagnosis	*DeMeester* count
H	382	sick	42,0
I	973	sick	19,5
J	987	sick	43,1
K	624	sick	18,7
L	1608	sick	41,8
M	1434	sick	15,9
N	601	sick	19,4
O	1198	sick	29,5
P	647	sick	21,5

count is due to the presence of components dependent not only on pH value but also on time. Hence, in some patients high *DeMeester* count results from the presence of only one dominant pathological symptom, which determines the high value, while in other patients all or most of the symptoms are present, but manifest less frequently.

In the next stage of research, the extreme differential coefficient W_x was compared between a wider group of patients - the validation group, in order to verify the results of the analysis performed on the test group. The test results are shown in Table IV.

As can be seen in Table IV except the case of patient H, the value of W_x for all patients is above 600. Therefore, on the basis of both the above observations and the results from the test group patients, to distinguish the sick from the healthy patients, the value of W_x coefficient was set to 600 units. This allowed to achieve a 77% sensitivity. It should also be noted that the value of extreme differential of more than 1000 units always points to the case of a sick person.

TABLE V
DATA CONCERNING INDIVIDUAL PATIENTS FROM THE TEST GROUP: PH COURSES DURING SLEEP

Patients ID	Pateints diagnosis	*DeMeester* count
1	12,4	healthy
2	11,3	healthy
3	68,0	sick
4	42,0	sick
5	94,0	sick
6	19,5	sick
7	18,7	sick
8	41,8	sick
9	15,9	sick
10	19,4	sick
11	19,6	sick
12	29,5	sick
13	107,8	sick
14	102,0	sick

A. Wavelet analysis of pH-metry courses during sleep

The next stage of the research was wavelet decomposition referred to the pH courses registered during patients sleep. The purpose of this approach was to check if the plotted sum of wavelet coefficients (as a function of the decomposition level) concerning healthy and sick patients differ from each other in such a way that they can be helpful from a diagnostic point of view. As previously, a continuous 128 lv wavelet decomposition was performed after which the sum of wavelet coefficients was plotted. In contrast to earlier studies - relating to the entire pH course - this analysis was carried out for the entire time during the patients sleep, without dividing it into pieces. Therefore, in this case the W_x coefficient was not calculated, and only a single ribbon curve was plotted for each patient. As previously the patient were divided into two groups: test and validation groups. The results for the test group is shown in Table V.

Analysis of these results, shown in Fig. 2 (for the patients listed in Table V) indicates that distinguishing a clear criterion for patient assessment using wavelet analysis based on pH courses of sleep is not possible. Charts of healthy and sick patients in many cases overlap, and therefore there is no direct way to observe a border of a clear division of the plane between the healthy and the sick. Regarding these observations and the medical *DeMeestera* criteria it can be stated, that a comparison of healthy and sick patients (relating to the sleep phase), is pointless in relation to cases in which the *DeMeester* coefficient does not exceed 25.

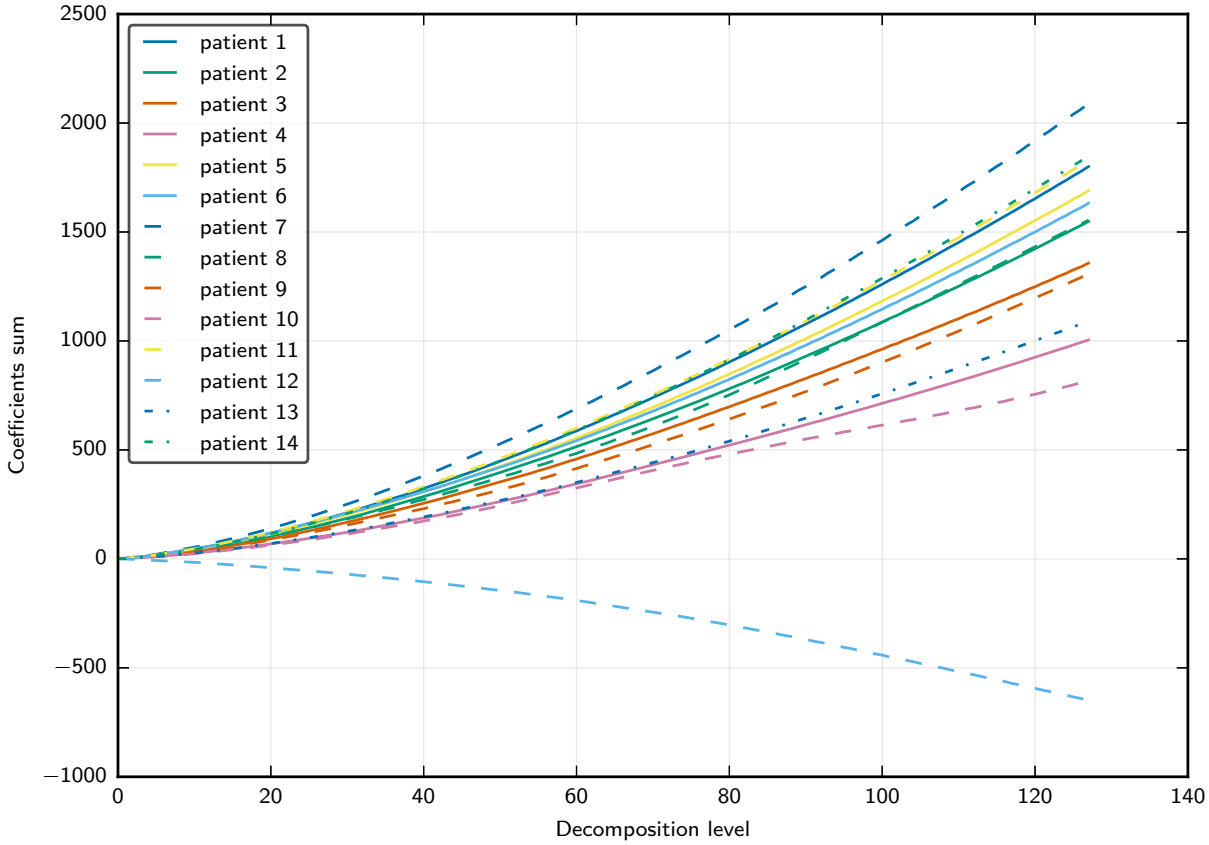


Fig. 2. Results of wavelet analysis of pH-courses during sleep

TABLE VI
VALUES OF W_x PARAMETER FOR PATIENTS FROM THE TEST GROUP: PH COURSES AFTER MEALS

Patients ID	W_x parameter	Pateints diagnosis	*DeMeester* count
A	821	sick	68,0
B	713	sick	94,5
C	1246	sick	107,8
D	887	sick	102,0
E	1303	sick	80,4
F	230	healthy	11,7
G	196	healthy	1,0

B. Wavelet analysis of pH-metry courses after meals

The basis for this approach were 20 minute fragments of pH recordings, that represented the changes in the pH after the patient stopped eating a meal. An example analysis is shown in Fig. 3. During that 20 minutes, the patients couldn't eat another meal. As previously the patients were divided into 2 groups: test and validation groups.

When analysing the results presented in Table VI it can be concluded that in the case of healthy patients extreme differential coefficient W_x achieves the lowest value in the group, whereas in the group of sick patients, the value of the coefficient is greater than 500 units. The conducted experiments did not confirm that the ribbons of wavelet coefficients sums are grouped in certain areas of the coordinate system separately for healthy and sick patients. Therefore the only clear difference between the ribbons describing healthy and sick patients is the W_x coefficient. On this basis - in order to verify the observed phenomenon, that in healthy patients the extreme differential is lower that sick patients - a threshold has been set, to discriminate those two groups of medical conditions. The threshold was set to 250 units. To verify the primary results, the threshold was applied to results obtained from the validation group. The results of these studies, together with *DeMeester* coefficients are shown in Table VII.

The results shown in Table VII lead to state that adopting a threshold of 250 units for W_x coefficient, to distinguish healthy from sick patients, is not fully satisfactory. This is due to the fact that for one of the healthy patient W_x is equal 405, although in can be noted that for this patient the *DeMeester* coefficient is 12.4, so very close to the borderline od 14.7. Moreover it can be seen that the lowest W_x was calculated for a patient whose *DeMeester* coefficient was equal 17. Therefore in the relevant group of 6 sick patients (whose *DeMeester* coefficients were higher than 20) only in 3 cases the W_x coefficient is above 400. For all sick patients, whose *DeMeester* coefficient were higher than 30 the W_x coefficient was significantly higher than 400. On this basis it can be stated that, as with the analysis of 21-hour pH courses, the

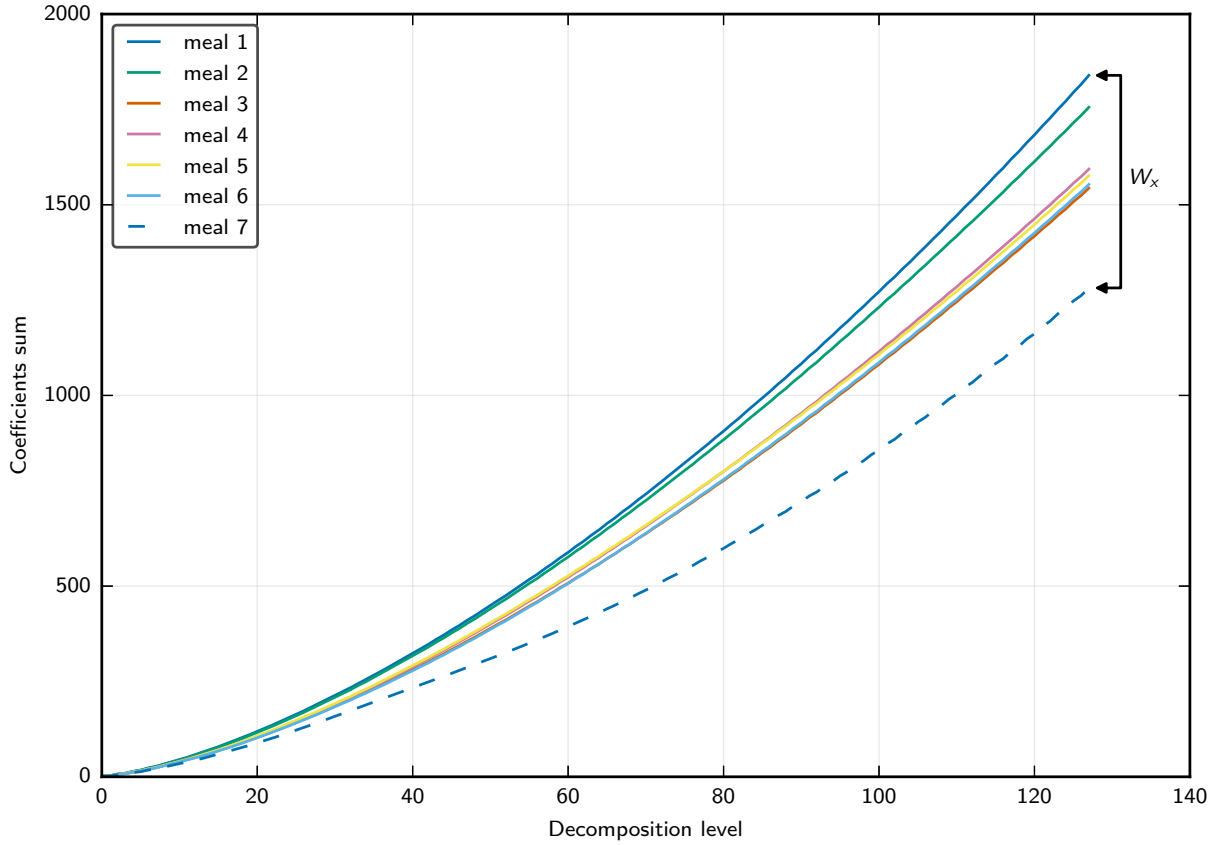


Fig. 3. An example wavelet analysis of 20 minute pH-metry courses taken after meals

TABLE VII
VALUES OF W_x PARAMETER FOR PATIENTS FROM THE VALIDATION GROUP: PH COURSES AFTER MEALS

Patients ID	W_x parameter	Pateints diagnosis	*DeMeester* count
H	405	healthy	12,4
I	821	sick	68,0
J	592	sick	42,0
K	713	sick	94,0
L	486	sick	19,5
M	266	sick	18,7
N	757	sick	41,8
O	543	sick	15,9
P	730	sick	19,4
R	385	sick	19,6
S	1988	sick	29,5
T	603	sick	107,8
U	887	sick	102,0
Q	206	sick	17,0

values of wavelet coefficients depend directly on the nature of changes in the pH course. Thus, both courses that in a short amount of time change their values quickly and courses with a small amount of fluctuation, but constant low pH value may eventually affect the resulting high value of the calculated *DeMeester* coefficient. The sensitivity in this case was 71%.

VI. CONCLUSION

Summing up the results of the research, it can be stated that: A) the method of wavelet analysis of esophageal pH courses registered during the patients sleep can not be used as a tool supporting the diagnosis of reflux diseases, B) both the method of 21-hour pH courses wavelet analysis and the 20-minute pH courses registered after meals wavelet analysis do not give fully satisfactory results, and therefore can not be used as the only method of diagnosing reflux diseases, but they can be used as an additional source of information, support the decision.

The proposed methods can be used as a preliminary assessment procedure when analysing esophageal pH courses, which would be subject to verification by a medical specialist in the course of further analysis. A particular advantage of the presented methods is that its implementation on a computer is quick, and the extreme differential W_x calculation process takes a small amount of time. This in turn leads to a fast diagnosis suggestion that is computed and available for the gastroenterologist. Significant advantages and disadvantages developed methods are shown in Table. VIII.

The conducted experiments showed that wavelet analysis can be successfully applied to evaluate esophageal pH courses in the means for supporting diagnosis of reflux diseases of the gastrointestinal tract. Further research on the methods pro-

TABLE VIII
ADVANTAGES AND DISADVANTAGES OF USING WAVELET DECOMPOSITION TO ANALYSE ESOPHAGEAL PH COURSES

W_x advantage	W_x disadvantage
use only pH courses	sensitivity level around 71% to 77%
easy to implement as a computer program	no possibility to asses the level of the reflux disease advancement
possibility to apply regardless of the pH recording equipment manufacturer	no possibility to check the number or the time of reflux episodes
possibility to fully automate the process of initial diagnostics/classification	possibility to use only as a additional, aiding tool in the diagnostics process

posed should focus primarily on determining accurate extreme differential coefficients W_x to distinguish between the results of healthy and sick patients. Moreover effort should be put on researching the dependence of the thresholds on other, not mentioned here, factors, eg. age, ancestry or general health of patients, to improve the sensitivity of the test. Research showed that improving the sensitivity of the method could be obtained by: increasing the number of patients (collected data), researching other wavelets, researching the corelation between GERD symptoms and pH data for each patient and optimizing the threshold for other specific factors in order to better adjust the threshold value to each patient (like age, sex, general health, etc). Since using wavelet decomposition is a new approach to the topic of GERD diagnostics it is difficult to state which path will lead to better results without commencing more research. Hence, with this state of knowledge, it is difficult to predict the best or optimal course of action. Therefore future work in this topic will present the latest findings in all or in the most promising of the proposed paths.

Further studies should be performed in order to apply classification algorithms to the found wavelet parameters. This can improve the sensitivity of the results as well as allow to find correlations between GERD symptoms and the pH-courses. Such algorithms were successfully applied in other similar research like [27], [28]. Such course can lead in the near future to develop a full medical computers system to aid the diagnostic process of GERD detection, that can be wildly used in the healthcare system. Such system would improve the quality of diagnosis, lower the cost of the diagnostic process and could be wildly used by medical staff. Such systems are being developed in a vast field of medical and healthcare areas: [29], [30].

REFERENCES

[1] G. Redlarski, P. M. Tojza, Computer Supported Analysis of the Human Body Surface Area, International Journal of Innovative Computing, Information and Control, vol. 9, no. 5, 2012

[2] T.Yamada, Podrecznik Gastroenterologii, Czelej, Lublin, 2016

[3] R. Tutuian, M.F. Vela, E. Hill, I. Mainie, A. Agrawal, D. Castell, Characteristics of Symptomatic Reflux Episodes on Acid Suppressive Therapy, Am. J. Gastroenterol., vol. 103, no. 5, 2008, pp. 1090:1096, DOI:DOI:10.1111/j.1572-0241.2008.01791.x

[4] I. Segal, C.S. Pitchumoni, J.Sung, Gastroenterology and hepatology manual: a clinicians guide to a global phenomenon, "McGraw Hill, 2011

[5] T.Yamada, Postepy w Gastroenterologii, Czelej, Lublin, 2006

[6] P.M. Tojza, J. Jaworski, D. Gradolewski, G. Redlarski, Mechatronics, Ideas for Industrial Applications (chapter: Platform Supporting the Esophageal Impedance Analysis), Springer International Publishing, 2015

[7] P.M. Tojza, D. Gradolewski, G. Redlarski, An Application Supporting Gastroesophageal Multichannel Intraluminal Impedance-pH Analysis, SCITEPRESS - Sci. Technol., 2014

[8] G. Redlarski, P.M. Tojza, Computer application supporting upper gastrointensinal tract disease diagnosis based on pH-metry analysis, Pomiary Autom. Kontrola, vol. 59, no. 3, 2013, pp. 193:195

[9] A. Krogulska, K. Wasowska-Krolikowska, Refluks zoladkowo-przelykowy a refluks krtaniowo-gardlowy - znaczenie w laryngologii, Otolaryngologia, vol. 8, no. 2, 2009, pp.42-52

[10] G. Porro, Gastroenterologia i hepatologia, Czelej, Lublin, 2003

[11] T. Yamada, Textbook of Gastroenterology, Blackwell Publishing, 2009

[12] D. Sifrim, F. Fornari, Esophageal impedance-pH monitoring, Dig. Liver Dis, vol. 40, 2008, pp. 161:166

[13] P.J. Kahrilas, Will impedence testing rewrite the book on GERD?, Gastroenterolog, vol. 120, no. 7, 2001, pp. 1862:1864, DOI: 10.1053/gast.2001.25290

[14] A. Lazarescu, D. Sifrim, Ambulatory Monitoring of GERD: Current Technology, Gastroenterol. Clin. North Am., vol. 37, no. 4, 2008, pp. 793:805, DOI:10.1016/j.gtc.2008.09.006

[15] J. M. Pritchett, M. Aslam, J. C. Slaughter, R. M. Ness, C. G. Garrett, M. F. Vaezi, Efficacy of Esophageal Impedance/pH Monitoring in Patients With Refractory Gastroesophageal Reflux Disease, on and off Therapy, Clin. Gastroenterol. Hepatol., vol. 7, no. 7, 2009, pp. 742:748, DOI:10.1016/j.cgh.2009.02.022

[16] S. S. Shay, S. Bomeli, J. E. Richter, Reflux event (RE) clearing: Multichannel intraluminal impedance (MII) compared to pH probe and manometry in fasting severe GERD patients, Gastroenterology, vol. 120, no. 5, 2001, pp. A431, DOI:10.1016/S0016-5085(08)82138-5

[17] D. Sifrim, R. Holloway, J. Silny, Z. Xin, J. Tack, A. Lerut, J. Janssens, Acid, nonacid, and gas reflux in patients with gastroesophageal reflux disease during ambulatory 24-hour pHimpedance recordings, Gastroenterology, vol. 120, no. 7, 2001, pp. 1588:1598

[18] H. L. Smith, G. W. Hollins, I. W. Booth, Epigastric impedance recording for measuring gastric emptying in children: how useful is it?, J. Pediatr. Gastroenterol. Nutr., vol. 17, no. 2, 1993, pp. 201:206

[19] R. Tutuian, D. O. Castell, Use of multichannel intraluminal impedance (MII) in evaluating patients with esophageal diseases. Part III: Combined MII and pH (MII-pH), Pract. Gastroenterol., vol. 27, no. 3, 2003, pp. 19–28

[20] J. T. Bialasiewicz, Falki i aproksymacje, Wydawnictwo Naukowo-Techniczne, Warszawa, 2000

[21] D.M.W. Powers, Evaluation: From Precision, recall and F-measure to ROC, informendess, markedness and correlation, J. Mach. Learn. Technol, vol. 2, no. 1, 2011, pp. 37-63

[22] M. Tokmakci, Analysis of the electrogastrogram using discrete wavelet transform and statistical methods to detect gastric dysrhythmia, J. Med. Syst., vol. 31, no. 4, 2007, pp. 295:302, DOI: 10.1007/s10916-007-9069-9

[23] S. Sharma, G. Kumar, Wavelet analysis based feature extraction for pattern classification from Single channel acquired EMG signal, Elixir Control Engg., vol. 50, no. 8, 2012, pp. 10320:10324

[24] S. Kara, F. Dirgenali, A system to diagnose atherosclerosis via wavelet transforms, principal component analysis and artificial neural networks, Expert Syst. Appl., vol. 32, no. 2, 2007, pp. 632:640, DOI:10.1016/j.eswa.2006.01.043

[25] L. Brechet, M. F. Lucas, C. Doncarli, D. Farina, Compression of biomedical signals with mother wavelet optimization and best-basis wavelet packet selection, IEEE Trans. Biomed.Eng, vol. 54, no. 12, 2007, pp. 2186:2192, DOI:10.1109/TBME.2007.896596

[26] C. Gordan, R. Reiz, ECG signals processing using Wavelets, IEEE, proceedings of the fifth IASTED, vol. 1, 2005,

[27] A. Pasieczna, J. Korczak, Classification Algorithms in Sleep Detection - A Comparative Study, Proceedings of the 2016 Federated Conference

on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 113-120 (2016), DOI: dx.doi.org/10.15439/2016F187

[28] A. Bujnowski, J. Ruminski, M. Kaczmarek, K. Czuszynski, P. Przystup, Cardiovascular data analysis using electronic wearable eyeglasses - preliminary study, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 1409-1412 (2016), DOI: dx.doi.org/10.15439/2016F512

[29] F. Babic, A. Jancus, K. Melisova, Customized Web-based System for Elderly People Using Elements of Artificial Intelligence, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 277-280 (2016), DOI: dx.doi.org/10.15439/2016F165

[30] B. Metelmann, C. Metelmann, Medical Simulation Center as a Model for Testing M-Health Concepts in Prehospital Emergency Medicine, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 1423-1426 (2016), dx.doi.org/10.15439/2016F540.

[31] M. Komenda, M. Karolyi, A. Pokorna, M. Vita, V. Kriz, Automatic Keyword Extraction from Medical and Healthcare Curriculum, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pages 287-290 (2016), DOI: dx.doi.org/10.15439/2016F156

11th Workshop on Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

TOPICS

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on coprocesors (GPU, Intel Xeon Phi, etc.)
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

EVENT CHAIRS

- **Bylina, Beata,** Maria Curie-Sklodowska University, Poland
- **Bylina, Jaroslaw,** Maria Curie-Sklodowska University, Poland
- **Stpiczyński, Przemysław,** Maria Curie-Sklodowska University, Poland

PROGRAM COMMITTEE

- **Amodio, Pierluigi,** Università di Bari, Italy
- **Anastassi, Zacharias,** De Montfort University, United Kingdom
- **Banaś, Krzysztof,** AGH University of Science and Technology, Poland
- **Brugnano, Luigi,** Universita' di Firenze, Italy
- **Fialko, Sergiy,** Tadeusz Kościuszko Cracow University of Technology, Poland
- **Fourneau, Jean-Michel**
- **Gansterer, Wilfried,** University of Vienna, Austria
- **Georgiev, Krassimir,** IICT - BAS, Bulgaria
- **Gravvanis, George,** Democritus University of Thrace, Greece
- **Kozielski, Stanislaw**
- **Księżopolski, Bogdan**
- **Kucaba-Pietal, Anna,** Politechnika Rzeszowska, Poland
- **Lirkov, Ivan,** Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Bulgaria
- **Luszczek, Piotr,** University of Tennessee, United States
- **Marowka, Ami,** Bar-Ilan University, Israel
- **Petcu, Dana,** West University of Timisoara, Romania
- **Ristov, Sashko,** University of Innsbruck, Austria
- **Satco, Bianca-Renata,** Stefan cel Mare University of Suceava, Romania
- **Sergeichuk, Vladimir,** Institute of Mathematics of NAS of Ukraine, Ukraine
- **Shishkina, Olga,** Max Planck Institute for Dynamics and Self-Organization, Germany
- **Srinivasan, Natesan,** Indian Institute of Technology, India
- **Tudruj, Marek,** Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland
- **Tůma, Miroslav,** Academy of Sciences of the Czech Republic, Czech Republic
- **Vazhenin, Alexander,** University of Aizu, Japan

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 297–306
DOI: 10.15439/2018F107
ISSN 2300-5963 ACSIS, Vol. 15

Computation of Gauss-Jacobi Quadrature Nodes and Weights with Arbitrary Precision

Dariusz W. Brzeziński
Institute of Applied Computer Science
Lodz University of Technology
18/22 Stefanowskiego St., 90-924 Łódź, Poland
Email: dbrzezinski@iis.p.lodz.pl

***Abstract*—In the paper there are presented efficient and accurate methods of Gauss-Jacobi nodes and weights computation. They include an enhancement for standard iteration method for Jacobi polynomials zeros finding, weight function formula transformation for increased accuracy of fractional derivatives computation and arbitrary precision application for mitigation of double precision arithmetic flaws. The results of numerical experiments presented in the paper prove high accuracy and efficiency of developed methods for computation of quadratures' nodes and weights, decreased amount of required iterations for polynomials zeros finding and elimination of truncation errors during weights computation. Accuracy of computations depends on height of precision applied for it, which is limited only by accessible hardware.**

I. Introduction

SPECIAL FUNCTIONS are part of mathematics that covers not only well known logarithmic, exponential and trigonometric functions, but also beta, gamma and zeta functions and orthogonal polynomials.

Special functions have numerous applications, not only in mathematics, but also in applied sciences, astronomy, heat conduction, electrical circuits, quantum mechanics and mathematical statistics. More about this subject can be found in [1].

Classical Jacobi orthogonal polynomials are applied in many important scientific areas that include functions' approximation in collocation points method for solutions of ordinary differential equations known as Sturm-Liouville problem and lately - fractional order derivatives and integrals computations by applying Gauss-Jacobi Quadrature [2], [3].

Methods of mathematical formulas implementations in computer programs are crucial part of numerical methods research due to their influence on general accuracy and efficiency of scientific computing. Especially in the case of a basic research as for example computation of polynomials values, their derivatives or their zeros, that can become a part of another computational methods.

Available research on these subjects focus on achieving the highest order of calculated polynomial [4], highest computational speed [5] and lowest computational complexity [6].

Besides of an interesting implementation of algorithms around orthogonal polynomials by applying Julia programming language [4], the majority of results published in scientific papers are obtained by applying computer implementations in Matlab, C++ or Python programming languages and the use of the double precision arithmetic.

The double precision arithmetic is optimized for speed and has many flaws influencing negatively accuracy of computations, e.g. limitations of number values which double precision variables can hold or no programmer influence on mathematical operations rounding.

In the meantime, computational capabilities of computers has been steadily increased and they presently enable using numerous enhancements for the uniform programming languages on the everyday basis. They include *infinite precision computing* for increasing accuracy and correctness of numerical calculations and Nvidia CUDA parallelization technology for their effectiveness, are the best examples in this context.

The term *Infinite Precision Computing* is just a metaphor suggesting, that precision of computation is only limited by an amount of accessible hardware. The more appropriate description of this aspect of computation would be *Arbitrary Precision Computing*.

Arbitrary precision computing has numerous advantages over standard double precision one, e.g. it makes possible for the user to choose a precision for each calculation and for each variable storing a value; it is also not depended on machine or IEEE standard types of data. Therefore, it opens brand new possibilities in terms of accuracy and correctness of scientific computing.

Having that in mind, the primary aim of the following paper is to present high-accuracy computing methods of Jacobi polynomial and its derivative, nodes (zeros of Jacobi polynomials) and weights of Gauss-Jacobi Quadrature.

The secondary aim is to present application usefulness of high-accurate computed Gauss-Jacobi Quadrature for fractional order derivatives and integrals computation.

Presented results of the experiments enable investigating in the future the possibility of their application in spectral methods for solutions of fractional order differential equations and the research of the high-accuracy computation on mitigation of Runge Phenomenon.

The paper is organized as follows: In section II, there are presented mathematical formulas for Jacobi polynomials and their derivatives computation. Section III and IV provides mathematical preliminaries about Gauss-Jacobi Quadrature together with detailed guidance how to adapt it for high-accuracy

fractional derivatives and integrals computation. Section VI describes standard methods for zeros of Jacobi polynomials finding (nodes of quadrature) and their enhancements for increasing computational accuracy and efficiency. Next section VII expands the information on methods of Gauss-Jacobi Quadrature weights computation with details on their enhancements. Both sections include numerous test plots and some preliminary results. Last section VIII presents practical testbed for developed methods in terms of accuracy and efficiency for fractional derivatives of two example functions computation. The paper ends with usual conclusions and future research.

II. Jacobi Polynomial and Its Derivative Computation

Jacobi orthogonal polynomials have two parameters usually denoted as α and β [7] and can be computed by applying Rodrigues' formula [8]

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2n\cdot n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{n+\alpha}(1+x)^{n+\beta}\right]. \tag{1}$$

Jacobi polynomials are orthogonal with respect to the weight function

$$w(x) = (1-x)^{\alpha}(1+x)^{\beta} \tag{2}$$

only for

$$\alpha,\beta > -1,\ -1 < x < 1$$

and particularly

1) For $\alpha = \beta = 0$ we obtain ultraspherical Jacobi polynomials - Legendre polynomials,
2) For $\alpha = \beta = \frac{1}{2}$ we obtain ultraspherical Jacobi polynomials - Chebyshev polynomials of second kind,
3) For $\alpha = \beta = -\frac{1}{2}$ we obtain ultraspherical Jacobi polynomials - Chebyshev polynomials of first kind,
4) For $\alpha = \beta$ we obtain Gegenbauer polynomial.

Jacobi polynomials $P^{(\alpha,\beta)}(x)$ of order n $P_n^{(\alpha,\beta)}(x)$ can be calculated by applying explicit form of the Rodriguez formula [9]

$$P_n^{(\alpha,\beta)}(x) = 2^{-n}\sum_{k=0}^{n}\binom{n+\alpha}{k}\binom{n+\beta}{n-k}(x-1)^{n-k}(x+1)^k, \tag{3}$$

wherein

$$P_0^{(\alpha,\beta)}(x) = 1,\ P_1^{(\alpha,\beta)}(x) = \frac{1}{2}(\alpha+\beta+2)\,x\frac{1}{2}(\alpha-\beta).$$

The derivative of Jacobi polynomial $P_n'^{(\alpha,\beta)}(x)$ of order n $P_n'^{(\alpha,\beta)}(x)$ can be calculated by applying the following formula

$$\frac{d}{dx}\left[P_n^{(\alpha,\beta)}(x)\right] = \frac{1}{2}(n+\alpha+\beta+1)\,P_{n-1}^{(\alpha+1,\beta+1)}. \tag{4}$$

However, application of formula (3) for computations is impractical. We can replace it by three-term recurrent formula (6) for this purpose resulting from theorem 1.1.1 published in [9].

According to it, more practical formula for Jacobi polynomial computation $P_n^{(\alpha,\beta)}(x)$, $n = 0,1,2,3,\ldots$ is

$$P_n^{(\alpha,\beta)}(x) = \sum_{i=0}^{n}\frac{(-1)^{n-i}(1+\beta)_i(1+\alpha+\beta)_{n+i}}{m!\,(n-i)!\,(1+\beta)_i(1+\beta+\alpha)_n}\left(\frac{x+1}{2}\right)^i, \tag{5}$$

where $(\alpha)_i = \alpha(\alpha+1)\cdots(\alpha+i-1)$, $\alpha_0 = 1$.

From (5) the following computational three-term recurrence formula can be then derived

$$\begin{aligned} P_0^{(\alpha,\beta)}(x) &= 1, \\ P_1^{(\alpha,\beta)}(x) &= \frac{1}{2}\left[(\alpha-\beta)+(\alpha+\beta+2)\,x\right], \\ P_{n+1}^{(\alpha,\beta)}(x) &= (\alpha_n x+\beta_n)P_n^{(\alpha,\beta)}(x) - \gamma_n P_{n-1}^{(\alpha,\beta)}(x), \\ & n = 1,2,\cdots, \end{aligned} \tag{6}$$

where

$$\begin{aligned} \alpha_n &= \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+1)(n+\alpha+\beta+1)}, \\ \beta_n &= \frac{(2n+\alpha+\beta+1)(\alpha^2-\beta^2)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}, \\ \gamma_n &= \frac{(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}. \end{aligned}$$

Two example plots for Jacobi polynomial $P_n^{(1.5,-0.5)}(x)$ and its derivative $P_n'^{(1.5,-0.5)}(x)$, $x \in \langle -1,1\rangle$ of order $n = 1,2,\ldots 5$ are presented in Figures 1 and 2.

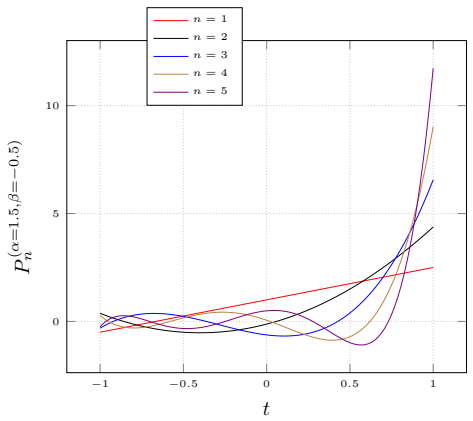


Fig. 1. Graph of Jacobi polynomial $P_n^{(1.5,-0.5)}(x)$, $x \in \langle -1,1\rangle$ of order $n = 1,2,\ldots 5$

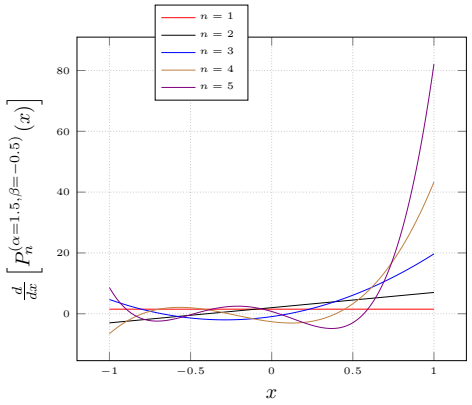


Fig. 2. Graph of Jacobi polynomial 1^{st} derivative $\frac{d}{dx}\left[P_n^{(1.5,-0.5)}(x)\right]$, $x \in \langle -1, 1\rangle$ of order $n = 1, 2, \ldots 5$.

III. Gauss-Jacobi Quadrature

A weight function enabling elimination of problems with integration of functions with singularities at both ends of integration interval is so called *Jacobi weight* (2).

By using some of its properties, there can be increased the accuracy of numerical integration, e.g. for computation of derivatives and integrals of fractional orders. Detailed description will be presented in subsection of section IV.

Jacobi polynomials are orthogonal in respect to weight function (2).

By applying Gauss-Jacobi Quadrature definition, formula for finite integral approximation assumes the following form

$$\int_{-1}^{1} (1-x)^{\alpha} (1+x)^{\beta} \cdot f(x)\, dx \approx \sum_{k=1}^{n} w_k f(x_k), \quad (7)$$

where the nodes of the quadrature x_k are the zeros of Jacobi polynomial $P_n^{(\alpha,\beta)}(x_k)$ of order n.

The weights w_k can be computed by applying the following formula

$$w_k = 2^{\alpha+\beta+1} \frac{\Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!\Gamma(\alpha+\beta+n+1)(1-x_k^2)\left[P_n^{(\alpha,\beta)'} x_k\right]^2}, \quad (8)$$

where $P_n'(x_k)$ is 1^{st} derivative $P_n^{(\alpha,\beta)'}(x_k)$ of Jacobi polynomial of order n and $\Gamma(.)$ is Euler Gamma function.

IV. Formulas for Fractional Order Derivatives and Integrals Approximation

Fractional order derivatives and integrals can be approximated by applying numerous formulas representing various approaches to this problem [10]. The most popular are as follows.

Riemann-Liouville integral of fractional order $\nu > 0$

$${}_0^{RL}\mathrm{I}_t^{(\nu)} f(t) = \frac{1}{\Gamma(\nu)} \int_0^t \frac{f(\tau)}{(t-\tau)^{1-\nu}} d\tau, \quad (9)$$

Riemann-Liouville derivative of fractional order $\nu > 0$

$${}_0^{RL}\mathrm{D}_t^{(\nu)} f(t) = \frac{d^n}{dt^n}\left[\frac{1}{\Gamma(n-\nu)} \int_0^t \frac{f(\tau)}{(t-\tau)^{\nu-n+1}} d\tau\right], \quad (10)$$

Caputo derivative of fractional order $\nu > 0$

$${}_0^{C}\mathrm{D}_t^{(\nu)} f(t) = \frac{1}{\Gamma(n-\nu)} \int_0^t \frac{f^{(n)}(\tau)}{(t-\tau)^{\nu-n+1}} d\tau \quad (11)$$

with the following conditions: $f(t) = 0$ for $t \le 0$, $f(0) = 0, f^{(1)} = f^{(2)} \ldots f^{(n)} = 0$.

The following formula presents the relationship between formula (10) and (11)

$${}_0^{RL}\mathrm{D}_t^{(\nu)} f(t) = {}_0^{C} D_t^{(\nu)} f(t) + \sum_{k=0}^{n-1} \frac{t^{k-\nu}}{\Gamma(k-\nu+1)} f^{(k)}(0). \quad (12)$$

Inserting formula (11) to the right side of equation (12) enables derivation of an equivalent to (10) formula for Riemanna-Liouville'a derivative of fractional order

$${}_0^{RL}\mathrm{D}_t^{(\nu)} f(t) = \sum_{k=0}^{n-1} \frac{t^{k-\nu} f^{(k)}(0)}{\Gamma(k-\nu+1)} + \frac{1}{\Gamma(n-\nu)} \int_0^t \frac{f^{(n)}(\tau)}{(t-\tau)^{\nu-n+1}} d\tau. \quad (13)$$

In formulas (9)-(13) ν is a real number such as $n-1 < \nu < n$, n denotes an integer number $n = \lceil \nu \rceil$.

The practical application advantage of Caputo fractional derivative (11) over Riemann-Liouville fractional derivative (10) is, that the first one enable defining initial conditions in terms of classical, integer order derivatives [11]. Therefore, the Riemann-Liouville derivative definition is used more often in theoretical consideration, in which initial conditions must be defined in terms of fractional order integrals [12].

V. Approximation of Fractional Order Derivatives and Integrals by Applying Gauss-Jacobi Quadrature

Using the weight function (2) and integration formula (7), we can "remove" the kernel of the integrand from the formula (9)

$${}_0^{RL}\mathrm{I}_t^{(\nu)} f(t) = \frac{1}{\Gamma(\nu)} \int_0^t \underbrace{(t-\tau)^{\nu-1}}_{\text{kernel}} f(\tau)\, d\tau,$$

substituting $\lambda = \nu - 1, \beta = 0$, we obtain

$$\int_{-1}^{1} (1-t)^{\lambda} f(t)\, dt \approx \sum_{k=1}^{n} w_k f(t_k) = \sum_{k=1}^{n} \frac{2^{\nu}}{(1-t_k^2)\left[P_n^{(\lambda,0)'} t_k\right]^2} f(t_k), \quad (14)$$

where w_k are the weights (8).

Transforming integration interval $[0,t]$ into $\langle -1,1\rangle$

$$\left(\frac{t-t_0}{2}\right)^{\nu}\int_{-1}^{1}\frac{f(u)}{(1-u)^{\lambda}}du$$

where

$$f(u)=f\left(\left(\frac{t-t_0}{2}\right)u+\left(\frac{t+t_0}{2}\right)\right),$$

we obtain formula (15), which can be applied for computing Riemann-Liouville and Caputo fractional derivatives with high accuracy. A formula for fractional integrals computation can be derived in a similar way.

In the formula (15) the difficult part of the integrand - the kernel - equipped with singularity and high increases of function values, is computed using different, much more accurate method - by applying formula for the weights w_k [2].

$$\begin{aligned}{}_{t_0}^{RL}\mathrm{D}_t^{(\nu)}f(t)&=\sum_{k=0}^{n-1}\frac{(t-t_0)^{k-\nu}f^{(k)}(t_0)}{\Gamma(k-\nu+1)}+\frac{1}{\Gamma(n-\nu)}\left(\frac{t-t_0}{2}\right)^{n-\nu}\int_{-1}^{1}\frac{f^{(n)}(u)}{(1-u)^{\nu-n+1}}du\\&=\sum_{k=0}^{n-1}\frac{(t-t_0)^{k-\nu}f^{(k)}(t_0)}{\Gamma(k-\nu+1)}+\frac{1}{\Gamma(n-\nu)}\left(\frac{t-t_0}{2}\right)^{n-\nu}\underbrace{\sum_{k=1}^{n}\frac{2^{\nu}}{(1-u_k^2)\left[P_n^{(\nu-n+1,0)'}u_k\right]^2}}_{w_k}f^{(n)}(u_k),\ n=\lceil\nu\rceil.\end{aligned}\tag{15}$$

VI. Methods of Finding Zeros of Jacobi Polynomials

Formula (7) suggests that the construction of Gauss-Jacobi Quadrature is limited to finding zeros x_k of Jacobi polynomial $P_n^{(\alpha,\beta)}(x_k)$ of order n and determining its derivative $P_n^{(\alpha,\beta)'}(x_k)$.

Polynomial of order n has n distinct zeros [13]. This rule extends for systems of orthogonal polynomials, including Jacobi polynomials. Proof of this theorem can be found in [14].

Chebyshev polynomials of I, II, III and IV kind are special cases of Jacobi polynomial for α and β with values $-0.5,-0.5,\ 0.5,0.5,-0.5,0.5$ and $0.5,-0.5$ respectively.

Zeros of Chebyshev polynomials called Chebyshev points are given by the following formulas [15]:

$$\begin{aligned}x_k&=\cos\frac{(k-0.5)\pi}{n},\\x_k&=\cos\frac{k\pi}{n+1},\\x_k&=\cos\frac{(k-0.5)\pi}{n+0.5},\\x_k&=\cos\frac{k\pi}{n+0.5},\ k=1,2,\dots n.\end{aligned}\tag{16}$$

Finding zeros of Jacobi polynomial with other values α and β is not easy.

However, a standard method for finding zeros of polynomials is an iteration algorithm called Newton-Raphson algorithm [16].

Iteration algorithms usually require first raw approximation for finding a zero. In case of Jacobi polynomial, it can be for example a zero of Chebyshev polynomial of 1^{st} kind (16).

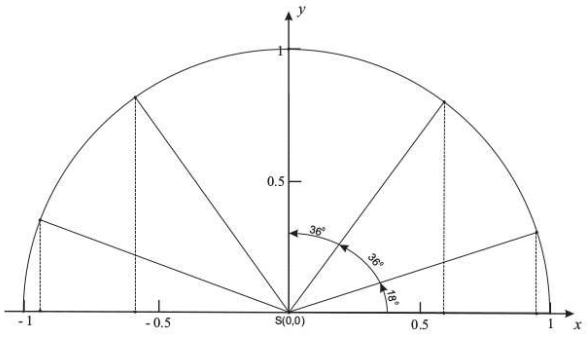


Fig. 3. Chebyshev points,$n=5$ of Chebyshev polynomial of the I kind.

Then, the Newton-Raphson method is used for finding highly accurate location of that zero.

This method is usually fast-convergent, especially for orthogonal polynomials.

Let s denote consecutive iteration. Each iteration of standard Newton-Raphson method requires computation of polynomial value and its derivative. Both values can be used for approximating k^{th} zero in the following way

$$x_k^{s+1}=x_k^s-P_n(s_k^s)/P_n'(x_k^s),\tag{17}$$

where $P_n^{(\alpha,\beta)}(x)$ is Jacobi polynomial computed using recurrent relationship (6).

Its derivative $P_n'^{(\alpha,\beta)}(x)$ can be computed by another recurrence relation

$$P_{n+1}'=P_n+(x-\alpha)P_n'-\beta_nP_{n-1}'.\tag{18}$$

A. Accuracy of the Standard Iteration Method

For the purpose of assessing accuracy of finding zeros of Jacobi polynomial P_n for arbitrary selected values of n by

applying standard iteration method, relative error measure e_r and norm $\|e_r\|_{L_\infty}$

$$\|e_r\|_{L_\infty} = max_i \frac{\|x_i - \hat{x}_i\|}{\|x_i\|}. \tag{19}$$

are used.

The norm (19) assess similarity of two vectors, e.g. with zeros values by applying standard iteration method $\hat{x}_i$ and their exact values. In both cases, there are applied standard double precision for computations and first raw approximations of zeros by applying formulas (16) proposed in [17].

TABLE I
Relative error $\|e_r\|_{L_\infty}$ for selected order n of Jacobi polynomial.

n	Chebyshev I	Chebyshev IV
50	2.89e-15	5.66e-15
100	9.55e-15	1.04e-14
200	1.37e-14	1.10e-14
300	3.18e-14	2.02e-14
400	2.19e-14	4.11e-14
500	2.00e-14	1.19e-14
1000	5.36e-14	7.39e-14

Application of first raw approximation of zeros by applying formulas (16) in Newton-Raphson method enabled finding n distinct zeros with double precision exactness of each zero after 8-10 iterations.

B. Enhancements of the Standard Iteration Method

1) More accurate first raw approximations of zeros: Application of more accurate first raw approximations of zeros can reduce an amount of required iterations for finding an exact location of a zero in Newton-Raphson method.

As it is to see in figure 4 with plots of Jacobi polynomials of order 5 and 6, in the middle part, the polynomials are similar to sine and cosine functions. Near boundary, at $x = 1$ they become compressed. Therefore, we can draw a conclusion that an amount of zeros of Jacobi polynomial increases towards end of the interval $[-1, 1]$. This conclusion suggests that we should use different formulas for first raw approximations for zeros in Newton-Raphson method for middle and boundary parts of Jacobi polynomial.

According to [18] an universal formula for the middle part of the Jacobi polynomials, zeros $4, n-2$ for $k = 1, 2, 3, \ldots, n$ is (20).

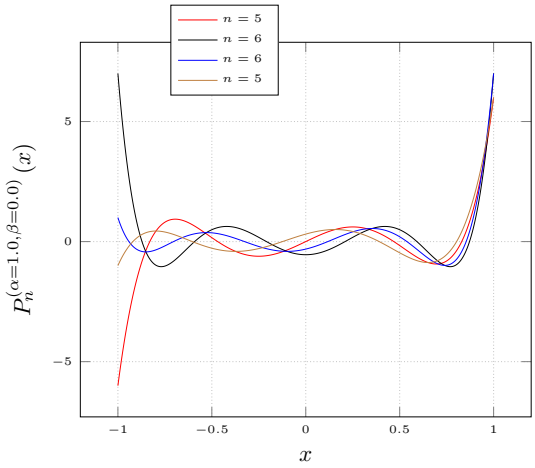


Fig. 4. Jacobi polynomials $n = 5$ and $n = 6$.

$$x_k = -\cos(\theta_k + \delta\theta_k) + O\left(n^{-4}\right) \tag{20}$$

where

$$\theta_k = \pi(2k + \beta - 0.5)/\sigma,$$
$$\delta\theta_k = \left[\left(0.25 - \beta^2\right)\cot\left(\frac{\theta_k}{2}\right) - \left(0.4 - \alpha^2\right)\tan\left(\frac{\theta_k}{2}\right)\right]/\sigma^2,$$
$$\sigma = 2n + \alpha + \beta + 1.$$

For approximating zeros for nodes $[1, 2, n-1, n]$ i.e. for both ends of the interval we can use formula by [18] that uses zeros of Bessel functions of order 5 $J^5_{\alpha,k}$, e.g. [19]

$$x_k = \cos(\theta_k + \delta\theta_k) + O\left(J^5_{\alpha,k} n^{-7}\right) \tag{21}$$

where

$$\theta_k = \frac{J^5_{\alpha,k}}{\nu},$$
$$\delta\theta_k = -\theta_k\left[\frac{4 - \alpha^2 - 15\beta^2}{720\nu^4}\left(\frac{J^2_{\alpha,k}}{2}\right) + \alpha^2 - 1\right],$$
$$\nu = 0.5\sqrt{\sigma^2 + (1 - \alpha^2 - 3\beta^2)/3}.$$

2) Application of reflection formula: For finding zeros of Jacobi polynomial for $\alpha \neq \beta$, there can be applied reflection formula

$$P_n^{(\alpha,\beta)}(-x) = (-1)^n P_n^{(\beta,\alpha)}(x),$$
$$\frac{d}{dx}\left[P_n^{(\alpha,\beta)}(-x)\right] = (-1)^{n-1}\frac{d}{dx}\left[P_n^{(\beta,\alpha)}(x)\right], \tag{22}$$

which results from formula for Jacobi polynomial of order n (1).

It enables reducing computational effort of polynomial values to the right part of the interval, i.e. $x \in [0, 1]$. It means that we only require to compute zeros from the right part of the interval and copy them to the second one $x \in [-1, 0]$.

3) Application of arbitrary precision: enables increasing overall accuracy of computations.

To be able to solve a difficult numerical problem according to a set goal, we have to make some crucial decisions regarding applied hardware, programming tools and techniques for that purpose. This includes a selection of an appropriate computer programming language, mathematical libraries and hardware.

The selection of uniform C++ equipped with the standard mathematical library as a main programming tool is not enough nowadays to take full advantage of available hardware. And it is the main task for a computer scientist, because the newest hardware gives the opportunity to solve many problems, which appeared "unsolvable" not long time ago. The application of *infinite precision computing* for increasing the accuracy and the correctness of numerical calculations and Nvidia CUDA parallelization technology for their effectiveness, are the best examples in this context.

In this paper there is presented application of arbitrary precision for increasing accuracy of computations.

The standard double precision computer arithmetic was replaced by arbitrary precision for most parts. This move made possible unlocking full potential of developed algorithms by using available hardware.

Double precision arithmetic commonly applied in scientific numerical calculations is optimized for speed and has many flaws which influence negatively the accuracy of computations, e.g. limitations of number values which double precision variables can hold or no programmer influence on mathematical operations rounding.

However, it is the lack of clarity in handling of intermediate results which troubles the most, i.e. the floating-point standard only defines that the results must be rounded correctly to the destination's precision and not defines the precision of destination variable. This choice is commonly made by a system or a programming language. The user can not influence it in any way. Therefore, the same program returns significantly different results depending on the implementation of the IEEE standard.

Arbitrary precision makes possible for the user to choose a precision for calculation and for each variable storing a value and it is nor machine or IEEE standard types depended. It is only limited by accessible hardware.

Arbitrary precision can be applied for calculating important constants like π or increase general accuracy of the mathematical computations. Its application purpose is above all to increase accuracy of numerical calculations, e.g. by eliminating under- and overflows, increasing accuracy of a polynomial zeros finding and derivatives and integrals calculating.

Still, application of arbitrary precision has drawbacks:

Arbitrary precision is simulated and therefore, depending on chosen precision, calculations with the help of it require more time to complete than by applying standard data types optimized to run on standard processors - even with the use of FPGAs (field programmable gate arrays), which can be fully programmed by the user.

Another challenge is a requirement of special computational algorithms which can handle different data structures.

Nevertheless, arbitrary precision application already became a part of standard computations without the consent of the user. The process named *constant folding with arbitrary precision* is used in preprocessing phase to increase the accuracy of constants before they can be handled with standard precision data types. This procedure [20] involves replacing constant expressions with their final value in order to reduce the need of recomputing the same result every time the program executes the code line containing the constant. When the compiler flag -01 is inserted GCC complier uses the GNU MPFR library with version 4.3 to handle constant folding and evaluate mathematical applied to constants at compile time at arbitrary precision.

The GNU MPFR library is an arbitrary precision package for C/C++ [21] and is based on the GNU Multiple-Precision Library (GMP) [22]. MPFR supports arbitrary precision floating-point variables and provides exact rounding of all implemented mathematical functions [23]. The code is portable, i.e. it will produce the same result independently from the hardware.

The GNU MPFR library is written in C and thus it can not use operator overloading. Even the most basic arithmetic operations have to be conducted using function calls. Therefore MPFR includes multiple functions for each operation and for each supported data type.

C. Results

Table II presents accuracy of computed zeros of Jacobi polynomial in form of relative error (19) calculated in respect to the exact values obtained by applying Czebyshev points (16) for 50, 100, 500 and 1000 digits precision.

It is worth noting, that the proposed enhancements of the standard iteration method enables finding zeros with arbitrary precision. The level of exactness depends on how high precision is selected applied for computations.

Additionally, application of more accurate first raw approximations of zeros (20) and (21) before Newton-Rapshon method starts, decreases an amount of required iterations until exact zero position is found.

Computation time complexity of running program is presented in figure 5. The time depends directly on the hight of precision selected for computations: for polynomial order $n < 500$ and up to 100 digits precision, the time is similar to double precision computations, for $n > 500$ and more than 100 digits precision, the complexity is 2^n, and $n!$ is for 1000 and more digits precision.

VII. Methods of Jacobi Weights Computation

A. Standard Approach

A standard approach to the problem of Jacobi weights computation is the direct use of formula (8).

In this formula proposed in [24], weight is computed by using value of derivative of Jacobi polynomial of order n and a value of Jacobi polynomial of order $n-1$.

TABLE II
RELATIVE ERROR $\|e_r\|_{L_\infty}$ OF FOUND ZEROS OF $P_n(x)$ FOR SELECTED n IN RESPECT TO (16).

	Czebyszew I			Czebyszew II			Czebyszew III			Czebyszew IV		
n	50	time(s)	iter	100	time(s)	iter	500	time(s)	iter	1000	time(s)	iter
50	1.50e-49	0.047	6	3.77e-99	0.053	7	6.24e-499	0.102	7	8.31e-647	0.150	6
100	6.84e-49	0.0168	6	6.74e-99	0.195	7	1.40e-498	0.392	7	5.24e-641	0.562	7
200	9.52e-49	0.588	7	7.43e-99	0.765	6	1.35e-498	1.323	7	1.89e-641	2.073	7
300	1.66e-48	1.236	6	2.22e-99	1.673	6	1.31e-498	3.012	7	3.64e-640	4.514	7
500	1.49e-48	3.089	6	1.59e-99	4.349	6	9.05e-498	8.034	7	1.58e-648	12.598	7
1000	5.61e-48	11.867	6	2.60e-98	15.913	6	7.96e-498	34.37	7	8.99e-709	49.015	7
2000	7.00e-48	47.353	6	5.45e-98	63.445	6	3.78e-497	126.664	7	8.99e-709	49.173	7

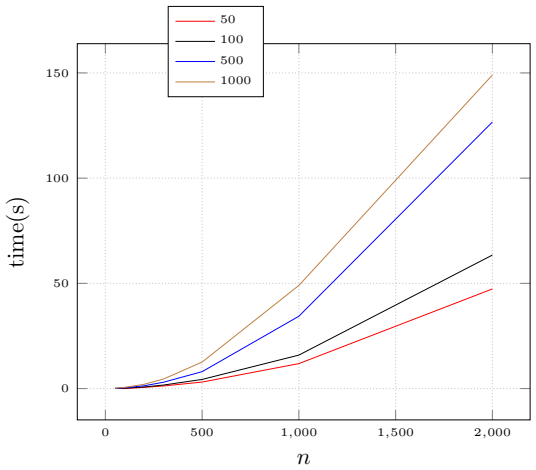


Fig. 5. Computation time complexity of finding zeros of Jacobi polynomials of selected order n with 50, 100, 500 and 100 digits precision.

However, due to $1-x_k^2$ expression occurrence in the denominator of (8), high truncation error is expected if standard double precision is applied.

Additionally, deducing from 4, an amount of zeros in Jacobi polynomials increases quadratically with increasing n (hence the distance between then decreases) towards bounds of integration interval $[-1,1]$. It causes the following problem: if standard double precision is applied, for enough large n, zeros become indistinguishable.

B. Enhancement of Standard Approach

Instead of formula (8), an equivalent formula is suggested to apply

$$w_k = 2^{\alpha+\beta+1}\frac{\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)}{n!\Gamma(\alpha+\beta+n+1)}\frac{1}{\left[\frac{d}{d\theta}P_n(\cos\theta_k)\right]^2}, \tag{23}$$

in which $\frac{d}{d\theta}P^{(\alpha,\beta)}$ is derivative of Jacobi polynomial of order n and $\theta_k = \cos^{-1}x_k$, x_k are zeros of Jacobi polynomial of order n.

The conversion into trigonometric functions space has been proposed by [25]. It enables omitting the expression $1-x_k^2$ and hence reduce truncation error at the same time.

C. Results

Table III presents accuracy of Jacobi weights computation in form of relative error (19) calculated in respect to the exact values obtained by applying Czebyshev weights [15] for 50, 100, 500 and 1000 digits precision. Application arbitrary precision enables computing Jacobi weights with high-accuracy. However, computational accuracy is not so straightforward depended on an amount of digits of precision applied for computations. It is caused by the fact that computational complexity is increased by zeros of finding of the polynomial of a given order.

General time complexity of Jacobi weights w_k computation presented in figure 6 depends directly on precision applied for computations: for $n<500$ and up to 100 digits precision it is similar to double precision, $n>500$ and more than 100 digits precision it is 2^n, and $n!$ for more than 1000 digits precision.

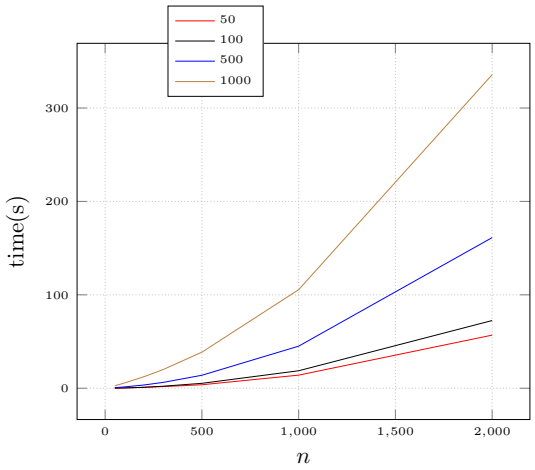


Fig. 6. Computation time complexity of Jacobi weights for selected order n with 50, 100, 500 and 100 digits precision.

VIII. FRACTIONAL ORDER DERIVATIVES AND INTEGRALS COMPUTATION

The most useful method of presenting practical capabilities of algorithms proposed in the following paper is computation of values of integrals and derivatives of fractional order of two exponential functions by applying formulas (9) and (13) with the help of Gauss-Jacobi Quadrature (15).

TABLE III
RELATIVE ERROR $\|e_r\|_{L_\infty}$ OF COMPUTED WEIGHTS w_k FOR SELECTED n IN RESPECT TO (16).

	Czebyszew I			Czebyszew II			Czebyszew III			Czebyszew IV		
n	50	czas(s)	iter	100	czas(s)	iter	500	czas(s)	iter	1000	czas(s)	iter
50	1.46e-46	0.060	8	1.63e-97	0.109	9	9.33e-326	0.640	10	1.14e-323	2.651	10
100	3.73e-46	0.256	8	2.49e-96	0.339	9	9.64e-317	1.440	10	4.51e-319	5.632	11
200	8.39e-45	0.910	8	1.31e-95	1.051	9	1.86e-318	3.469	10	1.97e-318	12.322	11
300	6.53e-45	1.781	8	1.23e-95	2.170	9	2.57e-317	6.252	10	3.40e-317	20.000	11
500	1.01e-45	3.728	8	3.91e-94	5.265	9	1.56e-317	13.895	11	2.06e-320	38.619	11
1000	8.47e-43	14.03	8	8.53e-94	18.710	9	2.72e-326	45.04	11	1.71e-349	105.491	11
2000	7.26e-42	56.852	8	1.14e-93	72.479	9	3.08e-318	161.266	11	4.74e-318	335.694	11

To assess accuracy of computations of fractional derivatives and integrals, it is required to computed exact values with high-accuracy for relative error computation. In case of fractional order derivative and integral computations, the effective accuracy assessment is difficult, sometimes not possible due to general lack of formulas for exact values.

Despite the availability of a handful of analytical formulas for fractional order $\nu = \frac{1}{2}$ and some computational formulas, they are accessible for selected types of functions only. Some other formulas are in form of series expansion only. As it is in case of exponential functions.

For the error (19) computation Mittag-Leffler function is used.

A. Mittag-Leffler Function Computation

The Mittag-Leffler function [26] is a direct generalization of the exponential function e^{at} and it plays a major role in fractional calculus. The one, two and three-parameter representations of the Mittag-Leffler function can be defined in terms of a power series as

$$E_\alpha(at) := \sum_{k=0}^{\infty} \frac{at^k}{\Gamma(\alpha k + 1)},\ \alpha > 0, \tag{24}$$

$$E_{\alpha,\beta}(at) := \sum_{k=0}^{\infty} \frac{at^k}{\Gamma(\alpha k + \beta)},\ \alpha, \beta > 0. \tag{25}$$

When $\beta = 1$, $E_{\alpha,1}(at) = E_\alpha(at)$.

$$E^\gamma_{\alpha,\beta}(at) := \sum_{k=0}^{\infty} \frac{(\gamma)_k}{\Gamma(\alpha k + \beta)} \frac{at^k}{k!},\ \alpha, \beta > 0, \tag{26}$$

in which $(\gamma)_k$ is Pochhammer symbol [27]

$$(\gamma)_k := \frac{\Gamma(\gamma + k)}{\Gamma(k)}.$$

When $\gamma = 1$, $E^1_{\alpha,\beta}(at) = E^\gamma_{\alpha,\beta}(at)$, and when $\gamma = \beta = 1$, $E^1_{\alpha,1}(at) = E_\alpha(at)$. Some particular cases of the Mittag-Leffler function are: $E_0(at) = \frac{1}{1-at}$, $E_1(at) = e^{at}$ $E_2(at) = cosh\sqrt{at}$, $E_{1,2}(at) = \frac{e^{at}-1}{t}$, $E_{2,2}(at) = \frac{sinh\left(at^{1/2}\right)}{at^{\frac{1}{2}}}$, $E_{\frac{1}{2},2}(at) = e^{at^2}\, erfc(-at)$. Papers [28] and [29] present comprehensive knowledge of computing the Mittag-Leffler function and its first derivative.

To calculate the Mittag-Leffler fractional order derivative/integral we combine the Riemann-Liouville fractional derivative of the power function $(t - t_0)^p$, $p \in \mathbf{R}$ and the Mittag-Leffler function (24) or (25)

$${}_{t_0}D_t^{(\nu)} E_{\alpha,\beta}(at) = t^{-\nu} \sum_{k=0}^{\infty} \frac{\Gamma(k+1)\, at^k}{\Gamma(k+1-\nu)\,\Gamma(\alpha k + \beta)} \tag{27}$$

and for calculations of fractional order integral of the Mittag-Leffler function, we apply the following formula

$${}_{t_0}D_t^{(-\nu)} E_{\alpha,\beta}(at) = at^\nu \sum_{k=0}^{\infty} \frac{(at^\nu)^k}{\Gamma(\alpha k + \beta + \nu)}. \tag{28}$$

B. Computing Environment Configuration

All computations described in the paper were conducted using PC computer with Intel i7 2600K Processor, 8 GB of RAM armed with full open-source operating system and compiler: Ubuntu 16.04 LTS 64-bit Linux OS and gcc 5.4.0 compiler.

The computer system can be described as high-performance, because its computational power is enormous. However, it dates from 2011. Therefore it also can be described as commonly used.

To complete the picture that is important for time complexity assessment, calculations were also conducted on an older notebook with Intel Core 2 Duo Processor 2.4 GHz with 4 GB of memory from 2008 with exactly the same software configuration.

C. Results

Figures 7 and 9 present plots of fractional integral and derivative of two exponential functions. Figures 8 and 10 present plots of relative error (19) for $n = 8, 16, 32$ computed with 100 digits precision.

As it is to see, steady 100 digits accuracy can be obtained by applying Jacobi polynomial of order $n = 32$ for computations.

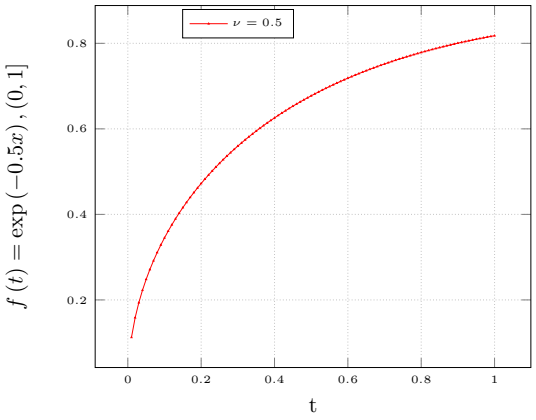


Fig. 7. Plot of fractional integral of order $\nu = 0.5$ of function $f(t) = \exp^{-0.5x}$ in interval $(0, 1]$.

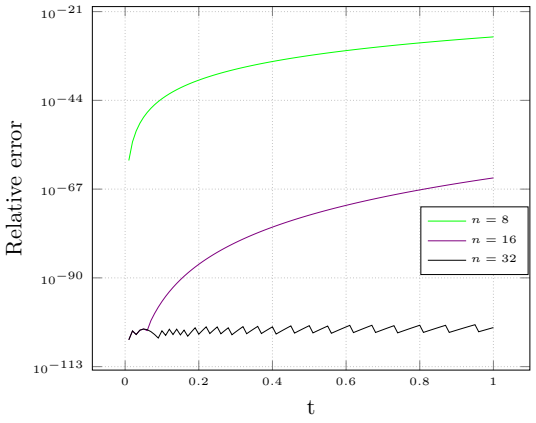


Fig. 8. Plot of relative error of fractional integral of order $\nu = 0.5$ of function $f(t) = \exp^{-0.5x}$ in interval $(0, 1]$.

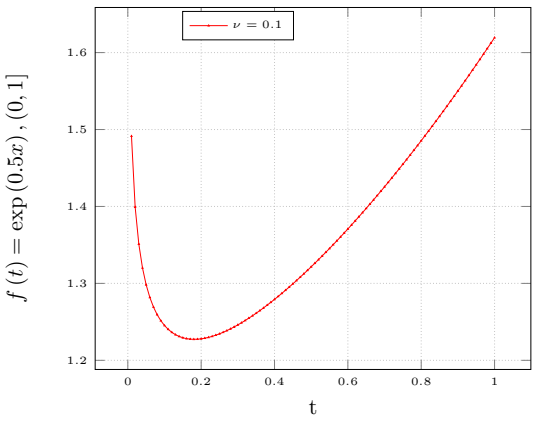


Fig. 9. Plot of fractional integral of order $\nu = 0.5$ of function $f(t) = \exp^{0.5x}$ in interval $(0, 1]$.

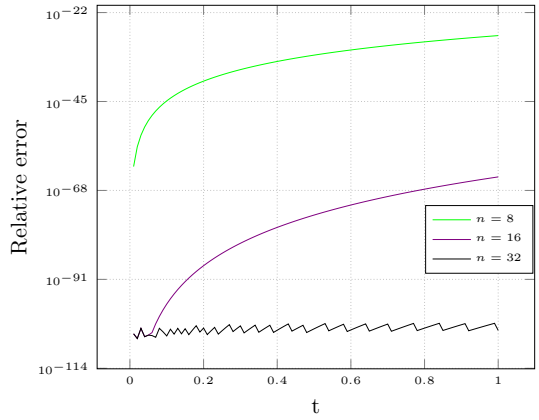


Fig. 10. Plot of relative error of fractional integral of order $\nu = 0.5$ of function $f(t) = \exp^{0.5x}$ in interval $(0, 1]$.

IX. Conclusions

The aim of the following research was to develop the most efficient and accurate numerical algorithms for Gauss-Jacobi Quadrature nodes and weights computation. In the paper, we discuss efficient mathematical formulas for nodes and wights computations and accurate methods of their computer implementations.

Results of numerical experiments presented in the paper prove that application of more accurate raw first approximations of zeros in the standard iteration method of polynomial zeros finding leads to significant decrease of an mount of iterations required for finding high-accurate zero location.

The proposed enhancements for the standard iteration method of determining zeros of Jacobi polynomial enables decreasing an amount of iterations required for finding each zero and increasing their accuracy many hundred times; changes to the weight function formula and its computation by applying cosine function enables massive reduction of truncation errors and increasing overall accuracy of computations.

The enhanced methods programmed by applying excellent arbitrary precision libraries GNU GMP and GNU MPFR together with C++ programming language enable computation of Gauss-Jacobi Quadratures and nodes and wights with arbitrary precision, i.e. with precision limited only by accessible hardware (computer memory).

Results of computations of fractional order derivatives and integrals of example exponential functions prove that modified Gauss-Jacobi Quadrature that uses high-accurately computed nodes and weight enables their computation with steady 100-digits precision with only 32 sampling points at most. It is worth nothing that standard numerical integration methods', e.g. Newton-Cotes quadratures' accuracy is limited to a few digits at best for the same computations [30], [31].

High accurately computed nodes of Jacobi polynomial are an excellent starting point for research on mitigation of Runge phenomenon. High-accurate methods of fractional order derivatives and integrals computation can be useful for constructing more efficient and accurate spectral methods

for solutions of fractional differential equations. This in turn enables more accurate simulations of physical processes and systems.

ACKNOWLEDGEMENT

The work was created as a result of the research project no. DEC-2016/23/D/ST6/01709 financed from the funds of the National Science Center, Poland.

REFERENCES

[1] S. Wolfram. (2005) The history and future of special functions. http://www.stephenwolfram.com/publications/history-future-special-functions/.

[2] D. W. Brzeziński and P. Ostalczyk, "High-accuracy numerical integration methods for fractional order derivatives and integrals computations," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 62, no. 4, pp. 723–733, 2014.

[3] D. W. Brzeziński, "Comparison of fractional order derivatives computational accuracy - right hand vs left hand definition," *Applied Mathematics and Nonlinear Sciences*, vol. 2, no. 1, pp. 237–248, 2017.

[4] A. Townsend, S. Olver *et al.* (2018) Fastgaussquadrature.jl. https://github.com/ajt60gaibb/FastGaussQuadrature.jl#fastgaussquadraturejl/.

[5] A. Glaser, X. Liu, and V. Rokhlin, "A fast algorithm for the calculation of the roots of special functions," *J. Sci. Comput.*, vol. 29, pp. 1420–1438, 2007.

[6] N. Hale and A. Townsend, "Fast and accurate computation of gauss-legendre and gauss-jacobi quadrature nodes and weights," Oxford Centre for Collaborative Applied Mathematics, 2012, oCCAM Preprint Number 12/79.

[7] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions. Applied Mathematics Series*. Cambridge University Press, 1968.

[8] G. Szegö, *Ortogonal Polynomials*. American Mathematical Society, Colloquiam Publications, Volume 23, 1939.

[9] D. Funaro, *Polynomial Approximation of Differential Equations*. Springer-Verlag., 1992.

[10] M. D. Ortigueira, J. A. T. Machado, and J. S. da Costa, "Which differintegration?" *IEE Proceedings - Vision, Image and Signal Processing*, vol. 152, no. 6, 2005.

[11] Y. Povstenko, *Linear Fractional Diffusion-Wave Equation for Scientists and Engineers*. Cham, Heidelberg, New York, Dodrecht, London: Birkhauser, Springer, 2015.

[12] J. Jiang, D. Cao, and H. Chen, "Boundary value problems for fractional differential equation with causal operators," *Applied Mathematics and Nonlinear Sciences*, vol. 1, no. 1, pp. 11–22, 2016.

[13] D. Xin, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press Press, 2000.

[14] E. D. Rainville, *Special Functions*. Chelsea Publications Company, 1960.

[15] J. C. Mason and D. C. Handcomb, *Chebyshev Polynomials*. Champan & Hall/CRC New York, 2003.

[16] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing, Third Edition*. Cambridge University Press, 2008.

[17] K. Petras, "On the computation of the gauss-legendre quadrature form with a given precision," *Jour. Comp. Appl. Math.*, vol. 112, pp. 253–267, 1999.

[18] W. Gautschi and C. Giordano, "Luigi gatteschi's work on asymptotics of special functions and their zeros," *Numerical Algorithms*, vol. 49, pp. 11–31, 2008.

[19] H. Gerber, "First hundert zeros of $j_0(x)$ accurate to 19 significant figures." *Math. Comp.*, vol. 23, pp. 319–322, 1969.

[20] N. Brisebarre and J. M. Müller, "Correctly rounded multiplication by arbitrary precision constants," *IEEE Transactions on Computers*, vol. 57, no. 2, pp. 165–174, 2008.

[21] J. M. Müller, N. Brisebarre, F. D. Dinechin, C. P. Jeannerod, V. Lefevre, G. Melquiond, N. Revol, D. Stehle, and S. Torres, *Handbook of Floating-Point Arithmetic*. New York, NY: Birkhauser, 2010.

[22] T. Granlund *et al.*, *gmp: GMP is a free library for arbitrary precision arithmetic (version 6.0.0a)*, 2015, `https://gmplib.org/`.

[23] N. Brisebarre and J. M. Müller, "Correct rounding of algebraic functions," *Theoretical Informatics and Applications*, vol. 47, pp. 71–83, 2007.

[24] V. I. Krylov, *Priblizhennoe wychislenie integralov, 2e izd.* Mockba: Nauka, 1967.

[25] P. N. Schwarztrauber, "On computing the points and weights for gauss-legendre quadrature," *SIAM Jour. Sci.Comput*, vol. 24, pp. 945–954, 2002.

[26] P. Humbert and R. P. Agarwal, "Sur la fonction de mittag-leffler et quelques-unes de ses géneéralisations," *Bull. Sci. Math. Ser. II*, vol. 77, pp. 180–185, 1953.

[27] R. K. Saxenna, A. M. Mathai, and H. J. Haubold, "On generalized fractional kinetic equations," *Physica A: Statistical Mechanics and its Applications*, vol. 344, pp. 657–664, 2004.

[28] R. Gorenflo, J. Loutchko, and Y. Luchko, "Computation of the mittag-leffler function and its derivative," *Fractional Calculus & Applied Analysis*, vol. 4, pp. 491–518, 2002.

[29] R. Garrappa, "Numerical evaluation of two and three parameter mittag-leffler functions," *SIAM J. Numer. Anal.*, vol. 53, no. 3, pp. 1350–1369, 2015.

[30] D. W. Brzeziński, "Accuracy problems of numerical calculation of fractional order derivatives and integrals applying the riemann-liouville/caputo formulas," *Applied Mathematics and Nonlinear Sciences*, vol. 1, no. 1, pp. 23–43, 2016.

[31] D. W. Brzeziński and P.Ostalczyk, "About accuracy increase of fractional order derivative and integral computations by applying the grünwald-letnikov formula," *Communications in Nonlinear Science and Numerical Simulation*, vol. 40, pp. 151–162, 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 307–310
DOI: 10.15439/2018F271
ISSN 2300-5963 ACSIS, Vol. 15

An effective sparse storage scheme for GPU-enabled uniformization method

Beata Bylina, Jarosław Bylina, Marek Karwacki
Marie Curie-Skłodowska University, Institute of Mathematics,
Pl. M. Curie-Skłodowskiej 5, 20-031 Lublin, Poland
Email: {beata.bylina,jaroslaw.bylina}@umcs.pl

***Abstract*—The authors developed a GPU approach to the uniformization method for the computing transient solution of Markov models. The authors use two techniques to reduce the memory size of storing matrices. One of them is a modification of a storage sparse matrix format HYB; second is to utilize two GPU cards and the multicore CPU. The modified HYB format is suitable for sparse Markovian transition rate matrices and oversized matrices on single GPU, also improving computation performance at the same time. The use of two GPUs enables processing matrices of even bigger sizes.**

I. Introduction

A POWERFUL tool used widely for modeling a lot of processes and systems (natural and artificial ones) are Markov chains.

In [2], we provided details of a heterogeneous (CPU-GPU) implementation of the uniformization method [6], [11], [12] for solving Markov chains. Markov chains transition rate matrices (which are very sparse) were stored with the use of the HYB format which is a hybrid of other well-known sparse formats (ELL and COO). The HYB format was chosen because it gave the best results in experiments described in [1].

However, matrices in question are usually very large and do not fit into one GPU memory. Thus, in [3], we presented an implementation of uniformization for many GPU. Notwithstanding, the communication between GPUs was slow and the results were not satisfactory. Hence, in [4] we described an effective storage scheme for Markov chains transition rate matrices, namely HYBIV, which reduced the memory requirements and the number of miss caches and thereby improved the overall performance. That format was not studied for uniformization though.

This work shows numerical experiments where that format (HYBIV) is used for uniformization and tested for four groups of transition rate matrices (from PRISM). It also compares the HYB format with HYBIV (on 1 and 2 GPUs) and with CSR (on CPU).

The structure of the article is following. Section II describes the conducted numerical experiments. Section III presents the memory usage for formats used in experiments. In Section IV, the performance time of the experiments is analyzed. Section V concludes the experiments and the paper.

II. Methodology of Numerical Experiments

The memory requirements were tested and compared in this section, as well as the time of the uniformization algorithm, with the use of three storage schemes:

- HYB — the original format from the CUSP library;
- HYBIV — the modified format;
- the well-known CSR format (as the most common and often the most efficient format for CPUs), on CPU, with the use of the MKL library [13].

We were interested in studying and comparing memory required by the original HYB format and for our modified HYBIV format — on one GPU and on two GPUs. Also, the times elapsed by the uniformization method with the use of these formats for one and two GPUs were compared. The comparison between times of these formats on GPUs with the use of the CUSP library and the CSR format on CPU with the use of the MKL library was done.

All the codes are written in C++. We tested the uniformization on CPU and GPU under Linux with `gcc` and NVIDIA `nvcc` compilers with optimization flag `-O3`. The experiments were run on an Intel system with 12 cores. The Intel system has two sockets with six-core Intel Xeon X5650 clocked at 2.67 GHz and 48 GB memory. In the performance evaluation were used NVIDIA GPUs (2× Tesla M2050 with 3 GB memory) and libraries: CUDA Toolkit 4.0, CUSP 0.2, MKL 10.3.

We tested the implementations on four widely used benchmark models: mutex [7], a Kanban system [5], a cyclic server Polling system [9], tandem queueing network [10].

These protocols were chosen due to their scalability and the possibility to verify their properties by numerical solving. The first model (MUTEX) is generated by the authors, which allowed scaling up this model. The remaining three models were generated using PRISM [8] (unfortunately, we were unable to generate matrices of bigger size), a probabilistic model checker developed at the University of Birmingham.

Tables I and II present details of test matrices, where:

- *name* is the name of the matrix with parameters describing the model
- n is the number of rows,
- nz is the number of non-zero elements,
- nz/n represents the matrix density,
- uv is the number of unique values of matrix' elements,
- x is the size of the ELL part,
- c is the size of the COO part.

TABLE I
PROPERTIES OF THE MATRICES FOR THE 'MUTEX' MODELS

#	name	n	nz	nz/n	uv	x	c
1	MUTEX__N_16__R_4	2517	20949	8.32	654	0	20949
2	MUTEX__N_12__R_7	3302	38966	11.80	1545	0	38966
3	MUTEX__N_12__R_8	3797	47381	12.48	1871	0	47381
4	MUTEX__N_12__R_9	4017	51561	12.84	2030	0	51561
5	MUTEX__N_12__R_10	4083	52947	12.97	2078	0	52947
6	MUTEX__N_12__R_11	4095	53223	13.00	2081	0	53223
7	MUTEX__N_20__R_4	6196	52596	8.49	1259	5	21616
8	MUTEX__N_16__R_5	6885	68997	10.02	2008	6	27687
9	MUTEX__N_20__R_5	21700	223140	10.28	5067	6	92940
10	MUTEX__N_16__R_6	14893	173101	11.62	4943	17	0
11	MUTEX__N_20__R_9	431910	7222550	16.72	132862	21	0
12	MUTEX__N_20__R_16	1047225	21972345	20.98	225590	21	0
13	MUTEX__N_24__R_10	4540386	86052066	18.95	329608	25	0
14	MUTEX__N_24__R_12	9740686	211067278	21.67	364622	25	0
15	MUTEX__N_24__R_13	12236830	278463166	22.76	379255	25	0
16	MUTEX__N_24__R_14	14198086	335339590	23.62	392677	25	0

TABLE II
PROPERTIES OF THE MATRICES FOR THE 'KANBAN', 'POLL' AND 'TANDEM' MODELS

#	name	n	nz	nz/n	uv	x	c
17	kanban_sm-1	160	776	4.85	73	0	776
18	kanban_sm-2	4600	32720	7.11	142	5	9845
19	kanban_sm-3	58400	504800	8.64	191	9	25910
20	kanban_sm-4	454475	4434325	9.76	200	10	245084
21	kanban_sm-5	2546432	27006448	10.61	200	11	1268920
22	poll17_sm	3342336	34537472	10.33	51	11	1730158
23	poll18_sm	7077888	76677120	10.83	45	12	2653722
24	tandem_sm-2047	8386560	37724163	4.50	14	5	0

We divide matrices into the COO and ELL parts with the HYB format using CUSP. With very small matrices format COO is faster, therefore the ELL part remains empty. In Table I, there are matrix properties describing MUTEX model; the matrices differ in size and also in the proportion of COO and ELL. Six matrices with empty ELL part were chosen, three matrices with not empty COO and ELL parts and seven matrices with empty COO part.

In Table II we describe properties of the matrices from three remaining models: 'kanban', 'poll' and 'tandem'. Among the describe matrices only one has an empty COO part, and only one an empty ELL part. The other have not empty COO and ELL parts.

III. GPU MEMORY REQUIREMENTS

Memory usage was checked by the function `cudaMemGetInfo`. This function returns the total GPU memory and free GPU card memory which gives us information about memory occupation by the application. The method of the measurement is not precise enough to enable comparison of such small matrices. It prints total GPU memory usage, not only of matrices but also all additional values. The 'basic' value of 63 MB consists of not only explicitly allocated data but also the inner CUDA variables. It is visible that minimal, constant size is about 63 MB (for $n = 13$). The memory usage depends directly on nz. With small differences in nz it may happen that the memory use will be smaller for a bigger matrix. It is difficult to find dependencies of n because the matrices have different sparsity patterns.

Tables III and IV show the memory usage in MB on one and two GPUs (respectively) for the 'kanban', 'poll' and 'tandem' models. This memory was counted by function `cudaMemGetInfo` and therefore we call this memory experimental. Basing on Tables III and IV we can say that:

- The HYBIV format on larger matrices required almost twice less memory than HYB which results from experimental data.
- Memory usage per GPU was smaller in HYB2 and HYBIV2 than in HYB and HYBIV but it was higher by half because some variables were stored on each GPU.
- The number of non-zeros (nz) and the number of unique elements (uv) had the biggest influence on the memory occupation.

TABLE III
EXPERIMENTAL MEMORY USAGE (IN MB) ON ONE GPU FOR THE 'KANBAN', 'POLL' AND 'TANDEM' MODELS ($M_{exp} = \frac{\mathbf{HYB}}{\mathbf{HYBIV}}$)

#	name	**HYB**	**HYBIV**	M_{exp}
17	kanban_sm-1	64.45	65.45	0.98
18	kanban_sm-2	64.45	64.45	1.00
19	kanban_sm-3	69.70	67.82	1.03
20	kanban_sm-4	118.45	95.20	1.24
21	kanban_sm-5	402.60	254.58	1.58
22	poll17_sm	509.73	315.46	1.62
23	poll18_sm	1075.04	621.11	1.73
24	tandem_sm-2047	542.49	334.46	1.62

TABLE IV
EXPERIMENTAL MEMORY USAGE (IN MB) ON TWO GPUS FOR THE 'KANBAN', 'POLL' AND 'TANDEM' MODELS

#	name	**HYB2**			**HYBIV2**		
		gpu1	+	gpu2	gpu1	+	gpu2
17	kanban_sm-1	64.45	+	64.45	64.45	+	64.45
18	kanban_sm-2	64.45	+	65.45	64.45	+	64.45
19	kanban_sm-3	69.70	+	68.70	67.57	+	67.57
20	kanban_sm-4	103.95	+	99.32	89.82	+	86.32
21	kanban_sm-5	301.97	+	280.59	217.84	+	197.09
22	poll17_sm	375.61	+	349.98	265.46	+	239.96
23	poll18_sm	757.91	+	703.91	507.37	+	453.37
24	tandem_sm-2047	526.49	+	462.49	406.48	+	342.48

IV. TIME

All the processing times are reported in seconds. The time is measured with an MKL function `dsecnd`. Computations were made in double precision. Let us assume that $pt = \pi(0) = [1, 0, \ldots, 0]^T$. The choice of e_1 as the starting vector may seem too special, but it reflects the fact that we order the state space by reachability and that the Markov chain starts in the first state before evolving into other states.

Tables V and VI show run-time on CPU (CSR format, SpMV operation from MKL library), on one GPU (HYB and HYBIV formats, SpMV operation from CUSP library), on two GPUs (HYB2 and HYBIV2 formats, modified SpMV operation from CUSP library) respectively for $t = 1$, $t = 100$ and $\varepsilon = 10^{-10}$ (t and ε are parameters of the uniformization method), for matrices from 'MUTEX' model with numbers: 13, 14, 15 and 16. The bold values denote the fastest computation times and '—' denotes that the matrices could not be stored in the device memory.

Figures 1–6 show SpMV run-time (in seconds) on CPU, on one GPU, on two GPUs for the 'MUTEX' model.

On the basis of the charts of run-time it can be concluded that computation on CPU on CSR format takes the longest

TABLE V
RUN-TIME (IN SECONDS) ON CPU (CSR, SPMV FROM MKL); ON ONE GPU (HYB AND HYBIV, CUSP); ON TWO GPUS (HYB2 AND HYBIV2, CUSP) — THE 'MUTEX' MODELS FOR $t = 10$ AND $\varepsilon = 10^{-10}$

#	name	CPU	HYB	HYB2	HYBIV	HYBIV2
13	MUTEX__N_24__R_10	635.18	290.60	213.88	272.15	**207.79**
14	MUTEX__N_24__R_12	1844.90	—	552.48	757.49	**536.78**
15	MUTEX__N_24__R_13	1910.35	—	—	1048.57	**724.38**
16	MUTEX__N_24__R_14	3420.17	—	—	1314.93	**891.61**

TABLE VI
RUN-TIME (IN SECONDS) ON CPU (CSR, SPMV FROM MKL); ON ONE GPU (HYB AND HYBIV, CUSP); ON TWO GPUS (HYB2 AND HYBIV2, CUSP) — THE 'MUTEX' MODELS FOR $t = 100$ AND $\varepsilon = 10^{-10}$

#	name	CPU	HYB	HYB2	HYBIV	HYBIV2
13	MUTEX__N_24__R_10	6293.35	2923.13	2073.12	2758.9	**2001.43**
14	MUTEX__N_24__R_12	15206.1	—	5386.58	7647.88	**5212.61**
15	MUTEX__N_24__R_13	20242.1	—	—	10628.6	**7076.16**
16	MUTEX__N_24__R_14	33267.8	—	—	13396.4	**8767.05**

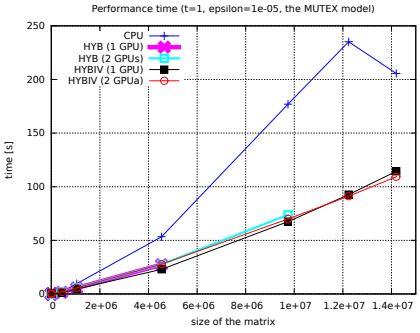


Fig. 1. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 1$, $\varepsilon = 10^{-5}$)

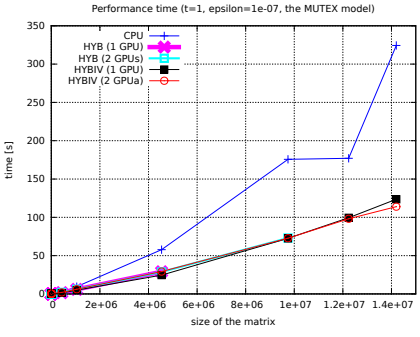


Fig. 2. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 1$, $\varepsilon = 10^{-7}$)

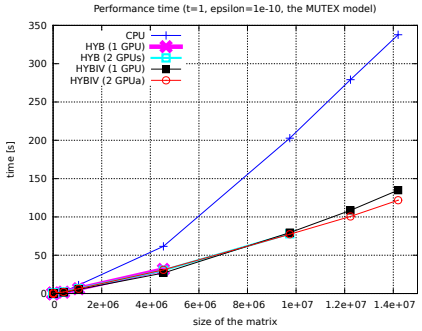


Fig. 3. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 1$, $\varepsilon = 10^{-10}$)

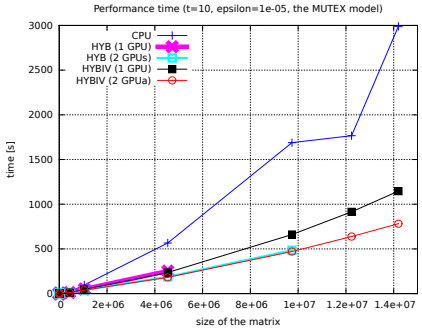


Fig. 4. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 10$, $\varepsilon = 10^{-5}$)

time, while GPU application speedup the run-time considerably.

The value of the variable l depends on t variable value. For 'MUTEX' model it is worth using two GPUs. Clearly, computations were done faster for proposed HYBIV format than for HYB format.

Tables VII and VIII show the time in seconds for the double precision uniformization method using the HYB and HYBIV storage formats on one GPU and two GPUs and the CSR

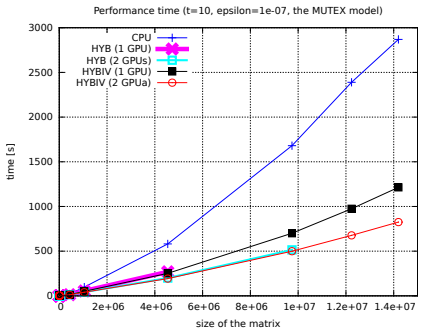


Fig. 5. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 10$, $\varepsilon = 10^{-7}$)

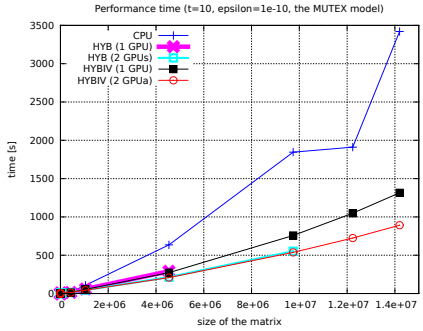


Fig. 6. Run-time (in seconds) on CPU (CSR, SpMV from MKL); on one GPU (HYB and HYBIV, CUSP); on two GPUs (HYB2 and HYBIV2, modified CUSP) — the 'MUTEX' model ($t = 10$, $\varepsilon = 10^{-10}$)

TABLE VII
SPMV RUN-TIME ON CPU (CSR, SPMV FROM MKL); ON ONE GPU (HYB AND HYBIV, CUSP); ON TWO GPUS (HYB2 AND HYBIV2, MODIFIED CUSP) — THE 'KANBAN', 'POLL' AND 'TANDEM' MODELS, $t = 1$, $\varepsilon = 10^{-10}$)

name	CPU	HYB	HYB2	HYBIV	HYBIV2
kanban_sm-1	**0.01**	0.02	0.02	**0.01**	0.02
kanban_sm-2	**0.01**	0.02	0.07	0.02	0.07
kanban_sm-3	0.03	0.04	0.12	**0.02**	0.12
kanban_sm-4	0.25	0.19	0.59	**0.06**	0.72
kanban_sm-5	1.58	1.04	3.40	**0.24**	5.98
poll17_sm	21.10	3.75	11.39	**2.10**	10.73
poll18_sm	35.70	8.37	26.26	**4.59**	24.59
tandem_sm-2047	1410.19	121.03	482.84	**88.31**	466.98

storage format from the MKL library on CPU for $\varepsilon = 10^{-10}$, $t = 1$ and $t = 10$ for the other three models - 'kanban', 'poll' and 'tandem'. The bold values denote the fastest computation times.

Run-times on two GPUs are slower than on one GPU for matrices which $\frac{nz}{n} < 16$ (for all matrices 'kanban', 'poll' and 'tandem' models and matrices number from 1 to 10 for 'MUTEX' model). There are too few computations to sensibly use two GPUs. For 'poll' and 'tandem' models we achieve considerable computation speedup on one GPU.

The best storage scheme — that is, the fastest and the most compact — for bigger transition rate matrices is HYBIV2 (HYBIV on 2 GPUs). The HYB storage format performs not quite efficiently in many cases. It is because the granularity (one thread per row) of the sparse matrix-vector multiplication is not fine enough for them, so the bigger the matrix, the better the utilization of the GPU. The performance of HYBIV was a little better than HYB, because in HYBIV less data is stored in slow global memory. Using two GPUs we almost doubled the performance in comparison to single GPU. For smaller matrices, HYBIV and splitting data across two GPUs were not useful.

TABLE VIII
SPMV RUN-TIME (IN SECONDS) ON CPU (CSR, SPMV FROM MKL); ON ONE GPU (HYB AND HYBIV); ON TWO GPUS (HYB2 AND HYBIV2) — THE 'KANBAN', 'POLL' AND 'TANDEM' MODELS, $t = 10$, $\varepsilon = 10^{-10}$)

name	CPU	HYB	HYB2	HYBIV	HYBIV2
kanban_sm-1	**0.01**	0.04	0.05	**0.01**	0.06
kanban_sm-2	**0.02**	0.05	0.26	0.05	0.26
kanban_sm-3	**0.06**	0.08	0.30	**0.06**	0.31
kanban_sm-4	1.06	0.36	1.14	**0.20**	1.24
kanban_sm-5	6.84	1.80	5.20	**0.85**	7.90
poll17_sm	197.15	25.98	73.42	**18.82**	68.49
poll18_sm	330.33	56.86	151.52	**40.85**	137.82
tandem_sm-2047	14475.40	1220.19	4873.97	**898.60**	4716.94

V. CONCLUSION

In this article, we investigated the use of a modified sparse memory format on the GPU in a practical problem, namely calculating probabilities from Markov transition matrices. Our results showed that in the case of small size matrices, we did not achieve high performance in the HYBIV format or significant memory savings. However, the proposed method reduces the memory size for storing larger matrices. In addition, the use of HYBIV does not degrade performance and the use of two GPUs allows the processing of larger matrices than one GPU. In our future work, we try to transfer codes to the CUDA version beyond 4 and use streams to optimize overall performance.

REFERENCES

[1] Bylina, B., Bylina, J., Karwacki, M.: Computational Aspects of GPU-accelerated Sparse Matrix-Vector Multiplication for Solving Markov Models. Theoretical and Applied Informatics 23, 127–145 (2011)
[2] Bylina, B., Karwacki, M., Bylina, J.: A CPU-GPU Hybrid Approach to the Uniformization Method for Solving Markovian Models — A Case Study of a Wireless Network. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2012. CCIS, vol. 291, pp. 401–410. Springer, Heidelberg (2012)
[3] Bylina, B., Karwacki, M., Bylina, J.: Multi-GPU Implementation of the Uniformization Method for Solving Markov Models. In: Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS) 2012, pp. 533–537 (2012)
[4] Bylina, J., Bylina, B., Karwacki, M.: An Efficient Representation on GPU for Transition Rate Matrices for Markov Chains; Lecture Notes in Computer Science 8384, 663–672. Springer, Heidelberg (2014)
[5] G. Ciardo and M. Tilgner.: On the use of Kronecker Operators for the Solution of Generalized Stochastic Petri Nets. ICASE Report 96-35, Institute for Computer Applications in Science and Engineering, 1996.
[6] N. J. Dingle, P. G. Harrison, W. J. Knottenbelt: *Uniformization and hypergraph partitioning for the distributed computation of response time densities in very large Markov models*, Journal of parallel and distributed computing, 64 (2004), 908-920
[7] Fernandes, P., Plateau, B., Stewart, W.J.: Efficient Descriptor-Vector Multiplication in Stochastic Automata Networks. J. ACM 45, 381–414 (1998)
[8] M. Z. Kwiatkowska, G. Norman, D. Parker: PRISM: Probabilistic Symbolic Model Checker, Computer Performance Evaluation, Modelling Techniques and Tools 12th International Conference, TOOLS 2002, LNCS 2324, pp.200-204, Springer, 2005.
[9] Resing, J. A. C. (1993). Polling systems and multitype branching processes. Queueing Systems 13 (4): 409–426
[10] H. Hermanns, J. Meyer-Kayser and M. Siegle. Multi Terminal Binary Decision Diagrams to Represent and Analyse Continuous Time Markov Chains. In B. Plateau and W. Stewart and M. Silva (editors), Proc. 3rd International Workshop on Numerical Solution of Markov Chains (NSMC'99), pages 188-207, Prensas Universitarias de Zaragoza. 1999.
[11] R. B. Sidje: *Expokit: A software package for computing matrix exponentials*, ACM Trans. Math. Software, 24 (1998), pp. 130–156.
[12] R. B. Sidje, K. Burrage, S. MacNamara: *Inexact Uniformization Method for Computing Transient Distributions of Markov Chains.* SIAM J. Scientific Computing 29(6): 2562–2580 (2007).
[13] Intel Math Kernel Library, http://software.intel.com/en-us/articles/intel-mkl/

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 311–318
DOI: 10.15439/2018F40
ISSN 2300-5963 ACSIS, Vol. 15

Multithreaded Parallelization of the Finite Element Method Algorithms for Solving Physically Nonlinear Problems

Sergiy Fialko
Tadeusz Kościuszko Cracow University of Technology
ul. Warszawska 24, 31-155 Kraków, Poland
Email: sergiy.fialko@gmail.com

Viktor Karpilowskyi
IT company SCAD Soft
ul. Osvity 3a, office 1, 2, Kiev, Ukraine
Email: kvs@scadsoft.com

***Abstract*—The parallelization of the leading procedures of the finite element method applied to solving physically nonlinear problems of structural mechanics is considered.**

I. Introduction

THE SOLUTION of physically nonlinear problems of structural mechanics by the finite element method requires a large number of calculations. The vast majority of such problems are solved on multi-core shared memory computers - desktops, laptops and multiprocessor workstations with SMP (Symmetric Multiprocessing) architecture. When using the finite element method, the most time-consuming procedures are the assembling of a tangent stiffness matrix, the evaluation of an internal force vector, and the solution of a system of linear algebraic equations with a sparse symmetric matrix. Solver PARDISO [1] from the Intel Math Kernel Library (MKL) [2] or PARFES [3], [4] is used to solve systems of linear algebraic equations with a symmetric sparse tangent stiffness matrix. These solvers have successfully proven themselves on multi-core computers of SMP architecture and demonstrate stable acceleration with an increase in the number of cores during the factorization stage. In addition, forward and backward substitutions are also parallelized. Therefore, this paper considers parallel algorithms for the assembling of a tangent stiffness matrix and the calculation of an internal force vector.

A. Assembling of the stiffness matrix

1) Related works: In [13] is presented a parallel node-by-node assembling approach, when each block of the global stiffness matrix, corresponding to given node, is processed on the same processor. The blocks of an element matrices related to given node are evaluated on this processor. There are no overlapping of blocks in the global stiffness matrix and no communication between processors are needed.

In [14] is considered only banded matrices. The set of consecutive rows are taken as a synchronization region. Such a partitioning the matrix into a sufficient number of synchronization regions allows to avoid simultaneous modification of the same elements by the different threads.

The algorithm [15] assembles the stiffness matrix by groups of rows related to each node of the finite element mesh. Each processor will only assemble the rows related to a specific group of nodes. Therefore, no synchronization is required because each processor updates only their addresses of the memory.

The method [16] creates for each element the list of the processors onto which the associated columns in global stiffness matrix composing this element have been mapped. If this list consists in only one processor, the finite element is totally local to this processor, and non totally local otherwise. The totally local elements are mapped to each processor. The remaining finite elements are processed on several processors and need a synchronization.

In [17] before assembling is precomputed a coloring of the finite elements such that no two elements of the same color share any given degree of freedom. The appropriate heuristic coloring algorithm is presented. The several algorithms realising assembling on GPU, are presented.

We present the procedure of the stiffness matrix assembling (section II.B) which require no synchronization between threads and demonstrates an almost perfect load balance even for different types of finite elements – triangular and quadrilateral shell elements, spatial bar elements and so on. It is the typical situation in structural analysis when design model consists of different types of finite elements, and evaluation of their stiffness matrices requires the quite different computational efforts.

2) Assembling procedure: The procedure for a tangent stiffness matrix assembling is as follows:

$$\mathbf{K}_t = \sum_{e=1}^{Ne} \mathbf{P}_e^T \mathbf{K}_{e,t} \mathbf{P}_e, \tag{1}$$

where $\mathbf{K}_{e,t}$ is a tangent stiffness matrix of the e-th finite element, $\mathbf{P}_e$ is a permutation matrix and N_e is a number of finite elements in the design model. In the case of physically

This work was supported by IT company SCAD Soft

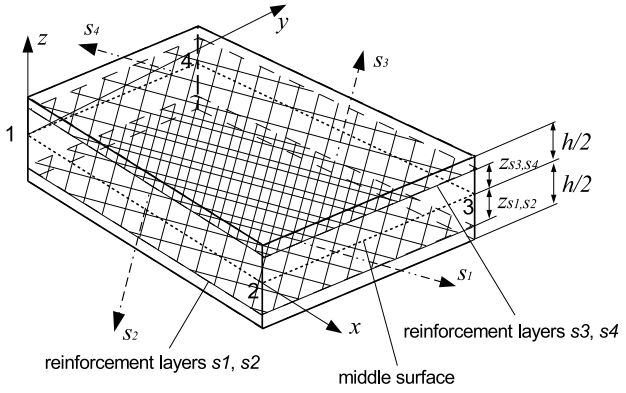


Fig. 1. The flat shell finite element

nonlinear problems, the calculation of the tangent stiffness matrix for the thin plane shell finite element (Fig. 1) is represented as a sum of the following type integrals:

$$\mathbf{K}_e = \int\limits_{\Omega} \mathbf{B}^T(\Omega) \left[\int\limits_{-h/2}^{h/2} f(\Omega, z)\, dz \right] \mathbf{B}(\Omega)\, d\Omega, \qquad (2)$$

where Ω is an in-plane domain ($d\Omega = dxdy$), h is the shell thickness, $\mathbf{B}(\Omega)$ is a deformation matrix [5], $s_1, s_2, ...$ are the axes of the reinforcement layers, $z_{s1}, z_{s2}, ...$ are the distances between reinforcement layers $s_1, s_2, ...$ and middle surface.

The trapezoid method is applied to calculate the integral over the shell thickness. In this case, the shell is divided into 10 - 40 layers through thickness. In addition, reinforcing rods of the same direction form a layer of reinforcement. Usually, the number of reinforcement layers is 4, although it may be arbitrary. The integral over the domain of the finite element Ω is computed by the Gauss method using the 2×2 integration scheme. The components of the stress and strain tensors are calculated at each Gaussian point. The number of such points for a given type of finite element is $2 \times 2 \times$ (the number of layers plus the number of reinforcement layers). The tangent stiffness matrix for other types of finite elements is defined similarly.

Thus, the procedure for the tangent stiffness matrix assembling requires significant computational effort. The tangent stiffness matrix assembling time is of the same order as the factorization time of this matrix.

B. Internal force vector evaluation

The internal force vector is calculated as follows:

$$\mathbf{f}^{int} = \sum_{e=1}^{Ne} \mathbf{P}_e^T \mathbf{r}_e \mathbf{P}_e, \qquad (3)$$

where

$$\mathbf{r}_e = \mathbf{K}_e \mathbf{q}_e, \qquad (4)$$

$\mathbf{q}_e$ is a nodal displacement vector and $\mathbf{r}_e$ is a nodal reaction vector for the e-th finite element. The stiffness matrix $\mathbf{K}_e$ has to be calculated for each finite element in the expression (3). It is not a tangent stiffness matrix, but it is a full stiffness matrix [6], [7], [9].

This paper is devoted to the technique of multithreaded parallelization of problems (1) and (3).

II. Multithreaded Parallelization

A. Internal force vector evaluation

First of all, we consider parallelization of the problem (3), the corresponding Algorithm 1 is presented below.

Algorithm 1 Internal force vector evaluation

1: Initialization.
   $\mathbf{rr}_{ip} \leftarrow 0$, $ip \in [0,\ np-1]$, $\mathbf{f}^{int} \leftarrow 0$

2: **for parallel** $e = 1$ **to** N_e **schedule(dynamic) do**
3: ip = $omp\_get\_thread\_num()$
4: *Compute a transformation matrix* $\mathbf{T}_e$
   $\mathbf{u}_e^{glob} \leftarrow \mathbf{u}$
   $\mathbf{u}_e^{loc} = \mathbf{T}_e \mathbf{u}_e^{glob}$
5: *Compute a nodal reaction vector* $\mathbf{r}_e^{loc}$, *using the constitutive relations.*
   $\mathbf{r}_e^{glob} = \mathbf{T}_e \mathbf{r}_e^{loc}$
   $\mathbf{rr}_{ip} \leftarrow + \mathbf{r}_e^{glob}$
6: **end for**

7: **for** $ip = 0$ **to** $np - 1$ **do**
8: **for parallel** $eqn = 1$ **to** Neq **schedule(dynamic, chunk) do**
9: $\mathbf{f}_{eqn}^{int} += \mathbf{rr}_{ip,eqn}$
10: **end for**
11: **end for**

At the initialization stage (point 1), we dynamically allocate memory for vectors $\mathbf{rr}_{ip}$, where ip is the thread number and np is the number of threads. After Algorithm 1 is finished, vector $\mathbf{f}^{int}$ will store internal forces. Vectors $\mathbf{rr}_{ip}, ip \in [0, np-1]$, and $\mathbf{f}^{int}$ have the dimension Neq - the number of equations in the finite element model. All these vectors are zeroed.

At the second stage (points 2 – 6) we run a parallel loop **for** over the number of finite elements N_e, where e is a number of the current finite element. We obtain the thread number ip (point 3) end evaluate the coordinate transformation matrix $\mathbf{T}_e$ (point 4). Then we put elements of the displacement vector $\mathbf{u}$, corresponding to the degrees of freedom of the finite element e, to vector $\mathbf{u}_e^{glob}$. Vector $\mathbf{u}$ of dimension Neq (number of equation) contains the nodal displacements and rotations in the global coordinate system (CS) for the entire finite element model. Vector $\mathbf{u}_e^{glob}$ of dimension $nstAct$ contains the nodal displacements and rotations of the finite element e. The displacements and rotations in the global CS are transformed into the displacements and rotations in the local CS of the e-th finite element with the help of the coordinate transformation matrix $\mathbf{T}_e$.

The evaluation of the nodal reaction vector $\mathbf{r}_e^{loc}$ in the local CS is performed according to (4) (point 5). The constitutive relations, reflecting the mechanical rules, are applied to evaluate the full stiffnes matrix $\mathbf{K}_e$. Then, the transformation of the reaction vector, given in the local CS, to the global CS is performed ($\mathbf{r}_e^{glob} = \mathbf{T}_e \mathbf{r}_e^{loc}$). After this, the elements of the reaction vector $\mathbf{r}_e^{glob}$ are added ($\leftarrow +$) to the corresponding elements of the vector $\mathbf{rr}_{ip}$.

At the third stage (points 7 – 11), partially prepared reaction vectors $\mathbf{rr}_{ip}, \quad ip \in [0, np-1]$ are combined in the vector $\mathbf{f}^{int}$ of internal forces. We parallelize an inner loop (points 8 – 9) covering the equations of the entire model. To avoid incoherence in the processor caches when writing data to vector $\mathbf{f}^{int}$ at multithreading, the chunk size **chunk** is assumed to be sixteen. In doing so, we assume that the size of the cache line is 64 or 128 bytes, that is, eight or sixteen double words respectively. It is important to do it to avoid a degradation of performance at multithreading on processors, which are not protected against such incoherence at the hardware level.

B. Stiffness matrix assembling

Unlike the internal forces evaluation algorithm discussed in the previous subsection, the tangent stiffness matrix assembling algorithm stores a large array to memory - a nonzero structure of sparse matrix that comprises only nonzero elements. The symmetric sparse matrix in compressed column format (CCS) is represented in the form of three arrays: $Space[pos], ind[pos]$ and $Pos[j]$. Array $Space[pos]$ comprises nonzero elements located column-by-column from the diagonal element to the last nonzero element of the column. Array $ind[pos]$ stores the i subscript for the corresponding element k_{ij} of matrix $\mathbf{K}_t$ located in $Space[pos]$. Array $Pos[j]$ points to the position pos of the diagonal element k_{jj} of the column j in the arrays $Space$, ind. Here $j \in [1,\ Neq], \quad i \geq j$. For instance, the CCS format for the matrix (5) is presented in (6).

$$\begin{pmatrix} k_{11} & & & & & \\ 0 & k_{22} & & & & \\ 0 & 0 & k_{33} & & & \\ k_{41} & 0 & 0 & k_{44} & & \\ 0 & k_{42} & 0 & 0 & k_{55} \end{pmatrix} \tag{5}$$

pos	0	1	2	3	4	5	6
$Space$	k_{11}	k_{41}	k_{22}	k_{42}	k_{33}	k_{44}	k_{55}
ind	1	4	2	4	3	4	5
Pos	0	2	4	5	6	7	
j	1	2	3	4	5	–	

(6)

Subscript j is a column number. The last element in Pos is required to properly obtain the number of nonzero elements in the last column using the following expression $nonzero_j = Pos[j+1] - Pos[j]$, where $nonzero_j$ – the number of nonzero entries in the column j counting from the diagonal element to the lowest one. The nonzero structure of a symmetric sparse matrix is presented by the undirected adjacency graph, where vertexes correspond to columns of matrices or to diagonal elements and the edges – to nonzero off-diagonal elements [10].

The reordering of the adjacency graph reduces the number of fill-in in the factorized matrix. The symbolic factorization procedure [10] enables to create a nonzero structure of the factorized sparse matrix without numerical factorization. It is a very fast procedure, which operates only on the adjacency graph of the initial matrix $\mathbf{K}_t$ and obtains a factor-graph. As a result, the arrays Pos and ind are filled. Therefore, the assembling procedure fills the array $Space$. Algorithm 2 demonstrates a trivial solution of a such problem.

Algorithm 2 Assembling of the tangent stiffness matrix with using of critical sections (trivial solution)

```
1:  for parallel e = 1 to N_e schedule(dynamic) do
2:    ip = omp_get_thread_num()
3:    evaluate a finite element matrix K_{e,t} and the list of
      global equation numbers list_glob_eqns
4:    for jeqn = 1 to nstAct do
5:      jglobeqn = list_glob_eqns(jeqn)
6:      prepare the inverse data structure to avoid a search
        procedure
7:      for pos = Pos[jglobeqn] to Pos[jglobeqn+1] - 1
        do
8:        iglobeqn = ind[pos]
          Col[ip][iglobeqn] = pos
9:      end for
        fill Space:
10:     for ieqn = jeqn to nstAct do
11:       iglobeqn = list_glob_eqns(ieqn)
          pos = Col[ip][iglobeqn]
12:       begin_critical_section
13:         Space[pos] += K_{e,t}[ieqn, jeqn]
14:       end_critical_section
15:     end for
16:   end for
17: end for
```

The parallel loop **for** is performed over the number of finite elements (point 1). The thread number ip is defined (point 2) and the finite element tangent stiffness matrix $\mathbf{K}_{e,t}$ with the list of global equation numbers $list\_glob\_eqns$ are obtained (point 3).

The loop **for** is performed over the local equation number $jeqn$ (point 4 – 15). A column number $jglobeqn$ of the global tangent stiffness matrix $\mathbf{K}_t$ (point 5) is defined for each column number $jeqn$ of the matrix $\mathbf{K}_{e,t}$. Then, the inverse data structure $Col[ip][iglobeqn]$ is prepared to avoid a time-consuming search of the position pos in the array $Space$ (points 7 – 9).

Loop **for** (points 10 – 15) runs along the column $jeqn$ of the matrix $\mathbf{K}_{e,t}$. The position number pos is extracted from $Col[ip][iglobeqn]$, where $ieqn$ is a row number of the matrix $\mathbf{K}_{e,t}$ and $iglobeqn$ is a row number of the global tangent stiffness matrix $\mathbf{K}_t$ (point 11 and a line below). To avoid the situation, when several threads simultaneously modify the

same element of the array $Space[pos]$, a critical section is used (points 12 – 14).

Using a critical section in such a situation is not a good idea, because the speedup with the increasing number of threads begins to degrade very fast. The application of interlocked functions [11] instead of a critical section does not improve the speedup much.

Therefore, we do not use Algorithm 2. The following approach is proposed instead. We divide the finite elements into groups so that each group is simultaneously processed by different threads and writes only in its positions pos of the array $Space$. In other words, a situation, when different threads simultaneously write data to the same addresses of the array $Space$, should be eliminated.

To create such groups, we prepare an adjacency graph for the finite elements of the design model (Fig. 2). The vertices of the graph present the finite elements and the edges – the nodes in which the adjacent finite elements are coupled.

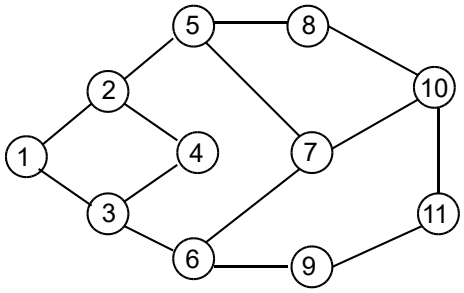


Fig. 2. The finite element adjacency graph. The vertices present the finite elements and edges – the nodes of the design model.

Then, we search for a pseudo peripherical vertex and create a structure of levels with a root in such a vertex [10] (Fig. 3).

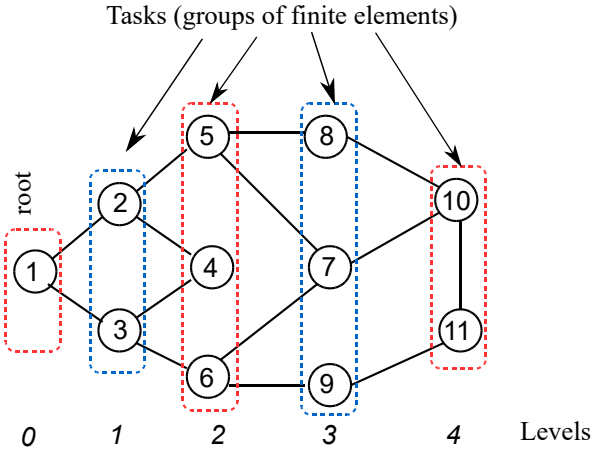


Fig. 3. The structure of levels for the finite element adjacency graph with a root in a (pseudo) peripherical vertex

The root of the level structure is placed in the pseudo peripheral vertex in order for the structure to be maximally elongated (containing as many levels as possible) and to minimize the number of vertices on each level. We call the vertices belonging to a given level a computational task or just a task. If we choose tasks of only even levels, then the vertices of each task are not connected in any way with the vertices of the remaining selected tasks. In other words, the finite elements belonging to each even-level task do not have any common nodes with finite elements belonging to tasks of other even levels. Thus, we guarantee that the finite elements belonging to different even levels make contributions to different degrees of freedom of the design model, that is, the modification of different elements of the $Space$ array is performed. Therefore, each even-level task can be performed simultaneously at multithreaded data processing. In the same way, as in the case of even levels, each task belonging to odd level can be solved independently from the other tasks of odd levels.

To achieve an acceptable load balance between threads, the Algorithm 3 is applied.

Algorithm 3 Mapping tasks onto threads

1: *Prepare finite element adjacency graph*
2: *Find a (pseudo) peripherical vertex*
3: *Create a structure of levels for the finite element adjacency graph with the root in a (pseudo) peripherical vertex*

Prepare the weights for even levels:
4: **for** *lev=0;* **to** *NoLevels-1, lev += 2* **do**
5: $levelWeight_{lev} = \sum_{e \in lev} nstAct_e^2$
6: **end for**

7: *Sort the even levels in descending order of their weights. Map tasks of even levels onto threads:*
$sumWeigh\,[ip] \leftarrow 0, \quad ip \in [0, np-1]$
8: **for** *lev=0;* **to** *NoLevels-1, lev += 2* **do**
9: *Find the thread number* min_ip *having a minimum sum of already mapped weights,* $min\_ip \in [0, np-1]$
10: $queue\,[min\_ip] \leftarrow \forall e \in lev$
$sumWeigh\,[min\_ip] += levelWeight_{lev}$
11: **end for**

Prepare the weights for odd levels:
12: **for** *lev=1;* **to** *NoLevels-1, lev += 2* **do**
13: $levelWeight_{lev} = \sum_{e \in lev} nstAct_e^2$
14: **end for**

15: *Sort the odd levels in descending order of their weights. Map tasks of odd levels onto threads:*
$sumWeigh\,[ip] \leftarrow 0, \quad ip \in [0, np-1]$
16: **for** *lev=1;* **to** *NoLevels-1, lev += 2* **do**
17: *Find the thread number* min_ip *having a minimum sum of already mapped weights,* $min\_ip \in [0, np-1]$
18: $queue1\,[min\_ip] \leftarrow \forall e \in lev$
$sumWeigh\,[min\_ip] += levelWeight_{lev}$
19: **end for**

The points 1 – 3 of the presented algorithm have been discussed above. After creating the level structure with a root

at a (pseudo) peripherical vertex, we obtain the weight of each even level (points 4 – 6). Here, the $NoLevels$ is the number of levels of a level structure and $levelWeight_{lev}$ is defined as a sum of weights for all vertices (finite elements) belonging to the level lev. The dimension of the finite element tangent stiffness matrix $\mathbf{K}_{e,t}$ is denoted as $nstAct_e$ and the weight of the vertex is accepted as a number of elements $nstAct_e^2$ in $\mathbf{K}_{e,t}$.

Then, we sort all vertices belonging to the even levels, in descending order of their weights, and zero the sum of weights $sumWeight\,[ip]$ mapped onto thread ip, $\quad ip \in [0, np-1]$ (point 7). For each even level we find a thread min_ip which has a minimum sum of already mapped weights (point 9), put all vertices of the given level to the $queue\,[min\_ip]$ and correct $sumWeight\,[min\_ip]$ (point 10).

Finally, we apply the same approach to all the odd levels and obtain $queue1\,[ip]\,, ip \in [0, np-1]$ (points 12 – 19).

Therefore, all even levels are mapped onto np threads and are presented by $queue\,[ip]\,, \quad ip \in [0, np-1]$ queues. Similarly, all odd levels are presented by $queue1\,[ip]\,, \quad ip \in [0, np-1]$ queues.

Sorting in descending order improves a load balance between threads. A similar approach has been used in [3], [4] to achieve a load balance between threads in solver PARFES and in block incomplete Cholesky factorization solver [8].

The Algorithm 4 demonstrates a tangent stiffness matrix assembling using multithreading without any synchronization allowing a high speedup with the increasing thread number.

In the first parallel region (points 1 – 18) the **while** loop runs for each thread ip until the $queue\,[ip]\,, \quad ip \in [0, np-1]$ is empty. These queues contain parallel tasks for even levels of level structure. In each **while** loop, the nearest finite element number e is retrieved (point 4) and the tangent stiffness matrix $\mathbf{K}_{e,t}$ with the list of global equation numbers are evaluated (point 5). The loop **for** is executed over the columns of the matrix $\mathbf{K}_{e,t}$ (points 6 – 16), where *nstAct* is a dimension of $\mathbf{K}_{e,t}$. The $Col\,[ip]$ array stores a position number *pos* of the nonzero entry $Space\,[pos]$ with the global equation number *iglobeqn* (points 9 – 11). It allows us to avoid a time-consuming search of pos, corresponding to the global equation number $iglobeqn$. The loop **for** (points 12 – 15) fills the $Space\,[pos]$ with elements of the matrix $\mathbf{K}_{e,t}$.

The second parallel region (points 19 – 35) does the same as the previous parallel region, but operates with queues $queue1\,[ip]$, containing parallel tasks for odd levels of a level structure. In contrast to Algorithm 2, Algorithm 4 does not contain any synchronization objects due to the approach presented above. This allows us to hope a high speedup with the increasing thread number, even when the number of threads is large. This approach formed the basis for the master's degree thesis in computer science [12], in which one of the authors, S. Fialko, was a scientific leader.

Algorithm 4 Assembling of the tangent stiffness matrix using the proposed approach

1: **parallel region**
2: $ip = omp\_get\_thread\_num()$
3: **while** $queue\,[ip]\,is \quad not \quad empty$ **do**
4: $e \leftarrow queue\,[ip]$ *retrieve element number* e
5: *evaluate a finite element matrix* $\mathbf{K}_{e,t}$ *and the list of global equation numbers* $list\_glob\_eqns$
6: **for** $jeqn = 1$ **to** $nstAct$ **do**
7: $jglobeqn = list\_glob\_eqns\,[jeqn]$
8: *prepare the inverse data structure to avoid a search procedure*
9: **for** $pos = Pos[jglobeqn]$ **to** $Pos[jglobeqn+1]-1$ **do**
10: $iglobeqn = ind[ieqn]$
$Col\,[ip]\,[iglobeqn] = pos$
11: **end for**
fill Space:
12: **for** $ieqn = jeqn$ **to** $nstAct$ **do**
13: $iglobeqn = list\_glob\_eqns\,[ieqn]$
$pos = Col\,[ip]\,[iglobeqn]$
14: $Space\,[pos] += \mathbf{K}_{e,t}\,[ieqn, jeqn]$
15: **end for**
16: **end for**
17: **end while**
18: **end of a parallel region**

19: **parallel region**
20: $ip = omp\_get\_thread\_num()$
21: **while** $queue1\,[ip]\,is \quad not \quad empty$ **do**
22: $e \leftarrow queue1\,[ip]$ *retrieve element number* e
23: *evaluate a finite element matrix* $\mathbf{K}_{e,t}$ *and the list of global equation numbers* $list\_glob\_eqns$
24: **for** $jeqn = 1$ **to** $nstAct$ **do**
25: $jglobeqn = list\_glob\_eqns\,[jeqn]$
prepare the inverse data structure to avoid a search procedure
26: **for** $pos = Pos[jglobeqn]$ **to** $Pos[jglobeqn+1]-1$ **do**
27: $iglobeqn = ind\,[ieqn]$
$Col\,[ip]\,[iglobeqn] = pos$
28: **end for**
fill Space:
29: **for** $ieqn = jeqn$ **to** $nstAct$ **do**
30: $iglobeqn = list\_glob\_eqns\,[ieqn]$
$pos = Col\,[ip]\,[iglobeqn]$
31: $Space\,[pos] += \mathbf{K}_{e,t}\,[ieqn, jeqn]$
32: **end for**
33: **end for**
34: **end while**
35: **end of a parallel region**

III. Numerical Results

We consider a design model of a reinforced concrete floor slab, comprising 65 117 equations (Fig. 4). The triangular and

quadrilateral finite elements, taking into account the physical nonlinearity, are used. The supports, modeling the walls, are shown in blue color. The uniform normal pressure simulates the dead and operational loads.

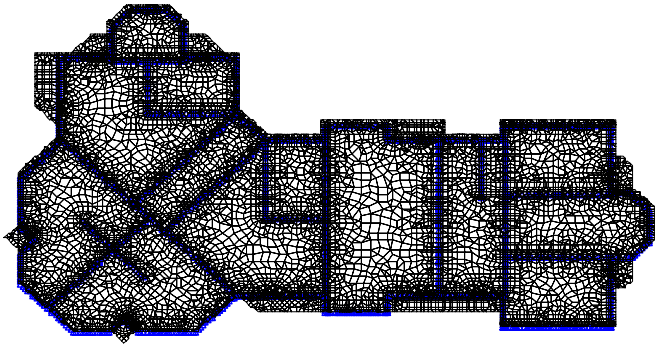


Fig. 4. The design model of a reinforced concrete floor slab

The computer with sixteen-core AMD Opteron 6276 processor, 2.3/3.2 GHz, 64 GB DDR3 RAM, OS Windows Server 2008 R2 Enterprise SP1, 64-bit, is used. Table I depicts a tangent stiffness matrix assembling time (s) during the solution of the entire nonlinear problem for the different number of threads np in the case with no optimization, optimization using a (pseudo) peripherical vertex and optimization using a (pseudo) peripherical vertex and weights of levels.

The "no optimization" case means that a root for the level structure is taken in the vertex 1. The pseudo peripherical vertex is not found. The weights of levels are not used. The queues $queue\,[ip]$ and $queue1\,[ip]\,,\quad ip \in [0, np-1]$ are prepared using a cyclic mapping of levels onto threads.

In the "use a peripherical vertex" case the pseudo peripherical vertex is found, but the weights of levels are not used. The cyclic mapping of levels onto threads is applied. Taking the pseudo peripherical vertex as a root of the level structure results in an increase of the levels from 110 to 114 for the given problem. It should be pointed out that for other problems the choice of a root in a pseudo peripherical vertex has a considerably larger impact on the increase of the level number. Therefore, for the considered problem of the "use a peripherical vertex" option on the reduction of the computing time is not observed.

Algorithm 3 is applied in the "plus the use of the weights of levels" case. The red color means that the load imbalance between threads exceeds 15%. The shortest computing time on a large number of threads is achieved when all optimizations are applied.

Table II shows the distribution of sum of weights among threads for the "no optimization" and "use peripherical vertexes and weights of levels" cases. Here, ip is a thread number, a red color indicates a thread with a maximum computational effort and a green color indicates a thread with a minimum computational effort. The difference between the maximum sum of weights and the minimum one is a measure of imbalance between threads. These results demonstrate that at a maximum number of threads the proposed approach, corresponding to Algorithm 3 and Algorithm 4, has a imbalancee between threads of about 11%. The "no optimization" approach has a imbalance of about 36%. We estimate a imbalance as (maximum sum of weights – minimum sum of weights)/maximum sum of weights in percent. Therefore, the above optimizations play an important role in improving a load balance between threads.

The Fig. 5 demonstrates a speedup with the increasing thread number in the range of the physical core number for the given processor: $S_{np} = T_1/T_{np}$, where T_1 is a time when using one thread and T_{np} is a time on np threads. The "ideal" curve corresponds to an ideal speedup, passing through the points (0, 0), (1, 1), (2, 2), ... However the given processor has a turbo core mode. When a small number of cores are loaded, the clock frequency is 3.2 GHz. When the number of loaded cores is a maximum, the clock frequency reduces to 2.3 GHz. Therefore, an ideal speedup is unreachable.

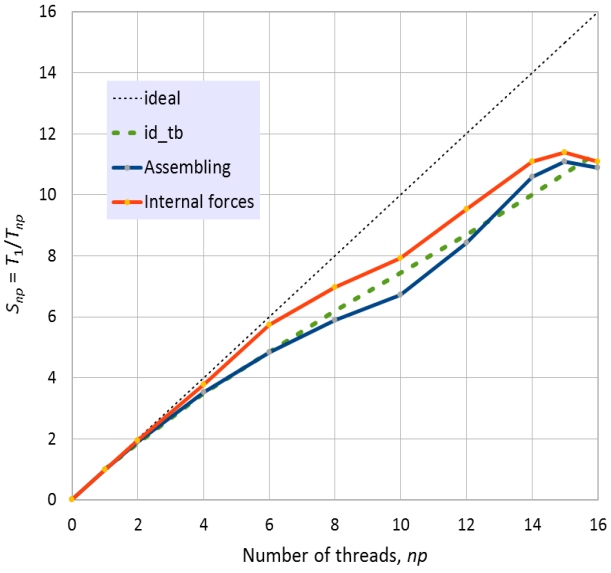


Fig. 5. The speedup with the increasing thread number for the evaluation of the internal forces and the tangent stiffness matrix assembling

The curve "id_tb" approximates an ideal speedup caused by the turbo core mode using a square parabola. The points (0, 0), (1, 1) and (16, x) are used to define such a parabola. The ordinate x is obtained as follows. A processor with sixteen cores without a turbo core mode should work with a clock frequency 3.2 GHz and should achieve a speedup of 16 times. A processor in the turbo core mode works with a clock frequency 2.3 GHz and has a speedup of x times. So, from the proportion, we obtain x = 2.3/3.2*16 = 11.5 times.

The "Internal forces" curve demonstrates a speedup of the internal force vector evaluation with the increasing thread number (Algorithm 1). The "Assembling" curve depicts a speedup of the tangent stiffness matrix assembling procedure with the above optimizations, presented by Algorithms 3

TABLE I
THE TANGENT STIFFNESS MATRIX ASSEMBLING TIME DEPENDING ON THE NUMBER OF THREADS

nos of threads	no optimization	use a peripherical vertex	plus the use of the weights of levels
1	84.7	85.9	86.7
2	46.6	46.2	45.6
4	26.3	26.8	27.6
8	17.7	18.2	16.6
12	13.6	14.4	12.8
16	11.3	11.6	10.4

TABLE II
SUM OF WEIGHTS PER EACH THREAD

thread number ip	no optimization		use a peripherical vertex and weights of levels	
	even levels	odd levels	even levels	odd levels
0	177264	183564	194760	196524
1	184644	187020	195012	197280
2	189108	208404	197820	198252
3	225216	242856	196488	195084
4	254772	248400	197532	206640
5	245304	228528	195840	194076
6	223272	218484	193392	194148
7	193572	174780	208512	210492
8	175464	169524	204840	194652
9	162432	163980	195048	211284
10	164268	168444	216396	205812
11	181224	185976	202140	202428
12	195804	202824	197460	194112
13	219816	229248	201348	193788
14	218160	190908	206784	196596
15	186840	187344	193788	199116

and 4. We obtain an acceptable correlation between "Internal forces" and "Assembling" curves with the "id_tb" curve. A steady increase in the speedup, with the exception of the last point (np = 16), confirms the effectiveness of the proposed approaches.

It should be noted that if we define speedup as $\frac{T_1 \cdot \Delta t_{np}}{T_{np} \cdot \Delta t_1}$, where Δt_1 and Δt_{np} are the time of a single processor's tick when using a single thread and when using the np threads correspondingly, then this ratio is not dependent on the processor clock speed, and the curve describing an algorithm's acceleration can be compared with the ideal speedup (curve "ideal"). However, for users, the definition of speedup as $S_{np} = T_1/T_{np}$, based on real-time execution of the tasks, is more understandable, therefore we use such a definition and above approach.

IV. CONCLUSION

Two different approaches for multithreaded parallelization of similar procedures – internal force vector evaluation and tangent stiffness matrix assembling have been considered.

The first approach requires the allocation of an additional vector of dimension Neq (number of equations) for each thread. Therefore, the amount of additional core memory is $Neq \times np$ words of double. On the other hand, such an approach is relatively simple and fully eliminates incoherences in caches of different processor cores.

The second approach, based on creating a finite element adjacency graph and preparing a level structure, ensures the independence of computational tasks, belonging to only even levels or only odd levels of a level structure, allows us to reject any type of synchronization and obtain a stable speedup with an acceptable correlation in comparison with an ideal speedup, taking into account the turbo core mode. Taking a pseudo peripherical vertex of the adjacency graph as a root of the level structure results in an increase of the levels number, so, the number of computational tasks increases too and each task becomes shorter. Together with a specific mapping-tasks-onto-threads algorithm, using the weights of computational tasks, this approach significantly improves the load balance between threads (Tables I, II) and helps to achieve a stable speedup.

On the other hand, the second approach does not guarantee the absence of incoherence in the processor caches. Numerous tests, performed on different computers, demonstrate the reliability of this approach. Moreover, the above example as well as other tests, performed on the AMD Opteron processor, which does not have hardware protection against performance degra-

dation due to incoherence in the processor caches, demonstrate stable speedup with the increasing thread number. This approach could be applied to the multithreaded parallelization of the internal force vector evaluation procedure, but we wanted to compare the efficiency of two different approaches to justify the reliability of the more complicated second method.

ACKNOWLEDGMENT

The authors are deeply grateful to IT company SCAD Soft for the financial support of this research and for providing a collection of problems.

REFERENCES

[1] O. Schenk, K. Gartner, "Two-level dynamic scheduling in PARDISO: Improved scalability on shared memory multiprocessing systems," *Parallel Computing,* vol. 28, 2002, pp. 187–197, https://doi.org/10.1016/S0167-8191(01)00135-1

[2] Intel Math Kernel Library Reference Manual. URL: https://software.intel.com/en-us/mkl-developer-reference-c-intel-mkl-pardiso-parallel-direct-sparse-solver-interface (Last access: 17.04.2018).

[3] S. Yu. Fialko, "Parallel direct solver for solving systems of linear equations resulting from finite element method on multi-core desktops and workstations", *Computers and Mathematics with Applications,* 70, 2015, pp. 2968–2987, doi:10.1016/j.camwa.2015.10.009

[4] S. Fialko, "PARFES: A method for solving finite element linear equations on multi-core computers", *Advances in Engineering Software,* 40, (12), 2010, pp. 1256–1265. https://doi.org/10.1016/j.advengsoft.2010.09.002

[5] K. J. Bathe, *Finite Element Procedures,* New Jersey: Prentice Hall; 1996.

[6] S. Yu. Fialko, "Quadrilateral finite element for analysis of reinforced concrete floor slabs and foundation plates", *Applied Mechanics and Materials,* 725–726, 2015, pp. 820 – 835, doi: 10.4028/www.scientific.net/AMM.725-726.

[7] S. Yu. Fialko, V. S. Karpilowskyi, "Triangular and quadrilateral fat shell fnite elements for nonlinear analysis of thin-walled reinforced concrete structures in SCAD software," *In: Petraszkiewicz and Witkowski (eds). Shell Structures: Theory and Applications,* V. 4., Taylor and Francis Group, London, 2018, pp. 367–370.

[8] S. Yu. Fialko, V. S. Karpilowskyi, "Block subspace projection preconditioned conjugate gradient method for structural modal analysis", *in Proceedings of the Federated Conference on Computer Science and Information Systems,* ISSN 2300-5963 ACSIS, Vol. 11, pp. 497–506. DOI: 10.15439/2017F64 .

[9] S. Yu. Fialko, *Application of finite element method to analysis of strength and bearing capacity of thin-walled concrete structures, taking into account the physical nonlinearity,* Moscow: Publishing House SCAD SOFT, Publishing House ASV; 2018 (Russian).

[10] A. George, J. Liu, E. Ng, Computer Solution of Sparse Linear Systems, 1994. URL: http://web.engr.illinois.edu/~heath/courses/cs598mh/george_liu.pdf

[11] Interlocked variable access. URL: https://msdn.microsoft.com/en-us/library/windows/desktop/ms684122(v=vs.85).aspx

[12] M. Olczyk, "The procedure of parallel assembling of stiffness matrix in FE analysis for applying to the solution of nonlinear algebraic equation systems", *Master's degree work, Cracow University of Technology,* Cracow, Polish, 2017.

[13] D. Th. Nguyen, *Parallel-Vector Equation Solvers for Finite Element Engineering Applications,* Springer Science+Business Media, LLC: New Yourk; 2002. DOI 10.1007/978-1-4615-1337-7.

[14] Yu.V. Khalevitsky, N.V. Burmasheva, A.V. Konovalov, "An approach to the parallel assembly of the stiffness matrix in elastoplastic Problems", *Mechanics, Resource and Diagnostics of Materials and Structures (MRDMS-2016), AIP Conf. Proc. 1785, pp. 040023-1–040023-4; Published by AIP Publishing. 978-0-7354-1447-1/$30.00* doi: 10.1063/1.4967080.

[15] M. N. De Rezendea, J. B. de Paiva, "A parallel algorithm for stiffness matrix assembling in a shared memory environment", *Computers & Structures,* 76, (5, 15), 2000, pp. 593–602. https://doi.org/10.1016/S0045-7949(99)00181-9.

[16] D. Goudin, J. Roman, *A scalable parallel assembly of irregular meshes based on a block distribution for a parallel direct solver*, In: Applied Parallel Computing, New paradigms for HPC in industry and academia, *5th International Workshop, PARA 2000, Bergen, Norway, June 2000, Proceedings*, Springer, Lecture Notes in Computer Science, V. 1947, pp. 113 – 116. URL: https://link.springer.com/chapter/10.1007/3-540-70734-4_15 (Last access: 13.07.2018)

[17] C. Cecka, A. Lew, E. Darve, "Introduction to Assembly of Finite Element Methods on Graphics Processors", *IOP Conf. Series: Materials Science and Engineering*, 10, 012009, 2010, pp. 1 – 10. doi:10.1088/1757-899X/10/1/012009.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 319–327
DOI: 10.15439/2018F28
ISSN 2300-5963 ACSIS, Vol. 15

On the energy consumption of Load/Store AVX instructions

Thomas Jakobs, Gudula Rünger
Chemnitz University of Technology,
Department of Computer Science
09111 Chemnitz, Germany
E-mail: [thomas.jakobs, ruenger]@cs.tu-chemnitz.de

***Abstract*—The energy efficiency of program executions is an active research field in recent years and the influence of different programming styles on the energy consumption is part of the research effort. In this article, we concentrate on SIMD programming and study the effect of vectorization on performance as well as on power and energy consumption. Especially, SIMD programs using AVX instructions are considered and the focus is on the AVX load and store instruction set. Several semantically similar but different load and store instructions are selected and are used to build different program versions of for the same algorithm. As example application, the Gaussian elimination has been chosen due to its interesting feature of using arrays of varying length in each factorization step. Five different SIMD program versions of the Gaussian elimination have been implemented, each of which uses different load and store instructions. Performance, power, and energy measurements for all program versions are provided for the Intel Sandy Bridge, Haswell and Skylake architectures and the results are discussed and analyzed.**

***Index Terms*—Energy consumption, power consumption, AVX instruction set, Gaussian elimination.**

I. Introduction

In addition to performance optimization of codes for numerical algorithms, there is a growing need to optimize the power and energy consumption of such programs as well. Reasons are well-known and include aspects, such as budgeting, battery life, cooling capacity, or physical boundaries. A possibility to reduce the power and energy consumption of the execution of a program is the choice of energy-saving architectures and or architecture components, which have been developed by the hardware manufacturers for this purpose. An example for such an architecture components are SIMD or vectorization units supplied by recent processors. A CPU based SIMD execution unit calculates multiple elements in special registers simultaneously with one instruction in contrast to processing them sequentially. For the same computation, a vectorized program activates less transistors than a sequential program and, hence, vectorization provides a potential for performance and, power and energy optimization.

The exploitation of SIMD units for executing numerical algorithms requires to provide a suitable program which contains so-called vector operations that cause the architecture to exploit the vector units. One possibility to do this is to use the Intel AVX instruction set, which can be used on recent Intel architectures, such as the Intel processors Sandy Bridge, Haswell or Skylake. The design of such a vectorized program using AVX instruction includes several decisions when transforming a sequential program or algorithm into a vectorized program version. The decisions include the selection of program parts to be coded as vector operations but also the specific choice of AVX instructions to be used for the coding. Both, the strategy to include SIMD parallelization into a program as well as the choice of instructions, can have an effect on the performance as well as the power and energy consumption. This article concentrates on the AVX load and store instruction set and investigates the effect on the performance when different but semantically similar instructions are chosen.

The investigations of this article concentrate on different load and store instructions provided by the AVX instruction set, such as aligned, unaligned, streaming or masked load and store instructions. Using these alternatives for load and store, different SIMD program versions for the Gaussian elimination are built. The experimental investigation of these program versions exhibits interesting differences in the performance results, such as the different performance of aligned and unaligned load and store instructions. In detail, this article makes the following contributions:

- Several program versions for the Gaussian elimination are implemented with different AVX instructions.
- The performance, power and energy of the the five program versions with different AVX instructions have been investigated on three processor architectures.
- A detailed discussion and comparison of differences in performance, energy and power consumption of the implemented program versions is given.

The rest of this article is structured as follows: In Sec. II, we introduce AVX instructions, the Gaussian elimination, and the vectorized program versions of the Gaussian elimination. Section III introduces the execution environment and hardware architecture. Section IV presents the performance properties of the program versions. In Sec. V, we discuss the influence of different AVX instruction on the energy efficiency of a program. Section VI shows related research and Sec. VII concludes.

II. SIMD IMPLEMENTATION OF THE GAUSSIAN ELIMINATION

The vectorized implementation of the Gaussian elimination leaves room for different choices of load/store AVX instructions. This section introduces the AVX instructions, their expected influence, the Gaussian elimination, and the vectorized program versions.

A. Programming with AVX

Many of today's CPUs have vector or SIMD instruction sets, which support parallel executions by applying the same operation simultaneously to two, four, or more pieces of data. Two popular SIMD instruction sets are the *Streaming SIMD Extensions* (SSE) and the *Advanced Vector Extensions* (AVX), both implemented into various AMD® and Intel® processors. SSE and AVX provide instructions to load, modify, and store 128-bit (SSE) or 256-bit (AVX) vectors containing multiple elements, where the number of elements is specified by their particular size. In principle, program executions can reach a speedup equal to the number of data elements per vector, in practice several factors limit the optimal speedup. The performance properties and limitations for vectorization have been demonstrated for multiple examples, e.g. in [2]. Only few articles cover vectorization in the context of energy efficiency. Examples from Cebrián et. al. [3] or Lorenz et. al. [4] demonstrate a potential to increase the energy efficiency by the application of vectorization.

The SSE and AVX instructions can be used either as assembler instructions or embedded into C-style intrinsic functions. It is recommended to use intrinsic functions instead of assembler instructions, since it enables the compiler to apply further optimizations, such as dead code analysis. The intrinsic functions have the following format:

```
                <type>
_mm<size>_<operation>_<type_suffix>(
          <type> param1,..)
```

- `<type>` can either be a standard C-type (e.g. `float`) or a special vector type (e.g. `__m256` for 8 single-precision floating-point values) depending on the actual function purpose and definition.
- `<size>` denotes the number of bits used for the instruction, e.g. `256` for AVX.
- `<operation>` expresses the implemented operation, e.g. `add` for an addition.
- `<type_suffix>` denotes the type of data to operate on, e.g. `ps` for packed single precision.
- The number and type of parameters varies dependent on the instruction.

Table 1 lists the specific intrinsic functions used in this article with their semantic, and the values for latency and throughput on different processor architectures given in [1].

B. Alternative load/store instructions

The different semantic of the intrinsic functions can be investigated with regard to the energy efficiency. For this purpose the intrinsic functions in Tab. 1 are evaluated in the context of expected differences in measurement results. Such differences can arise from their difference in latency and throughput, and due to their requirement on memory alignment, i.e. the memory address being divisible by 32 byte. The following instructions are amenable to demonstrate such differences:

- **Unaligned** load/store: Load and store operations accessing unaligned memory are applicable for any memory position to load/store any consecutively stored elements. The cache line size is a multiple of the AVX register size of 256 bit, and thus the access of unaligned memory positions may contain a cache line border, resulting in a possible performance loss [5].
- **Aligned** load/store: Vectors aligned at 256 bit reside in the same cache line. Thus, the access is expected to lead to a higher performance of the program, compared to unaligned instructions. However, an aligned memory access has to be ensured by the programmer using methods such as loop peeling, special allocators, and/or alternative strategies.
- **Streaming** store: Streaming stores are special, aligned store instructions that bypass the caches when storing data. In detail, they are implemented using a "non-temporal hint" to indicate that no intermediate copy should be created in cache. Thus, streaming stores provide the possibility to issue a Write Through operation at the programming level.
- **Masked** load/store: Vector instructions usually use all elements for their calculations leading to a strict SIMD implementation and execution. A possibility to use only parts of a vector register is provided by masked instructions. Masked load/store instructions have an additional mask parameter which specifies the elements to be loaded/stored. The elements to be omitted, i.e. not to be loaded/stored, are specified by 0-bits in the mask parameter. Values omitted by loads are assigned zeros in the vector, whereas values omitted by stores are skipped while writing to memory. Values to be loaded/stored are identified by 1-bits in the mask parameter and treated accordingly.

```
1 for k = 0 < N-1; k+=1
2   //exchange rows with pivot element
3   for i = k+1 < N; i+=1
4     L[i*N+k] = A[i*N+k] / A[k*N+k]
5     for j = k+1 < N; j+=1
6       A[i*N+j] = A[i*N+j] - A[k*N+j] * L[i*N+k]
7     b[i] = b[i] - b[k] * L[i*N+k]
8 //Backward substitution
```

Listing 1: Algorithm of a Gaussian elimination adapted from [6] of which the vectorized program versions are derived.

C. The Gaussian elimination

The Gaussian elimination is an important kernel in scientific applications for solving linear equations. For the Gaussian elimination, a set of N linear equations with N

Intrinsic function	Instruction semantic	HSW Lat	HSW Tp	SKL Lat	SKL Tp
`_mm256_load_ps(*mem)`	Loads 8 float values from an aligned memory position `mem` into a vector variable.	1	0.5	1	0.25
`_mm256_loadu_ps(*mem)`	Loads 8 float values from an unaligned memory position `mem` into a vector variable.	1	0.5	1	0.25
`_mm256_maskload_ps( *mem,mask)`	Loads a specified selection (`mask`) of values from a memory position `mem` into a vector variable. Omitted values are 0.	8	2	11	1
`_mm256_broadcast_ss(*mem)`	Loads one float value (`mem`) into all elements of a vector variable.	-	-	-	-
`_mm256_loadu_si256(*mem)`	Loads 256 bits of integer values from an unaligned memory position `mem` into a vector variable.	1	0.25	1	0.25
`_mm256_store_ps(*mem,a)`	Stores the elements of a vector variable (`a`) into an aligned memory position `mem`.	1	0.5	1	0.25
`_mm256_storeu_ps(*mem,a)`	Stores the elements of a vector variable (`a`) into an unaligned memory position `mem`.	1	0.5	1	0.25
`_mm256_maskstore_ps( *mem,mask,a)`	Stores a specified selection (`mask`) of values from a vector variable `a` into a memory position `mem`. Omitted values are skipped.	-	2	-	1
`_mm256_stream_ps(*mem,a)`	Stores the elements of a vector variable (`a`) into an aligned memory position `mem` using a non-temporal hint.	-	1	-	1
`_mm256_mul_ps(a,b)`	Multiplies the elements of two vector variables (`a` and `b`) with each other.	5	0.5	4	0.5
`_mm256_sub_ps(a,b)`	Subtracts the elements of one vector variable (`b`) from another vector variable (`a`).	3	1	4	0.5

Table 1: AVX intrinsic functions used in this article with their respective values for Latency (Lat.) and Throughput (Tp.) for the Haswell (HSW) and Skylake (SKL) architectures from [1].

unknowns x_k and their respective coefficients a_{ik} and right hand side b_i, where $1 \leq i,k \leq N$ is given. The equation and solution for $Ax = b$, with $A \in \mathbb{R}^{N\times N}$ and $x, b \in \mathbb{R}^N$, has to be solved.

The Gaussian elimination can be divided into two phases: A forward elimination and a backward substitution. In the forward elimination the matrix A is transformed into an upper triangular form, such that $Ux = b'$ holds, where U is the matrix in upper triangular from.

The forward elimination is depicted in Listing 1 as pseudocode. The *k-loop* in Line 1 executes the steps of the forward elimination and iterates along the diagonal elements a_{kk} of matrix A. Each step starts with a pivot search in which the maximum value below (a_{ik}) the diagonal element a_{kk} is searched, and a row exchange of the *current* row $a_{k\_}$ with the row of the pivot $a_{piv\_}$ is done. In each step the *i-loop* in Line 3 calculates for each row $a_{i\_}$, with $i > k$, the elimination factor $l_{ik} = \frac{a_{ik}}{a_{kk}}$ (Line 4), which is stored in a separate matrix $L \in \mathbb{R}^{N\times N}$. Using the elimination factor l_{ik} all elements of row $a_{i\_}$ are calculated by $a_{ij} = a_{ij} - a_{kj} \cdot l_{ik}$, where $k+1 \leq j \leq N$ denotes the column that is iterated by the loop in Line 5. Afterwards, for each row the element b_i of the result vector b is calculated by $b_i = b_i - b_k \cdot l_{ik}$. After $N-1$ steps the former matrix A is transformed to upper triangular form U. The resulting matrix U is stored in the same memory location as the former matrix A, in which the lower triangular elements are *assumed* to be zero.

The second phase is the backward substitution in which the values for vector x are calculated. The elements of vector x are calculated in the order of $x_N, x_{N-1}, \ldots x_1$ according to $x_k = \frac{1}{a_{kk}} \left(b_k - \sum_{j=k+1}^{N} a_{kj} \cdot x_j \right)$.

The algorithm of Listing 1 is not optimized to preserve the ratio of memory- and computation-instructions.

```
1  for k = 0 < N; k+=1
2    //exchange rows with pivot element
3    for i = k + 1 < N; i+=1
4      L[k*N+i] = A[i*N+k] / A[k*N+k];
5      __m256 lv = _mm256_broadcast_ss(&L[k*N+i]);
6      for j = k + 1 < N - 8; j+=8
7        __m256 ak = _mm256_loadu_ps(&A[k*N+j]);
8        __m256 ai = _mm256_loadu_ps(&A[i*N+j]);
9        ak = _mm256_mul_ps(ak,lv);
10       ai = _mm256_sub_ps(ai,ak);
11       _mm256_storeu_ps(&A[i*N+j],ai);
12     if(j < N)
13       int loadmask = {0,0,0,0,0,0,0,0,
           -1,-1,-1,-1,-1,-1,-1,-1};
14       __m256i mask = _mm256_loadu_si256(
           (__m256i*)&loadmask[N-j]);
15       j = N - 8;
16       __m256 ak = _mm256_loadu_ps(&A[k*N+j]);
17       __m256 ai = _mm256_loadu_ps(&A[i*N+j]);
18       ak = _mm256_mul_ps(ak,lv);
19       ai = _mm256_sub_ps(ai,ak);
20       _mm256_maskstore_ps(&A[i*N+j],mask,ai);
21   //calculate b similarly
22 //Backward substitution
```

Listing 2: Vectorized implementation of the Gaussian elimination from Listing 1, which is the starting point for the subsequent program versions.

D. SIMD implementation versions of the Gaussian elimination

A vectorized implementation of the Gaussian elimination from Listing 1 is presented in Listing 2. The implementation uses the intrinsic functions introduced in Tab. 1 to calculate multiple elements of the row (of the *j-loop*) simultaneously. Lines 1 to 4 of Listing 1 remain unchanged. For the vectorized calculation the resulting value of l_{ik} is copied into all elements of a vector variable `lv` in Line 4. The *j-loop* in Line 6 is unrolled eight times to calculate

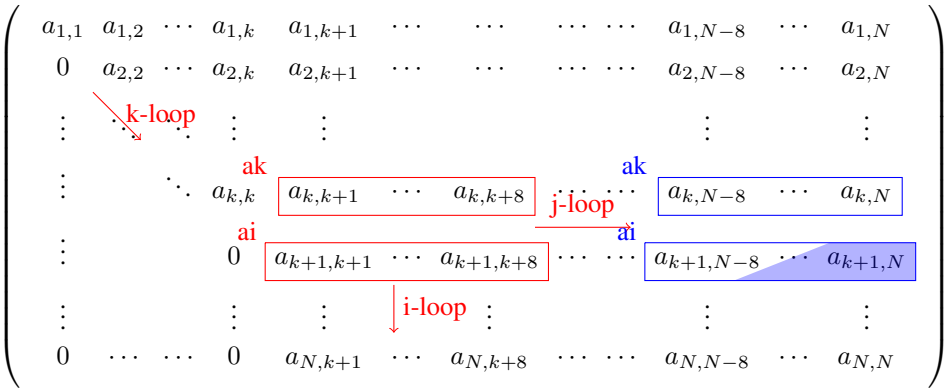


Figure 1: Matrix A in step k, with $i = k + 1$ and $j = k + 1$. The vector variables and the iteration directions according to Listing 2 are depicted in red. The remainder is calculated using the vectors highlighted as blue rectangles for which only the new elements (blue filled) are stored back to the array.

eight single-precision floating-point values simultaneously as demonstrated in Fig. 1. For the use with vector instructions the elements are loaded into vector variables (`ak` and `ai` in Lines 7 and 8, red rectangles in Fig. 1). The vector variables are used to calculate the new values for $a_{i,j}, a_{i,j+1}, \ldots, a_{i,j+7}$ (see red `ai` in Fig. 1) that are written back into the array in Line 11.

Since the number of iterations for the *j-loop* $(N-(k+1))$ is not guaranteed to be divisible by eight a special case has to be implemented. The remaining elements are handled in the block after Line 12 using the masked instructions. Usually, masked instructions are used to load the last elements of the row and values which would not be part of the row are omitted. We choose a different strategy to avoid two costly `maskload` instructions. For this strategy the last eight elements of each row are loaded, regardless of how many are actually needed (see blue rectangle in Fig. 1). After calculation, when storing the results, a mask is applied to write only those elements that are part of the remainder, discarding the twice used elements (blue fill of `ai` in Fig. 1). The mask is created by using an array (`loadmask` in Line 13) of 256 0-bits followed by 256 1-bits and loading the mask in Line 14 which contains the correct amount of 1-bits. The calculation of the $N-(k+1)$ elements of vector b is done similarly.

Different program versions of the Gaussian elimination are built with different load/store instructions. In the following we describe the five program versions:

- The **storeu** program version is the *starting point* implementation using unaligned loads and stores. The *storeu* program version is presented in Listing 2.
- The **store** program version uses aligned loads (`_mm256_load_ps`) and stores (`_mm256_store_ps`). Hence, the *j-loop* in Line 6 does not start at $j = k + 1$ but at the beginning of the aligned block of memory in which $k + 1$ resides. Thus, additional elements are calculated, but a peeling loop is avoided. The aligned load instructions are implemented in Line 8, and the aligned store instructions in Line 11 of Listing 2.
- The **stream** program version uses aligned loads identically to the *store* program version. However, the `store` instruction in Line 11 is replaced with a streaming store (`_mm256_stream_ps`) instruction which bypasses the cache while writing.
- The **maskload** program version replaces all load instructions with masked load (`_mm256_maskload_ps`) instructions. Explicitly, the `maskload` instruction is implemented in Lines 7, 8, 16 and 17 of Listing 2. The masks of these may be either all 1-bits or the correct mask for the remainder. The *maskload* program version illustrates the difference between masked loads and normal loads.
- The **SeqRem** program version replaces the vectorized remainder (Lines 12 to 20 in Listing 2) with a sequential loop that processes each remaining element sequentially as in Listing 1. A sequential remainder loop may be more efficient due to the expected higher cost of masked instructions.

The program versions cover a set of selectable instructions to implement a vectorized Gaussian elimination. The program versions differ only in the modifications shown.

III. Experimental Evaluation

The measurements for this article are conducted in the environment described in this section. Each measurement is conducted ten times and the presented values are averaged. For the measurements we use a matrix A that has $10\,000 \times 10\,000$ elements.

A. Execution Environment

For the measurements of this article we use three Intel® Core i7 processors with a similar specification but different architecture families. The three processors are: A Core i7-2600 of the Sandy Bridge architecture, a Core i7-4770K of the Haswell architecture, and a Core i7-6700 of the

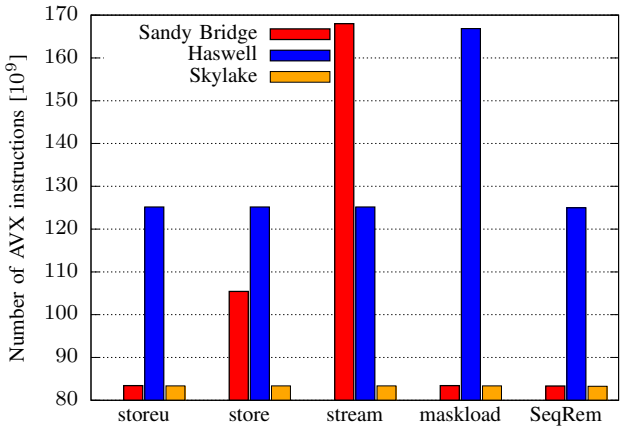


Figure 2: Number of AVX instructions actually executed in the CPU by the execution of the five program versions on the three processors.

Skylake architecture. All three processors allow processor frequency scaling with the `cpu-freq` tool. For the Haswell and Skylake processor scaling is enabled between 0.8GHz and 3.5GHz and for the Sandy Bridge processor the scaling is possible between 1.6GHz and 3.7GHz.

We measure the energy consumption of program executions using a simple interface to read the on chip energy values provided by the Intel® RAPL interface. Additionally, we measure other performance counters using the PAPI library version 5.5. We compiled the program versions using the Intel® C++ Compiler (`icc` version 17.0.0 [gcc version 4.9.0]) with the additional compiler flags `-O3` and `-restrict`. Additionally, we applied the compiler flags `-mavx` for Sandy Bridge architecture and `-march=core-avx2` for the other architectures. The usage of the `-march-avx2` flag implies the usage of FMA instructions rather than `sub` and `mul` by the compiler.

B. Issued AVX instructions

A variation of the number of issued AVX instructions occurs when executing the five program versions on the three processors. The number of issued AVX instructions can be measured during program execution and reflects the number of 256-bit instructions executed.

The number of issued AVX instructions is depicted in Fig. 2 for the different program versions from Sec. II and the three architectures. It is notable that most of the program executions have a nearly identical number of issued instructions. The expected behavior would be all issued instruction counts being nearly identical, since the number of instructions should be predefined by the number of assembler instructions that are generated by the intrinsic functions.

There are multiple exceptions to the expected behavior. For all program versions, the number of instructions issued on the Haswell architecture is about 50% higher, than for the other two architectures. Additionally, some program versions generate an untypical high number of executed

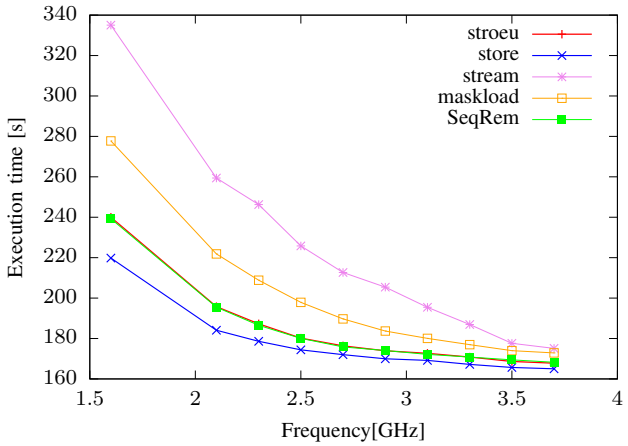


Figure 3: Execution time of the Gaussian elimination versions from Sec. II on Sandy Bridge architecture depending on CPU frequency.

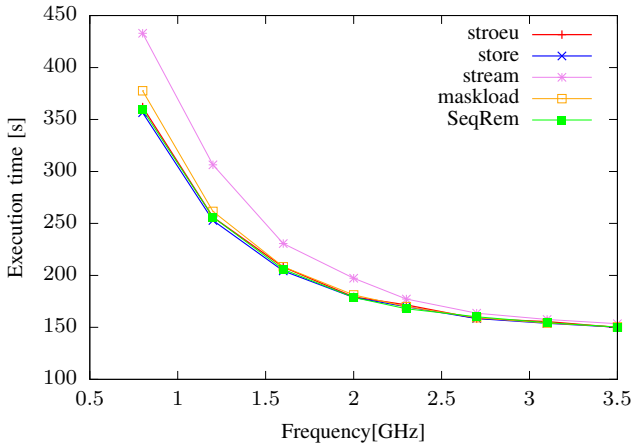


Figure 4: Execution time of the Gaussian elimination versions from Sec. II on Haswell architecture depending on CPU frequency.

instructions. These exceptions arise at the execution of the program version with a higher amount of `maskload` instructions for the Haswell architecture, the use of the `stream` instruction on the Sandy Bridge architecture and slightly more with the use of the aligned load and store operations for the Sandy Bridge architecture.

IV. Performance Evaluation of the Program Versions

In this section, we present the measurements and discuss the differences in execution time for the different program versions. The five program versions of Sec. II are executed on the three processors described in Sec. III.

A. Evaluating execution time

The Fig. 3, 4 and 5 display the execution times of the different program versions in dependence to the processor frequency. As expected, a higher processor frequency leads in all diagrams to a shorter execution time and the qualitative

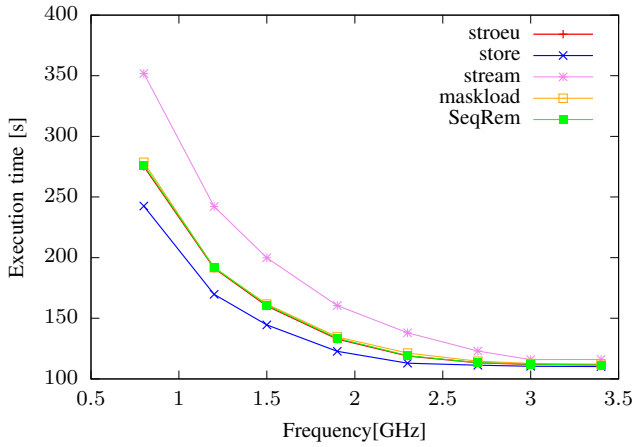


Figure 5: Execution time of the Gaussian elimination versions from Sec. II on Skylake architecture depending on CPU frequency.

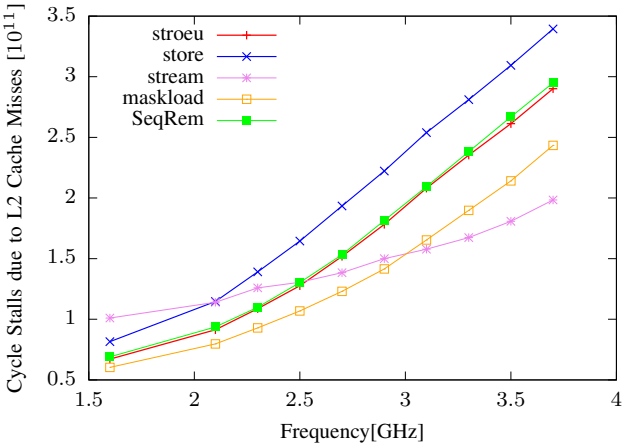


Figure 6: Cycle Stalls due to level 2 cache misses of the Gaussian elimination versions from Sec. II on Sandy Bridge architecture depending on CPU frequency.

behavior is similar. However, the effects of different AVX instructions on performance are thus revealed by the relation of the execution time of the program versions to each other. The program version using the `stream` instruction has the longest execution time for all architectures. This might be caused by the Write Through strategy that is applied with the `stream` instruction which bypasses the cache when writing data. Thus, the processor has to wait for completion of the write operation (*write-wait*) before another `stream` instruction can be executed.

The *write-waits* are a significant problem occurring for the Gaussian elimination, which has a high number of loads and stores with only little computation per element in between. In other algorithms, that implement a lower number of streaming stores the *write-waits* may be hidden by other computations, loads or Write Back stores.

The Sandy Bridge architecture was the first architecture to support AVX instructions and many improvements to hardware for AVX have been made since. This is also reflected in the example of the program version executing a higher number of `maskload` instructions. For the Sandy Bridge architecture the *maskload* program version takes a higher execution time than the remaining three program versions. For the Haswell architecture the execution time is only slightly higher and for the Skylake architecture there is no difference in execution time between a `maskload` (with a full true-mask) and a `loadu` instruction.

A comparison between the *storeu* program version and the *SeqRem* program version demonstrates the influence of a vectorized remainder against a sequential remainder loop. The measurement results of the *storeu* and the *SeqRem* program versions are nearly identical ($< 0.5\%$). In combination with the results of the *maskload* program version, this leads to the conclusion, that the use of masked instructions has a worse performance than regular load/store operations, but are at least as performant as a sequential program execution.

For the Sandy Bridge architecture the aligned `load` and `store` instructions exhibit a better performance than the unaligned ones. The difference between aligned and unaligned instructions is eliminated for the Haswell architecture. For the Skylake architecture the performance enhancement of aligned instructions is again visible. Presumably, the architectural changes between Sandy Bridge and Haswell architecture improved the unaligned instructions, whereas the architectural improvement from Haswell to Skylake architecture improved the aligned instructions.

B. Influence of Cache-Waits on execution time

The memory properties of a program can support the classification of its performance properties. One way to get information about the memory properties is to take a look at the cache usage of the program versions.

Figure 6 displays the number of CPU cycles stalled due to waits for pending operations on level 2 cache for the Sandy Bridge architecture. For most of the program versions from Sec. II the relation of the curves is directly inverse to their execution time of Fig. 3. This behavior is expected due to the limitation of memory bandwidth that gets visible more clearly if the program executes the implemented calculations faster.

The *stream* program version produces a lower rise in cycle stalls in dependence to the processor frequency than the other four program versions. The lower rise in the diagram can be explained with the operation inside the `stream` instruction: The Write Trough operation. The Write Trough operation does not write data into the cache but directly to main memory. Thus, waiting for a write operation is not counted as a wait for any cache and thus does not get counted for level 2 cache stalls. When regarding other resource stalls, i.e. general stalls, the results for the *stream* program version are much higher than for the other program versions. This observation reinforces the previous statement of the worse execution time of the `stream` instruction resulting from *write-waits*. The difference is presented for the Sandy Bridge architecture in Fig. 6 but is also observ-

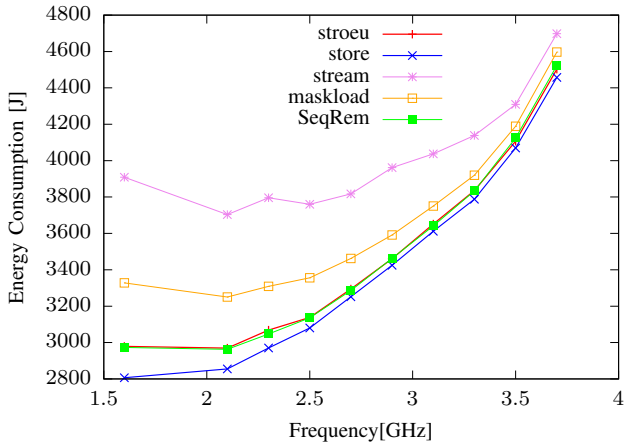


Figure 7: Energy consumption of the Gaussian elimination versions from Sec. II on Sandy Bridge architecture depending on CPU frequency.

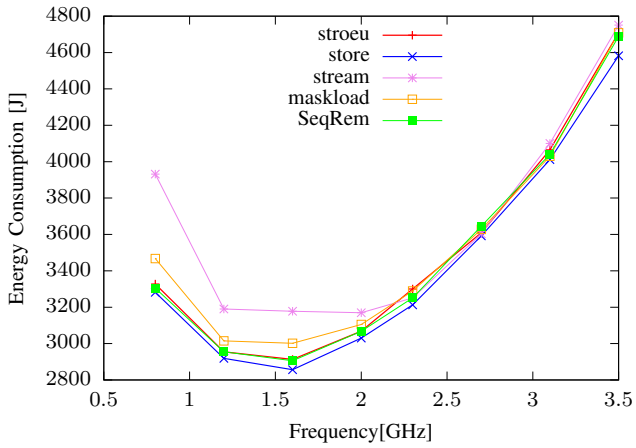


Figure 8: Energy consumption of the Gaussian elimination versions from Sec. II on Haswell architecture depending on CPU frequency.

able for the Haswell and Skylake architectures. Since the investigations for this article do not exploit parallelism, the level 3 cache displays a similar behavior.

V. Examining the energy efficiency

The main question of this article is the behavior of the energy efficiency of different AVX instructions. This section discusses the energy and power consumption of the five program versions from Sec. II. The results are presented as *Energy to Solution* and *Power to Solution* from which other metrics, such as Energy-Delay-Product, can be derived.

A. Evaluation of the energy consumption

One metric to discuss the energy efficiency of program execution is the energy consumed by the processor during the execution of the program version. The Fig. 7, 8 and 9 present the energy consumption of the program versions from Sec. II depending on the processor frequency. Similar

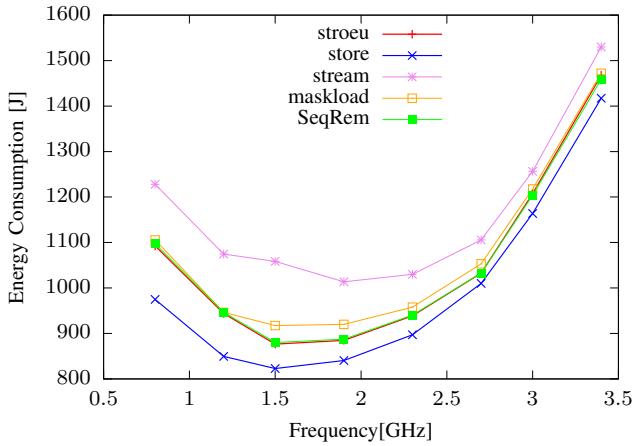


Figure 9: Energy consumption of the Gaussian elimination versions from Sec. II on Skylake architecture depending on CPU frequency.

to the results of the performance discussion, the *stream* program version has the highest (worst) energy consumption for all architectures. The *maskload* program version consumes more energy on the Sandy Bridge and Haswell architecture. Executing the remainder with a sequential remainder loop does not change the energy consumption against a vectorized remainder. In contrast to the performance discussion, the program version using aligned stores consumes less energy for all three architectures.

The lowest energy consumption is achieved with a frequency between 1.5GHz and 2.0GHz for all program versions and architectures. The dependency of the energy consumption on the processor frequency produces a U-Shape with deviations. Specifically, the U-Shape of the *stream* program version creates a U-Shape with a broader base, i.e. a flatter U-Shape. A similar behavior is shown by the *maskload* program version on the Haswell and Skylake architectures.

The different program versions produce their highest difference in energy consumption for the lower frequencies. For the higher frequencies the execution of the program versions is mostly affected by the memory transfer time, as already discussed in Sec. IV. The dependence on memory bandwidth limits the capabilities of vectorization and thus makes idle or waiting time a significant fraction of the program execution.

B. Differences in power consumption

In many cases the implementation with different vector instructions changes energy consumption in the same way as it changes the execution time of the program execution. Some exceptions can be found by regarding the power consumption of the program versions, where $power = \frac{energy}{time}$.

The power consumption is calculated from the averaged measurement results for execution time and energy consumption of each execution. The power consumption of a program execution strongly depends on the processor

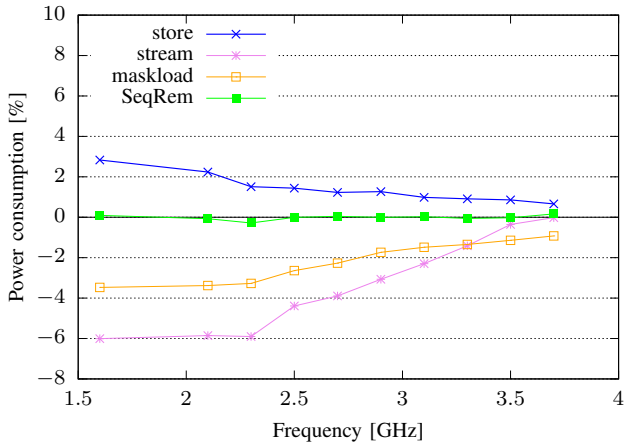


Figure 10: Difference of power consumption of the Gaussian elimination versions from Sec. II to the *storeu* program version in percent on Sandy Bridge architecture depending on CPU frequency.

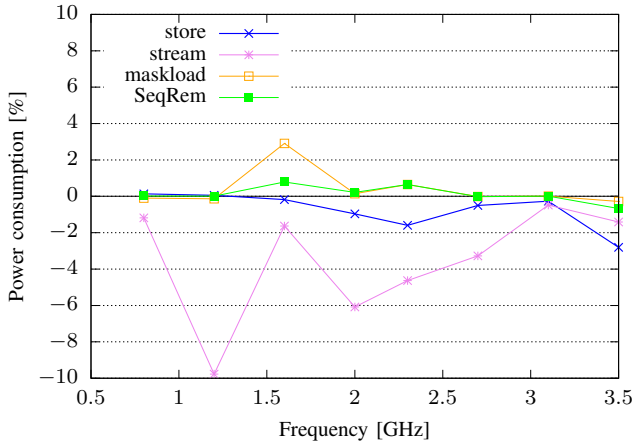


Figure 11: Difference of power consumption of the Gaussian elimination versions from Sec. II to the *storeu* program version in percent on Haswell architecture depending on CPU frequency.

frequency [7]. Thus, to discuss the differences between the five program versions the Fig. 10, 11 and 12 display the power consumption as difference from the *storeu* program version in percent. In general, a processor consumes more power when more transistors are active, i.e. the processor gets hotter, which often comes with a shorter execution time and less energy consumption.

Overall, the five program versions produce less differences for higher frequencies, which demonstrates the dependency on memory bandwidth rather than computing capabilities. Additionally, the sequential remainder loop program version has nearly no difference ($< 0.5\%$) to the *storeu* program version with a vectorized remainder.

The *stream* and *maskload* program versions produce a local maximum or peak behavior for the processor frequencies around 1.5GHz for the Haswell and Skylake architectures in Fig. 11 and 12. The peak results from the flatter U-

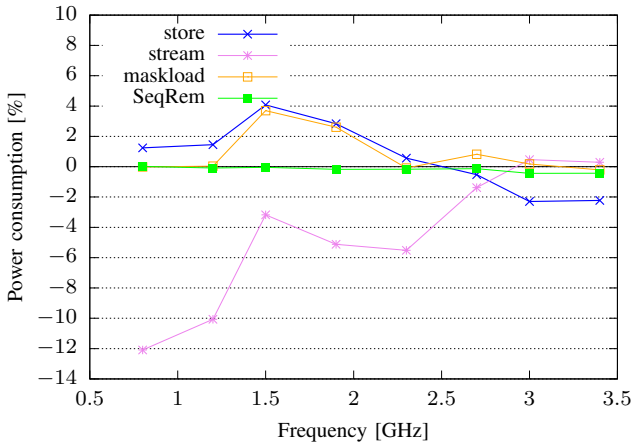


Figure 12: Difference of power consumption of the Gaussian elimination versions from Sec. II to the *storeu* program version in percent on Skylake architecture depending on CPU frequency.

Shape that these two program versions display in energy consumption, where the *storeu* program version (0%-Line) does not follow this behavior.

In Fig. 10 the *maskload* program version generates a lower power consumption for the Sandy Bridge architecture than the *storeu* program version. For the other two architectures the relation is directly inverse or no difference is shown at all. The reason for this is the higher energy consumption of the *maskload* program version on the Haswell and Skylake architecture for which the execution time is identical to the *storeu* program version. For the execution of the *maskload* program version on the Sandy Bridge architecture the execution time is higher (+16% for 1.6GHz) as well as the energy consumption is higher (+12% for 1.6GHz). The difference between these two increased values leads to a lower power consumption.

The program version implemented with aligned stores produces a lower energy consumption than the *storeu* program version for all architectures. However, the execution time is nearly identical on the Haswell architecture and lower on the other two architectures. This leads to a reduced power consumption of the *store* program version on the Haswell architecture. Additionally, for the Skylake architecture the *store* program version produces a constant difference in energy consumption and execution time to the *storeu* program version for higher frequencies. However, the point at which the constancy is occurring is at 2.0GHz for the energy consumption and at 2.7GHz for the execution time. This leads to an inverse relation of the *store* program version in Fig. 12 above 2.7GHz, for which the aligned stores are more power and energy efficient.

VI. Related Work

The energy and power efficiency is subject to different research fields. Especially, in the field of multi-threaded, parallel and distributed computing [3], [8], [9]. In our work we isolate the effects of vectorization on energy efficiency

and leave other techniques of parallel execution applicable on top.

Lien et al. [10] investigate the energy efficiency of multi-threaded vectorized programs. They use three different algorithms for their work: FFTW, Matrix-Multiplication, and blackscholes. They show that vectorization increases energy efficiency, even more with additional multi-threading. For our investigations we isolate the use of vectorization to reason about the impact of different instructions used.

Caminal et al. present an energy efficiency study of the ParVec Suite based on different vectorization strategies [11]. They demonstrate the need for easy user guided vectorization to reduce the energy consumption of program executions. Their main focus is on the results of different user guided vectorization techniques such as OmpSs and Mercurium. The focus of our work is the investigation of differences of vector instructions to create additional knowledge for the use in such user guided vectorization systems.

In [2] Kim et al. describe modifications of source code to increase the performance properties of vectorized programs. They specifically emphasize that the use of continuously stored data elements is one of the key factors for efficient vectorized programs. For this they extensively discuss the use of Struct-of-Arrays instead of Array-of-Structs store order. We considered their findings for creating our program versions and focus on the influence of different instructions.

Hofmann et al. examine the influence on performance of vector instructions with the RabbitCT benchmark in [12]. They demonstrate that the choice of instructions has an influence on the performance of the program execution. Their work examines the additional instructions introduced by the AVX2 instruction set and Intel Xeon Phi instructions. We extend this research by examining the performance and energy efficiency for regular load/store instructions.

VII. Conclusion

In this article, the influence of different load and store AVX instructions with different program versions of a Gaussian elimination are investigated. The investigations demonstrate that the choice of instructions influences the execution time, energy and power consumption. The number of issued AVX instructions may be different, depending on the processor architecture and set of instructions implemented. However, no influence of a different number of issued AVX instructions on the performance or energy properties of the program execution can be derived.

The processor development influences the properties of different instructions, such as for the masked instructions which are improved in the newer processors. The use of streaming store instructions produces the worst behavior due to the memory intensiveness of the Gaussian elimination. Additionally, remainder loops can be vectorized without performance or energy efficiency being negatively influenced. Aligned load and store instructions produce the best results in performance and energy consumption even if additional elements are computed.

A next step in our research will be to identify the root cause of the different number of instructions issued presented in Sec. III. Additionally, the influence of data size and cache usage can be examined with cache optimizations of the basic program versions and comparisons with established library implementations, such as in BLAS or LAPACK.

References

[1] Intel Corporation, "Intel Intrinsics Guide," Apr. 2018. [Online]. Available: https://software.intel.com/sites/landingpage/IntrinsicsGuide/#

[2] C. Kim, N. Satish, J. Chhugani, H. Saito, R. Krishnaiyer, M. Smelyanskiy, M. Girkar, and P. Dubey, "Closing the Ninja Performance Gap through Traditional Programming and Compiler Technology," Intel® Corporation, Tech. Rep., 2013. [Online]. Available: http://www.intel.com.br/content/dam/www/public/us/en/documents/technology-briefs/intel-labs-closing-ninja-gap-paper.pdf

[3] J. M. Cebrián, L. Natvig, and J. C. Meyer, "Improving Energy Efficiency through Parallelization and Vectorization on Intel Core i5 and i7 Processors," in *2012 SC Companion: High Performance Computing, Networking Storage and Analysis*, Nov. 2012. doi: 10.1109/SC.Companion.2012.93 pp. 675–684.

[4] M. Lorenz, L. Wehmeyer, and T. Dräger, "Energy Aware Compilation for DSPs with SIMD Instructions," in *Proceedings of the Joint Conference on Languages, Compilers and Tools for Embedded Systems: Software and Compilers for Embedded Systems*, ser. LCTES/SCOPES '02. New York, NY, USA: ACM, 2002. doi: 10.1145/513829.513847. ISBN 978-1-58113-527-5 pp. 94–101.

[5] Intel Corporation, "Intel C++ Compiler 17.0 Developer Guide and Reference," 2018. [Online]. Available: https://software.intel.com/en-us/node/682974

[6] G. H. Golub and C. F. Van Loan, *Matrix computations*, 4th ed. Baltimore, Md.: Johns Hopkins University Pr., 2013. ISBN 978-1-4214-0794-4

[7] T. Jakobs, M. Hofmann, and G. Rünger, "Reducing the Power Consumption of Matrix Multiplications by Vectorization," in *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, Aug. 2016. doi: 10.1109/CSE-EUC-DCABES.2016.187 pp. 213–220.

[8] M. Plauth and A. Polze, "Are Low-Power SoCs Feasible for Heterogenous HPC Workloads?" in *Euro-Par 2016: Parallel Processing Workshops*, vol. 10104. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-58943-5_61. ISBN 978-3-319-58942-8 978-3-319-58943-5 pp. 763–774.

[9] T. Rauber and G. Rünger, "Towards an Energy Model for Modular Parallel Scientific Applications," in *2012 IEEE International Conference on Green Computing and Communications*, Nov. 2012. doi: 10.1109/GreenCom.2012.79 pp. 523–532.

[10] H. Lien, L. Natvig, A. A. Hasib, and J. C. Meyer, "Case Studies of Multi-core Energy Efficiency in Task Based Programs," in *ICT as Key Technology against Global Warming*. Springer, Berlin, Heidelberg, Sep. 2012. doi: 10.1007/978-3-642-32606-6_4 pp. 44–54.

[11] H. Caminal, D. Caballero, J. M. Cebrián, R. Ferrer, M. Casas, M. Moretó, X. Martorell, and M. Valero, "Performance and energy effects on task-based parallelized applications: User-directed versus manual vectorization," *The Journal of Supercomputing*, Mar. 2018. doi: 10.1007/s11227-018-2294-9

[12] J. Hofmann, J. Treibig, G. Hager, and G. Wellein, "Comparing the Performance of Different x86 SIMD Instruction Sets for a Medical Imaging Application on Modern Multi- and Manycore Chips," in *Proceedings of the 2014 Workshop on Programming Models for SIMD/Vector Processing*, ser. WPMVP '14. New York, NY, USA: ACM, 2014. doi: 10.1145/2568058.2568068. ISBN 978-1-4503-2653-7 pp. 57–64.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 329–338
DOI: 10.15439/2018F169
ISSN 2300-5963 ACSIS, Vol. 15

On the Autotuning Potential of Time-stepping methods from Scientific Computing

Natalia Kalinnik[1], Robert Kiesel[2], Thomas Rauber[1], Marcel Richter[2] and Gudula Rünger[2]

[1] University Bayreuth
Email: {natalia.kalinnik,rauber}@uni-bayreuth.de

[2] Chemnitz University of Technology
Email: {robert.kiesel,m.richter,ruenger}@cs.tu-chemnitz.de

***Abstract*—Due to the ever changing characteristics of the newly provided hardware, there is the permanent requirement of designing and re-designing software adequately to meet the basic hardware conditions. Especially for well-established software, easy portability of the functional as well as the non-functional properties, such as runtime performance or energy efficiency, would be beneficial, so that the software adapts automatically to the given hardware conditions. In this article, we explore the autotuning potential of several methods from scientific computing. In particular, we consider time-stepping methods and investigate the effect of relevant tuning parameters of the different methods. We also address the question, whether offline or online autotuning approaches are appropriate for the specific method. The methods from scientific computing considered are particle simulation methods, solution methods for differential equations, as well as sparse matrix computations.**

***Index Terms*—offline and online autotuning, performance analysis, portability of efficiency, time-stepping methods.**

I. Introduction

IT IS a well-known fact that the life-span of software is usually much longer than the life-span of the hardware on which it is executed. The common practice is to port and tune the existing software for the new hardware generation by adapting it to the hardware details of the new architecture. Especially, if the software is a complex software system developed over many years, the effort of porting and tuning is high, but the development of a completely new software system would also be too time-consuming. Because of the advent of heterogeneous hardware platforms, the effort for porting, re-designing, tuning or re-developing software is even increasing. The challenge and a major research question is how complex software systems can be developed such that the runtime behavior of the software system can adapt or can be adapted to the ever-changing hardware of a varying heterogeneous type. Specifically, we consider the question whether and how software can gain portable efficiency by self-adaptation, also called autotuning.

In this article, we investigate the tuning potential of important simulation methods from scientific computing and address the question, which techniques can be used to support the tuning towards the development of flexible complex software systems. The simulation methods considered include sparse matrix computations, particle simulation methods, and solution methods for ordinary differential equations. The article provides a systematic investigation of the potential for self-adaptation towards a better runtime performance using offline and online autotuning.

Offline autotuning is performed in a separate offline phase, which is executed at software installation time before the actual software execution and in which the runtime behavior of the software on the given architecture is explored by detailed performance tests with different input scenarios. The test results are used to generate one or several implementations of the software that run efficiently on the given hardware platform for different input sets. After the generation of the software, no further adaptation is performed during production runs. Typical examples for offline autotuning are ATLAS [1] and PHiPAC [2] for dense matrix multiplication. Offline autotuning can be applied if the runtime behavior depends only or mainly on the size of the input set and other characteristics of the input set play only a minor role.

Online autotuning is integrated into the execution of the software. The software observes the performance behavior for the given input set during the runtime and adapts its behavior such that the performance is increased as much as possible. Thus, online autotuning is able to adapt the software behavior to the characteristics of the input set. Examples for online autotuning approaches are Active Harmony [3] or Periscope [4]. The challenge of online autotuning is to integrate the self-adaptation at runtime in such a way that the runtime performance is affected as little as possible.

This article investigates the autotuning potential of simulation methods from scientific computing and discusses the usage of offline and online autotuning approaches. In particular, we provide a detailed performance analysis of the different simulation methods, investigate relevant parameters which have a large influence on the performance, and analyze whether the parameters identified are amenable to autotuning and which autotuning methodology is suitable. Depending on the method from scientific

computing and the parameters identified, different autotuning approaches are best suited for different methods. The article derives a guideline, which method requires which degree for offline and online autotuning.

The rest of the article is structured as follows. Section II gives an overview of relevant aspects of autotuning approaches. Section III considers applications from different areas and discusses their autotuning potential. Section IV discusses related work. Section V summarizes the observations for the different applications and concludes the article.

II. Overview of Autotuning Approaches

In this section, we give an overview of key terms and techniques that are important for applying autotuning for time-stepping simulation methods.

A. Key aspects of self-adaptation

The main terms portability of efficiency and auto-tuning or self-adaptive software are often used as keywords today and we start by giving a precise definition for each of the terms.

An important goal in parallel scientific computing is *portability of efficiency* of simulation algorithms. A performance portable implementation of an application or algorithm can be defined as one that will achieve high performance across a variety of target systems [5]. Depending on the target system, high performance may be quite different, and the definition means high performance for each specific target system. For heterogeneous resources, a major challenge is the diversity of devices on different machines, which provide widely varying performance characteristics [6]. A program optimized for one processor processor may not run as well on the next generation of processors or on a device from a different vendor, and a program optimized for GPU execution is often very different from one optimized for CPU execution.

Adaptation and self-adaptation of software in general refers to the the ability of self-adjustment or self-modification of the software in accordance with changing conditions of environment or structure. This term has first been explored in [7]. Thus, self-adaptive software evaluates its own behavior during execution and changes its behavior when the evaluation indicates that it is not accomplishing what the software is intended to do, or when a better functionality or performance seems to be possible [8].

There are different levels on which software can be tuned or influenced towards a better performance:

- hardware-level: this includes hardware techniques such as dynamic voltage and frequency scaling (DVFS) for energy optimization;
- system-level: this includes scheduling approaches by a compiler or operating system to improve task executions on a parallel target system;
- software-level: this includes modifications at the software-level towards better performance, e.g., source-code transformations such as loop transformations; this captures also the case that the software can adapt itself using knowledge from earlier computing steps.

There are a variety of interactions between these different levels and the final performance improvements achieved for a specific situation may come from performance improvements at different levels. The main focus of this article are performance improvements at the software-level and the questions which methods of self-adaptation are suitable for which application areas.

B. Tuning parameters and possibilities

The efficiency of an application software can depend on many influencing parameters and program transformations, system parameters and characteristics of the input set may have a large influence on the resulting runtime performance or energy efficiency. In this subsection, we provide a short overview.

A new implementation version for a simulation method from scientific computing can be generated by applying correctness-preserving program transformations such as loop interchange, loop distribution, loop unrolling or loop tiling. Moreover, SIMD instruction or memory prefetching techniques can be applied. The result is an implementation version with the same input-output behavior but potentially improved non-functional properties according to a given optimization goal.

Some of the program transformations may be based on the use of transformation parameters. Examples are loop unrolling or loop tiling where an unroll factor or a blocking factor needs to be specified. The selection of a suitable set of program transformations along with their parameter values and an order in which the program transformations should be applied may significantly increase the performance or the energy efficiency. The application of the program transformation could be controlled by a performance model such as the ECM model [9].

The execution of a program implementation may also be influenced by configuration and system parameters. These include the usage of compiler options, the number of threads or processes used for the execution of the implementation and the mapping of these threads or processes to the resources of the given HPC system. For the case of energy efficiency as target function, the selection of the operational frequency for DVFS may also play an important role. For some of the simulation methods, also characteristics of the input set may play an important role for the resulting performance or energy efficiency. Examples are signal processing where the size of the input set may determine which algorithm is the most efficient one [10], [11] or spare linear algorithms where the sparsity and composition of the data structures such as vectors or matrices may play an important role [12], [13]. This behavior can also be observed for particle simulation methods that are considered in this article, where the initial

distribution of the particles may have a strong influence on which simulation method and which implementation leads to the fastest execution.

C. Offline and Online Autotuning approaches

The main emphasis of this article are methods of self-adaptation for the application class of time-stepping simulation methods from scientific computing, since these methods constitute an important class of HPC software. Depending on the characteristics of the specific simulation method and the underlying simulation algorithm, offline or online approaches may be suitable for self-adaptation. In the following, we give an overview.

1) Offline Autotuning: Analyzing the hardware and software for tuning and preselecting parameter sets are main aspects of the offline autotuning. The offline autotuning is performed before the first time step is executed and is performed once.

In this autotuning approach there should be an automated generation of test parameter sets for the identification of the influence of various platform and input parameters, e.g., processor frequency or number of threads. For the quantification of the influences, these parameters will be evaluated and analyzed. This evaluation and analysis can be done with some microbenchmarks. In the offline autotuning approach is also a creation of application-specific performance models, e.g., the Roofline model [14] or the ECM model[9], for the simulation application which can provide additional information. These performance models are evaluated by experiments and facilitate a selection of suitable program variants and configurations regarding all available program variants. The selected program variants are arranged in a decision tree which is provided to the online tuning step, whereby the leaves of the decision tree contain implementation variants considered for the execution and the inner nodes contain conditions to decide which leaves are appropriate. The conditions capture relevant information about the input data that is suitable to identify implementation variants than potentially lead to good runtime results under the conditions given. Depending on the application, the conditions may include the size of the input data and specific properties of the input data such as distribution characteristics for particle simulation methods, distribution patterns for sparse matrix computations, or access distances for solution methods for ordinary differential equations.

2) Online Autotuning: The monitoring and control of the self-adapting behaviour of time-stepping simulation methods are key components of the online autotuning approach. Many mechanisms should be provided to ensure a comprehensive online adaptation process for different simulation methods, which differ significantly in their computational structure, e.g. data arrangement or loop structures. Hence, the following online tuning mechanisms require to be as much application independent as possible in order to be applied to diverse time-stepping simulation methods.

The offline autotuning step pre-selected a diverse set of implementation variants that should be considered for the execution of the time-stepping simulation method. These implementation candidates are processed by a selection and preparation mechanism which ensures a later usage of these candidates. Architectural and algorithmic parameters are determined by an evaluation mechanism and include major properties of the actual input data. This evaluation step is performed before the first time step is executed. The set of determined parameters is used by a search mechanism, which works on parameter configurations and traverses the decision tree built in the offline tuning step. Thus, the search mechanism selects a final set of suitable implementation candidates based on the given parameter configuration. A generic iteration controller mechanism applies the final set of implementation candidates to the first time steps and compares their overall performance. This comparison process determines the best implementation candidate and the appropriate runtime parameters for the execution of the remaining time steps. The monitoring mechanism utilizes the initial comparison of the implementation candidates within the first time steps and observes the overall performance behaviour of the following time steps until the application finishes its execution. Hence, significant deviations of the performance measured can be detected and a new selection of a more suitable implementation candidate can be initiated to react on varying input data properties.

3) Cost estimations: The estimation of resources, e.g. time, energy or memory space, required to solve a given problem can be used to provide upper and lower limits for the appropriate resource. Hence, existing implementations can be rated based on these limits and are more or less suitable to solve a specific problem with actual input data on different HPC-systems or with different execution units, e.g. CPU, GPU or multiple HPC-systems. The cost estimation of a given algorithm or an implementation can be provided by a cost function. The granularity of cost functions can range from estimating whole programs of arbitrary size, which may result in quite complex or merely rough estimations, to specific parts of an implementation, e.g. time-step loops, single loop kernels or basic operations as vector, load or store operations. Since cost functions for loop kernels can be formulated quite accurately, offline autotuning approaches can benefit greatly of such estimation functions to predict the resource consumption of repeating calculations.

Time-stepping simulation methods should ensure an efficient calculation of each time step for varying input data. Thus, several implementations are provided for executing an actual time step, i.e. one loop iteration, and in general perform best with different input data and parameter configurations. Therefore, a selection of implementations has to be applied to limit the number of existing imple-

mentations to match a given requirement, e.g. time or energy consumption. Additionally, the input data may provide further selection criteria for the most suitable implementation variants based on the given cost functions. This leads to a significant reduction of the search space used within the autotuning process and, hence, ensures a faster convergence of the offline tuning step to find the optimal implementation variant for the current execution state.

III. Tuning examples and experimental results

To investigate the potential for self-adaptivity, we consider particle simulation methods, sparse matrix computations, and solution methods for differential equations and analyze their performance behavior with respect to varies tuning parameters, including the degree of parallelism, the operational frequency used, and application-specific parameters such as the grid size for particle simulation methods.

A. Particle simulation methods

We consider particle simulations with long-range interactions caused, e.g., by electrostatic or gravitational forces. The behavior of the particles is simulated by a series of time steps. In each time step, for each particle the simulation computes the forces caused by all other particles and determines the resulting new positions and velocities of the particles. This results in $O(N^2)$ complexity per time step if all interactions are taken into account. Many approaches have been proposed to reduce the complexity, including Fourier-based methods and hierarchical methods. Most of these advanced methods use a splitting approach and distinguish between long-range interactions from far-away particles and short-range interactions from nearby particles. Short-range interactions are typically computed exactly and long-range interactions are typically approximated.

The distinction between short-range interactions and long-range interactions is often performed by using a 3D grid structure with spatial cells. The short-range interactions cover all particles residing in the same or neighboring cells. The size of the spatial cells influences the computational behavior of the simulation and may also have an influence on the resulting accuracy of the simulation. The different advanced particle simulation methods mainly differ in the computation of the long-range interactions. In the following, we consider two different approaches, a Fourier-based approach and a hierarchical tree-based approach based on multipole expansions.

The performance behavior of the particle simulation methods considered depends on different hardware-specific and application-specific parameters. The hardware-specific parameters include the hardware platform used and the number of processes or threads employed. The application-specific parameters mainly include the separation between the short-range and long-range interactions. The number of particles and the initial distribution of the particles may also have a strong influence on the resulting performance of the different methods. In the following, we concentrate on the influence of the number of particles, the number of processors used, and the separation between the short-range and long-range interactions and the resulting grid sizes. The experiments are performed on an Intel Haswell system with two Xeon E5-2683 v3 processors, each equipped with 14 cores and a L3 cache of size 35,840 KB. The performance experiments are performed with two particle systems with non-uniform distribution: a small particle system with $300 \cdot 8^2 = 19,200$ particles and a large particle system with $300 \cdot 8^5 \approx 9.8$ million particles.

Fourier-based methods

Fourier-based methods compute the long-range interactions in Fourier space. Often, fast Fourier-transforms (FFT) are employed. The resulting computational complexity per time step is $O(N \cdot \log N)$, if the particles are sufficiently uniformly distributed [15]. In the following, we use an FFT-based particle simulation method for our experiments. For this method, the separation between the short-range and long-range interactions is determined by the grid sizes used. Figure 1 depicts the resulting execution times for both the small and the large particle system for different grid sizes. The execution times for 4 MPI processes are shown in the left diagram and the execution times for 56 MPI processes are shown in the right diagram. The diagrams show that different grid sizes lead to significantly different execution times due to the differences in the near-field and far-field computations. The optimum grid size that leads to the smallest execution time depends on the number of particles and the number of processes used. For both 4 and 56 MPI processes, the optimum grid size is 32 for the small particle system and 320 for the large particle system. Measurements have shown that these grid sizes are the optimum sizes also for other numbers of processes. For a sequential execution, the optimum grid size remains at 32 for the small particle system and changes to 448 for the large particle system (not shown in a figure).

The development of the execution time for different numbers of MPI processes is shown in Fig. 2 for the small particle system. It can be seen that the relative order between the resulting performance for the different grid sizes does not change with the number of processes.

Hierarchical multipole method

The fast multipole method (FMM) computes the particle interactions based on multipole expansions [16]. In each time step, the method calculates the (gravitational or electrostatic) potential at each particle position, which allows the computation of the new positions and velocities of the particles. The potential is split into a near-field and a far-field potential. To do so, an octree structure is defined, which results from a spatial decomposition

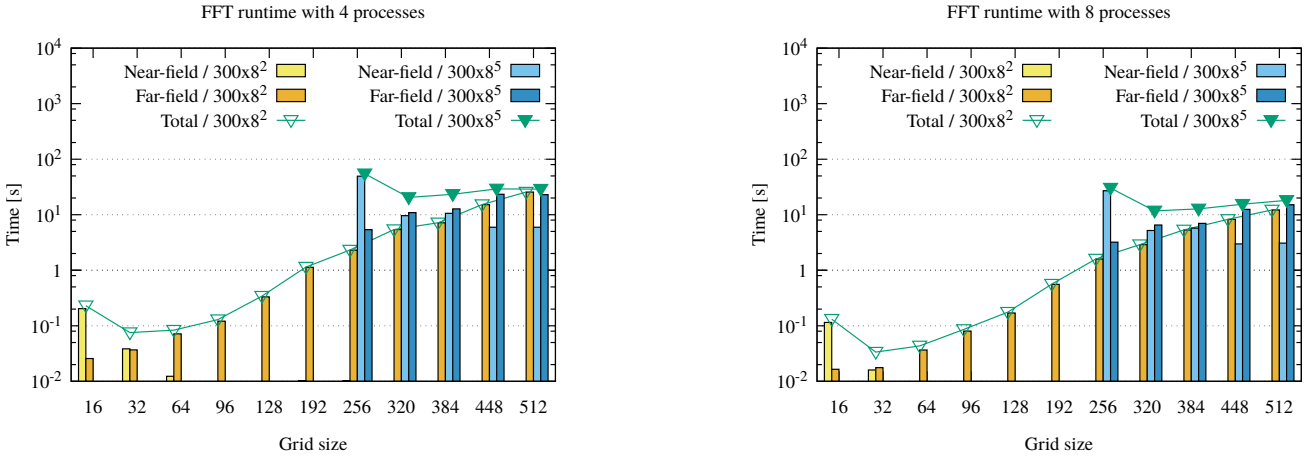


Figure 1. Execution times of the FFT-based particle simulation method for different grid sizes for the small and the large particle system for 4 (left) and 8 (right) MPI processes.

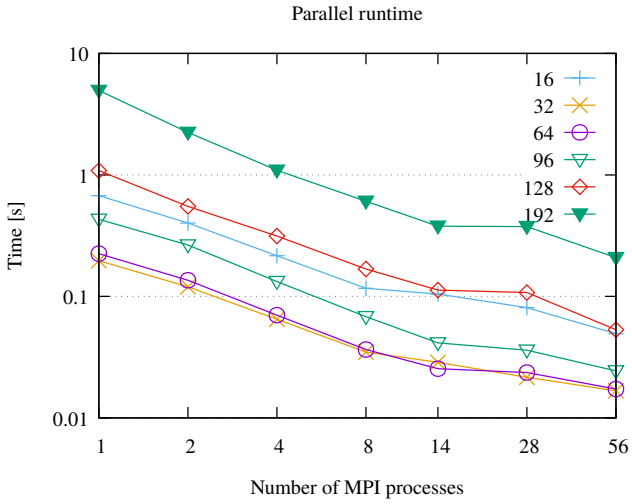


Figure 2. Development of the execution times of the FFT-based particle simulation method for different numbers of MPI processes and different grid sizes between 16 and 192.

into hierarchical boxes, and the particles are sorted into these boxes according to their current position. The spatial decomposition is controlled by a predefined maximum tree depth. For a particle p, the near-field potential is determined by computing the potential at the position of p caused by each of the particles in the same and in neighboring octree boxes. The far-field potential is computed by using approximations for the potential caused by all particles in a specific octree box. These approximations are computed for each octree level. The approximations at suitable octree levels are then used for approximating the far-field potential at a specific particle position. The maximum tree depth determines the separation in the near-field and the far-field potential. The resulting complexity is $O(N)$. In contrast to the FFT-based particle simulation, the FMM does not require that the particles are sufficiently uniformly distributed.

Figure 3 depicts the resulting execution times for both the small and the large particle system for different maximum tree levels. The execution times for 4 MPI processes are shown in the left diagram and the execution times for 56 MPI processes are shown in the right diagram. The figure shows that the maximum tree depth can have a significant influence on the resulting execution time. For the small particle system, the optimum maximum tree depth is 4 for 4 MPI processes, 8 for 8 MPI processes, and 3 for 56 MPI processes. For the large particle system, the optimum maximum tree depth is 7 for 4 MPI processes, 9 for 8 MPI processes, and 10 for 56 MPI processes.

Autotuning-Potential

The experiments of both long-range interaction approaches confirm that the runtime performance is influenced by the particle system size and the separation of the short-range and long-range interactions, i.e., by the grid size or the maximum tree depth. For the hierarchical method, the optimal separation setting to get the best runtime also depends on the number of MPI processes used. Some estimates, e.g., choose high number of MPI processes for better performance, can be done with offline autotuning before the first time step to start with acceptable parameters. To get the optimal parameters they have to be determined with online autotuning. Since the distribution of the particles in the particle system changes after each time step, it is also possible that the optimal parameter setting is changing over time. Thus the online autotuning has to be reapplied after several time steps to respond to imbalances and improve the overall runtime performance. Therefore the performance must be constantly checked for irregularities. For other optimization goals, e.g. energy efficiency, the autotuning-potential behaves the same.

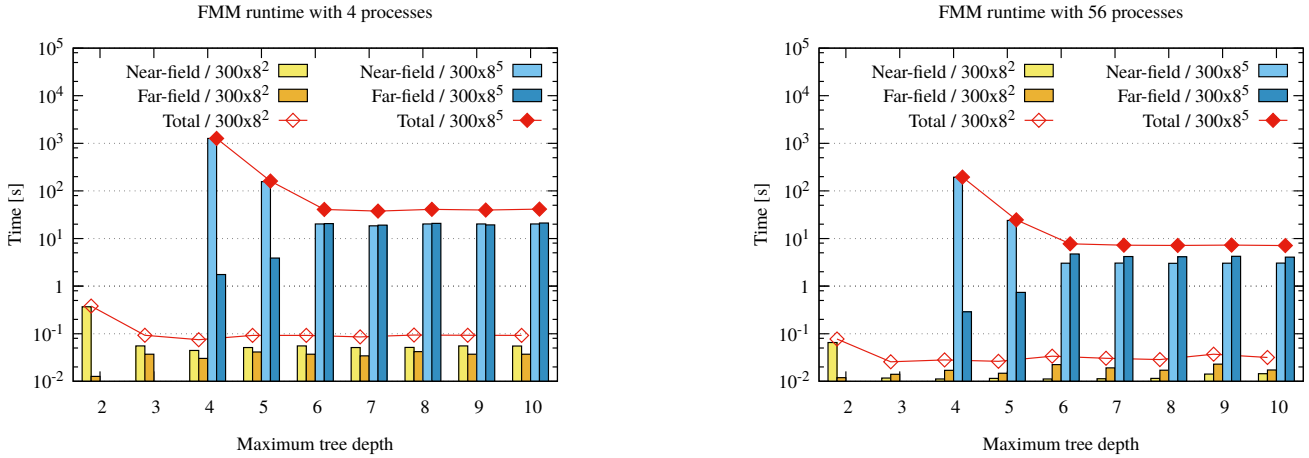


Figure 3. Execution times of the FMM particle simulation method for different maximum tree levels for the small and the large particle system using 4 (left) and 56 (right) MPI processes.

B. *Sparse matrix methods*

Methods performing operations on sparse matrices face additional challenges compared to methods processing dense matrices. Therefore typical calculations as sparse matrix-vector product (SpMV), sparse general matrix-matrix multiplication (GEMM) or solving sets of linear equations using Cholesky factorization have to consider carefully, which memory layout of the sparse input matrices provides the highest benefit in terms of the required performance and given limitations. The sparse property refers to matrices with a significant number of zero matrix elements and, thus, few actual non-zero matrix elements, which in general constitute a rather small fraction of the overall matrix elements.

In the case of the sparse GEMM operation, there is a upper runtime complexity of $O(n^3)$ for a dense matrix memory layout and a simple GEMM algorithm for dense matrices. However, the usage of more suitable sparse matrix memory layouts, called sparse matrix formats (SMF) in the following, is in general based on the exploitation of specific matrix properties of the input matrices and can lead to significant performance improvements. Since the reduction of the number of zero matrix elements stored leads to a better spatial locality when accessing the non-zero matrix elements, memory hierarchies, i.e. the cache hierarchy, can be utilized more efficiently. Hence, performance improvements in terms of time, power and memory consumption can be achieved.

Although the chosen SMF can have a great influence on the performance of the sparse GEMM, the set of parameters influencing performance rather consists of a diverse mix of software and hardware parameters. Solving sparse matrix methods efficiently on a given HPC system relies on several parameters, including the usage of different numbers of threads, the scaling of core and uncore frequency, the availability of SIMD processing units and a proper workload balance within the chosen implementation. Some of the listed parameter influences are considered subsequently with a comparison of three different SMF for a sparse GEMM operation. The SMF chosen are Compressed Sparse Row (CRS), Block Sparse Row (BSR) and Ellpack-Itpack (EIP). The thermal1 matrix with 574,458 non-zero matrix elements is used for the measurements and is taken from the SuiteSparse Matrix Collection [17]. Test systems are a Skylake system with 2× Xeon Gold 6130 processor each with 16 physical cores at 2.1 GHz and a Knights Landing system with a Xeon Phi 7250 processor with 68 physical cores at 1.4 GHz.

The effect of different numbers of threads are depicted in Fig. 4 and show different optimal thread utilizations of the SMF. Moreover, the effect of the nominal core frequency of two different HPC systems can be observed and shows implicitly an inefficient usage of power for certain proposed SMF, e.g. the EIP format can not utilize more than 32 threads for the given test matrix on the Knights Landing system. As an in-depth variable implementation parameter the block size for the BSR format has a significant influence on the achieved runtime as shown in Fig. 5. Comparing different modifiable frequency ranges, the results indicate that the runtime performance of different block sizes can differ such that a specific block size can perform better within a specific frequency range, i.e. the BSR-2 Version can achieve a better runtime within a low frequency range compared to the BSR-16 variant.

Autotuning-Potential

Since a great proportion of the significant parameters for sparse matrix methods are based on hardware properties or on the properties of the input matrices, which are invariant for most sparse matrix methods, a offline autotuning approach implicates the most benefits. Therefore, the application of the proper optimizations can improve the performance required, e.g. runtime, energy consumption or

memory bandwidth utilization. Accordingly, performance models as the Roofline model [14], which have to be created only once for a given hardware configuration, can be used to a great extent for decision making processes of choosing a proper implementation for the desired sparse matrix operation. Furthermore, other performance models, i.e. the Execution-Cache-Memory (ECM) model [18], [9], can provide additional information about memory bound algorithms, which applies to a great proportion of sparse matrix methods. Thus, predictions for a given algorithm can be made for different numbers of threads, so that several optimization goals can be pursued, e.g. optimizing the runtime or finding an optimal number of resources used to satisfy specific energy or memory constraints.

Additionally, an online tuning phase can be used to observe and react to imbalances, which may occur during the execution of the actual problem. This execution behaviour can result due to the initial parameter setup, which was determined in the offline phase. Therefore the necessity for further optimizations can be caused by multiple reasons, e.g. inappropriate estimation of cost functions for kernels executed or data transfer times while using multiple HPC-systems at once. As a result, the runtime or energy measurements may not match the expectations and lead to workload imbalances between different execution units, e.g. CPUs or GPUs. Hence, an online tuning step is desired to adjust the workloads or distribution of data to use all execution resources to full capacity. For a sparse matrix-vector multiplication (SpMV), such an online tuning approach was investigated by [19] and is based on a redistribution of workload between MPI processes if the runtime measurements for the processing of a given number of matrix rows are not similar to measurements of other MPI processes.

Sparse matrix methods benefit the most of an offline tuning phase. However, online tuning approaches can still be applied and are a fine tuning process of the parameters determined in the offline phase. Moreover, the application of online tuning may depend on the actual sparse matrix operation, which should be performed, or the execution units used, e.g. online GPU workload adjustments may get quite complex if the initial data distribution for multiple GPUs has to be modified. Nevertheless, the potential of a mixed tuning approach still contributes to the overall goal of utilizing the given hardware resources to full capacity while optimizing for a given performance goal, e.g. runtime or energy consumption.

C. Solving differential equations

Numerical solution methods for ordinary differential equations (ODEs) compute an approximate solution for a given ordinary differential equation of the form

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)) \text{ with } \mathbf{y}(x_0) = \mathbf{y}_0. \quad (1)$$

by performing a series of time steps one after another until the end of the predefined integration interval is reached

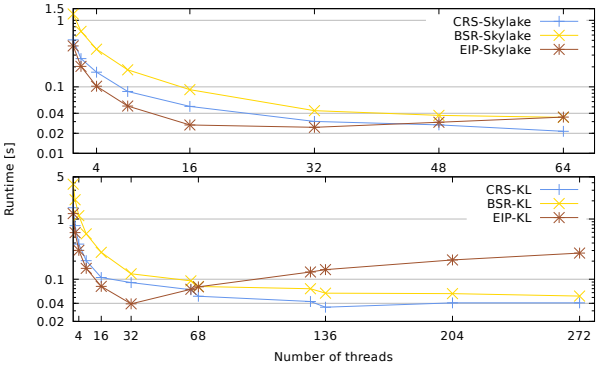


Figure 4. Runtime-Thread scaling of three sparse matrix formats for a Skylake system (top) and a Knights Landing system (bottom).

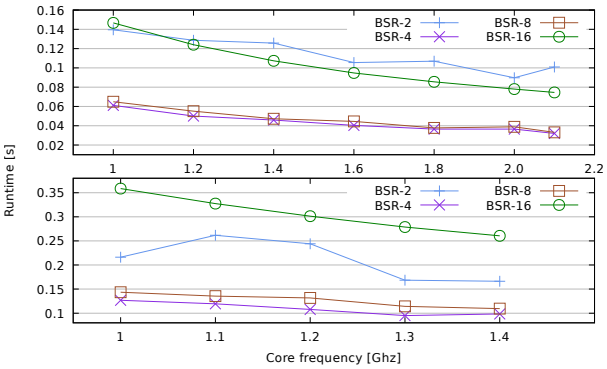


Figure 5. Runtime comparison of the Block Sparse Row matrix format with different block sizes as variable parameter for a Skylake system (top) and a Knights Landing system (bottom). Both diagrams refer to 64 threads used for the BSR variants.

[20]. As example method, we consider iterated Runge-Kutta (RK) methods which perform a fixed number m of computation steps in each time step using the approximation $\mathbf{y}_\kappa$ of the preceding time step. In each computation step, a fixed number s of stage vectors is computed using the stage vectors from the preceding time step and evaluating the function $\mathbf{f}$ defined by the differential equation to be solved. After the last computation step, the stage vectors are combined and an approximation $\mathbf{y}_{\kappa+1}$ for the next point in time is computed. An additional approximation of lower order can be computed additionally for error control and for the selection of the step size for the next time step. Overall, a four-dimensional loop structure results within each time step and many typical loop transformations such as loop interchange, loop unrolling, or loop tiling can be applied.

Taking the parameters of the transformations such as tile sizes and unrolling factors into account, a large number of code variants can be generated and it is not a priori clear, which of these variants will lead to the best performance on a given HPC system. This is an ideal situation for autotuning. A pure offline approach cannot be applied, since the function $\mathbf{f}$ of the differential equation to be solved

may have a large influence on the resulting computational behavior of the different code variants. However, **f** is not known in advance and a numerical solution method should be efficient for different differential equations. On the other hand, a pure online autotuning is also not feasible, since too many code variants would need to be tested. Some of these code variants could be quite slow, leading a further increase of the resulting overhead. Thus, a combination of offline and online autotuning is most promising.

The diagrams in Fig. 6 show the performance of different shared memory implementation variants of the iterated RK method (IRK) for the BRUSS2D example and for different system sizes N. The time per step and component is plotted against the increasing number of threads. BRUSS2D is derived from a reaction diffusion PDE by a spatial discretization on a $N \times N$ grid using the method of lines. The resulting ODE system has dimension $n = 2N^2$. The experiments have been done on a system with two Intel Xeon E5-2697 v3 processors, each equipped with 14 cores and 35 MB L3 shared cache. As RK method we use the LobattoIIIC (8) [20] method with $s = 5$ stages and $m = 7$ computation steps. The code variants [21] utilize data parallelism, the n equations of the ODE system are distributed among the available number of threads p. The variants differ in the loop and the data structures used. Consequently, they have different memory access patterns, resulting in different utilization of the cache and the memory hierarchy. The variants denoted with suffix 'mt' use loop tiling as an optimization technique to further improve the locality of the memory references. Further, four variants (PipeDb1m, PipeDb1mt, ppDb1m and ppDb1mt) exploit a special structure of the function **f** of BRUSS2D by overlapping of vectors and by using pipeline-like computational structure of the computation steps m [21]. These variants only require lock-based local synchronization with neighbor threads, whereas all other variants need global barrier operations.

The diagrams in Fig. 6 indicate that the performance of IRK variants depends on the runtime parameters, such as the dimension of the ODE system and the number of threads executing the program. In particular, following observations can be made: (1) For the same system size, but for different numbers of threads, the order of the implementation variants varies. For example, for $N = 460$ and thread numbers $p < 20$, the variant EAmt offers the best performance, closely followed by the variant E and A. For $p = 20$ all variants require nearly the same execution time. For even larger numbers of threads, the variants EAmt, A and E are the slowest variants. The main reason for the smaller efficiency of these variants for large numbers of threads are the costs of the barrier operations used for synchronization of the threads. The variants A, E, EAmt require two barrier operations per stage in each computation step. All other variants need either only one barrier operation per computation step or use more efficient lock-based synchronization. (2) For

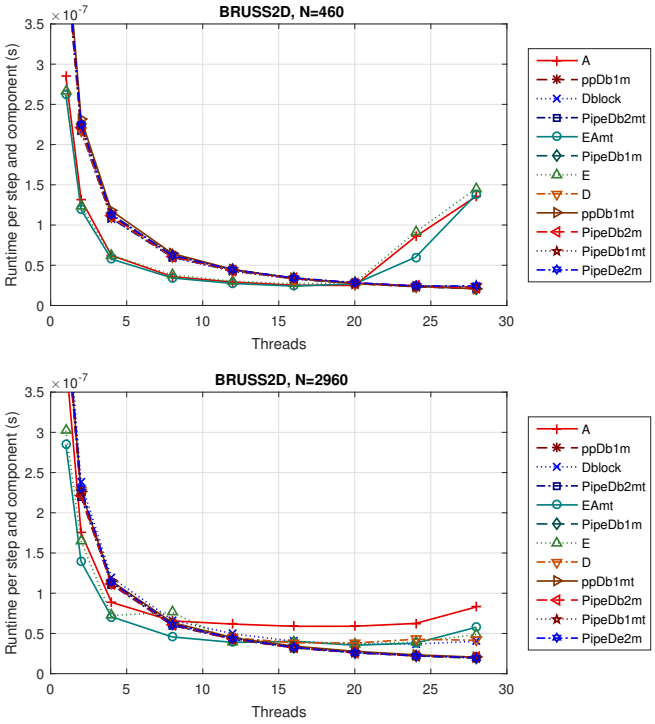


Figure 6. Execution time per step and component of iterated RK method for $N = 460$ (top) and $N = 2960$ (bottom) for different numbers of threads.

the same number of threads, but for different system sizes, different variants obtain the best performance. For example, for $p = 16$ and $N = 460$, the variants EAmt, E and A perform best, whereas for $p = 16$ and $N = 2960$ other variants run faster.

Autotuning-Potential

The experiments confirm that the performance of IRK variants is strongly influenced by the input data. Moreover, our experience shows that the performance of IRK variants also depends on the characteristics of the target architecture, such as the specific multi-core processor design, the cache architecture and the resulting memory latency and bandwidth. Thus, to obtain maximum performance, it is important that an IRK solver can adapt to the characteristics of the underlying architecture and of the ODE problem to be solved. Since usually these parameters are only known at runtime, the best variant cannot be determined at compile or installation time, and offline autotuning is not sufficient. For ODE solvers online autotuning has to be applied, but due to the large search space of candidate implementations and implementation parameters such as tile sizes for loop tiling or factors for loop unrolling, an online autotuning strategy should be supported by offline autotuning. For example, for multi-threaded implementations, offline measurements can be used to estimate the synchronization overhead of different

implementation variants. At runtime this information can be used to avoid the evaluation of variants if their synchronization overhead is too high to outperform variants with lower synchronization.

IV. Related work

The first autotuning approaches were offline approaches for numerical methods for dense linear algebra problems for which the properties of the HPC systems used play the most important role for the execution time. These approaches include ATLAS [1] and PHiPAC [2]. For some application areas, properties of the input data may have a significant influence on the resulting execution time [22]. In these cases, online approaches need to be employed or integrated. An example for such an application area is signal processing, where the size of the problem to be solved determines which algorithm is the most efficient one. Examples for approaches in this direction are FFTW [10] and SPIRAL [11], [23]. For sparse linear algebra problems, the distribution on the non-zeros in the matrix to be processed may have a large influence, see OSKI [12] and SALSA [13]. Several autotuning approaches for specific application areas have been developed on the basis of domain-specific languages (DSL) for the description of the processing algorithms. From these DSL descriptions, a compiler can generate different code variants that can be tested in an online phase. An important application area for these approaches are stencil computations, see PATUS [24], Pochoir [25] and Halide [26] for approaches in this direction. All approaches mentioned above are application-specific, i.e., they have been developed for a specific application area and exploit specific properties of this application area.

Several general autotuning framework have been proposed that work independent from a specific application area. Active Harmony [3] is such a general autotuning framework that aims at iterative parallel applications that can come from many application areas. It includes a source-to-source compiler tool to generate new code variants at runtime according to loop transformations as specified by the user. Active Harmony uses a pure online approach, no offline component is included. Another general autotuning framework is Perpetuum [27], which aims at the selection of tuneable parameters such as block sizes and number of threads. Parcae [28] provides a user-level runtime system and a compiler to translate parallel constructs and patterns into a task-based execution model. PetaBricks [22] and Periscope [4] are other approaches in this direction. A detailed survey of compiler autotuning approaches with an emphasis on machine learning is given in [29].

All approaches mentioned above are either offline or online approaches, depending on the needs of the specific applications area. None of these approaches employs both an offline and an online phase as it is attempted in this paper.

The generation of different code variants is an important part of many autotuning approaches. The polytope model [30] and compiler-based approaches [26] are often used in this context. The resulting number of code variants can be large and efficient heuristic search strategies are important, including Simulated Annealing or Nelder-Mead [31]. OpenTuner [22] provides several search strategies, but other techniques based on machine learning are also investigated [32]. Energy and performance autotuning for two irregular applications, graph community detection using the Louvain method (Grappolo) and high-performance conjugate gradient (HPCCG) have been investigated in [33] for OpenMP multithreaded programs using OpenTuner.

V. Conclusion

In this article we have investigated the tuning potential of important simulation methods from scientific computing. The simulation methods considered include sparse matrix computations, particle simulation methods and solution methods for ordinary differential equations. In particular, a detailed performance analysis has been performed with considerations of relevant parameters. The tuning potential has been investigated for offline and online tuning.

The investigation shows that sparse matrix computations are mainly amenable to offline autotuning, particle simulation methods require the use of online autotuning, and solution methods for ordinary differential equations need a combination of both offline and online autotuning. This is related to the complexity of the input data and the the number of implementation variants available.

The offline autotuning analyzes the hardware and software and pre-selects parameters. With performance models, e.g., the roofline model or ECM model, some predictions can be done. The three applications make different usage of this offline approach. While good parameters can be chosen for sparse matrix calculations, the offline approach is used for the differential equations to reduce the search space used in the online approach.

The online autotuning is good for adjusting the parameters and reacting to imbalances. While the importance of this approach is different for the applications, each time-stepping method can be adjusted at runtime to achieve the best performance. Insufficient decisions from the offline approach can be adjusted. To detect these imbalances, a monitoring must be performed. This monitoring should be application independent and can be done with established tools.

Thus, it is usually advantageous to use both approaches to tune methods from scientific computing. The offline approach to set start parameters as good as possible and to select suitable code variants, and the online approach to fine-tune parameters, react on imbalances and select the appropriate code variant.

Acknowledgement

This work has been supported by the German Ministry of Science and Education (BMBF), Project title SeASiTe (Self-Adaptation of Time-step-based Simulation Techniques on Heterogeneous HPC Systems) with project number 01IH16012A/B.

References

[1] R. Whaley, A. Petitet, and J. Dongarra, "Automated empirical optimizations of software and the ATLAS project," *Parallel Computing*, vol. 27, no. 1-2, pp. 3–35, 2001. doi: 10.1016/S0167-8191(00)00087-9

[2] J. Bilmes, K. Asanovic, C. Chin, and J. Demmel, "Optimizing Matrix Multiply Using PHiPAC: A Portable, High-performance, ANSI C Coding Methodology," in *Proc. of the 11th Int. Conf. on Supercomputing*, ser. ICS '97. New York, NY, USA: ACM, 1997. doi: 10.1145/263580.263662 pp. 340–347.

[3] A. Tiwari and J. K. Hollingsworth, "Online adaptive code generation and tuning," in *Proc. of the 2011 IEEE Int. Parallel & Distributed Processing Symp. (IPDPS 2011)*. IEEE, 2011. doi: 10.1109/IPDPS.2011.86 pp. 879–892.

[4] M. Gerndt, E. César, and S. Benkner, Eds., *Automatic Tuning of HPC Applications – The Periscope Tuning Framework*. Shaker Verlag, 2015, doi: 10.2370/9783844035179.

[5] M. Wolfe, "Compilers and More: What Makes Performance Portable?" April 19, 2016.

[6] P. Phothilimthana, J. Ansel, J. Ragan-Kelley, and S. Amarasinghe, "Portable Performance on Heterogeneous Architectures," in *Proc. of the Eighteenth Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS '13)*. ACM, 2013. doi: 10.1145/2451116.2451162 pp. 431–444.

[7] J. Aseltine, A. Mancini, and C. Sarture, "A survey of adaptive control systems," *IRE Transactions on Automatic Control*, vol. 6, no. 1, pp. 102–108, Dec 1958. doi: 10.1109/TAC.1958.1105168

[8] D. BAA-98-12, "DARPA Broad Agency Announcement on Self Adaptive Software," 1997.

[9] G. Hager, J. Treibig, J. Habich, and G. Wellein, "Exploring performance and power properties of modern multi-core chips via simple machine models," *Concurrency and Computation: Practice and Experience*, vol. 28, pp. 189–210, 2016. doi: 10.1002/cpe.3180

[10] M. Frigo and S. Johnson, "The design and implementation of FFTW3," *Proc. of the IEEE*, vol. 93, no. 2, pp. 216–231, 2005. doi: 10.1109/JPROC.2004.840301 Special issue on "Program Generation, Optimization, and Platform Adaptation".

[11] F. de Mesmay, Y. Voronenko, and M. Püschel, "Offline library adaptation using automatically generated heuristics," in *Int. Parallel and Distributed Processing Symp. (IPDPS)*, 2010. doi: 10.1109/IPDPS.2010.5470479

[12] R. Vuduc, J. Demmel, and K. Yelick, "OSKI: A library of automatically tuned sparse matrix kernels," in *Institute of Physics Publishing*, vol. 16, 2005. doi: 10.1088/1742-6596/16/1/071

[13] V. Eijkhout and E. Fuentes, "Machine learning for multi-stage selection of numerical methods," in *New Advances in Machine Learning*. INTECH, feb 2010, pp. 117–136, doi: 10.5772/9376.

[14] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, no. 4, pp. 65–76, Apr. 2009. doi: 10.1145/1498765.1498785

[15] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*. Bristol, PA, USA: Taylor & Francis, Inc., 1988, doi: 10.1137/1025102.

[16] L. Greengard, *The Rapid Evaluation of Potential Fields in Particle Systems*. Boston: MIT Press, 1988.

[17] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 1:1–1:25, Dec. 2011. doi: 10.1145/2049662.2049663

[18] J. Hofmann, J. Eitzinger, and D. Fey, "Execution-Cache-Memory Performance Model: Introduction and Validation," *ArXiv e-prints*, Sep. 2015.

[19] S. Lee and R. Eigenmann, "Adaptive runtime tuning of parallel sparse matrix-vector multiplication on distributed memory systems," in *Proceedings of the 22Nd Annual International Conference on Supercomputing*, ser. ICS '08. ACM, 2008. doi: 10.1145/1375527.1375558 pp. 195–204.

[20] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*. Berlin: Springer–Verlag, 1993, doi: 10.1137/1032091.

[21] M. Korch and T. Rauber, "Locality optimized shared-memory implementations of iterated Runge-Kutta methods," in *Euro-Par 2007. Parallel Processing*, ser. Springer LNCS, vol. 4641. Springer, 2007. doi: 10.1007/978-3-540-74466-5_78 pp. 737–747.

[22] J. Ansel, "Autotuning programs with algorithmic choice," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, Feb. 2014. [Online]. Available: http://groups.csail.mit.edu/commit/papers/2014/ansel-phd-thesis.pdf

[23] M. Püschel, J. Moura, J. Johnson, D. Padua, M. Veloso, B. Singer, J. Xiong, F. Franchetti, A. Gacic, Y. Voronenko, K. Chen, R. Johnson, and N. Rizzolo, "SPIRAL: Code generation for DSP transforms," *Proc. of the IEEE, special issue on "Program Generation, Optimization, and Adaptation"*, vol. 93, no. 2, pp. 232– 275, 2005. doi: 10.1109/jproc.2004.840306

[24] M. Christen, O. Schenk, and H. Burkhart, "PATUS: A code generation and autotuning framework for parallel iterative stencil computations on modern microarchitectures," in *Proc. of the 25th IEEE Int. Parallel and Distributed Processing Symp.*, May 2011. doi: 10.1109/IPDPS.2011.70

[25] Y. Tang, R. A. Chowdhury, B. C. Kuszmaul, C.-K. Luk, and C. E. Leiserson, "The Pochoir stencil compiler," in *Proc. of the Twenty-third Annual ACM Symp. on Parallelism in Algorithms and Architectures (SPAA '11)*. ACM, 2011. doi: 10.1145/1989493.1989508 pp. 117–128.

[26] J. Ragan-Kelley, C. Barnes, A. Adams, S. Paris, F. Durand, and S. Amarasinghe, "Halide: A language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines," in *Proc. of the 34th ACM SIGPLAN Conf. on Programming Language Design and Implementation (PLDI'13)*, 2013. doi: 10.1145/2499370.2462176 pp. 519–530.

[27] T. Karcher and V. Pankratius, "Run-time automatic performance tuning for multicore applications," in *Euro-Par 2011. Part I.*, ser. LNCS, E. Jeannot, R. Namyst, and J. Roman, Eds., no. 6852. Springer, 2011. doi: 10.1007/978-3-642-23400-2_2 pp. 3–14.

[28] A. Raman, A. Zaks, J. Lee, and D. August, "Parcae: A System for Flexible Parallel Execution," in *Proc. of the 33rd ACM SIGPLAN Conf. on Programming Language Design and Implementation*, ser. PLDI '12, 2012. doi: 10.1145/2254064.2254082 pp. 133–144.

[29] A. H. Ashouri, G. Palermo, J. Cavazos, and C. Silvano, *Automatic Tuning of Compilers Using Machine Learning*, 1st ed. Springer, 2017. doi: 10.1007/978-3-319-71489-9.

[30] D. Feld, T. Soddemann, M. Jünger, and S. Mallach, "Facilitate SIMD-Code-Generation in the Polyhedral Model by Hardware-aware Automatic Code-Transformation," in *Proc. of the 3rd International Workshop on Polyhedral Compilation Techniques*, A. Größlinger and L.-N. Pouchet, Eds., Berlin, Germany, Jan. 2013. doi: 10.13140/2.1.5066.3368 pp. 45–54.

[31] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965. doi: 10.1093/comjnl/7.4.308

[32] S. Muralidharan, M. Shantharam, M. Hall, M. Garland, and B. Catanzaro, "Nitro: A framework for adaptive code variant tuning," in *28th IEEE Int. Parallel and Distributed Processing Symp. (IPDPS 2014)*, May 2014. doi: 10.1109/IPDPS.2014.59 pp. 501–512.

[33] A. Panyala, D. Chavarra-Miranda, J. B. Manzano, A. Tumeo, and M. Halappanavar, "Exploring performance and energy tradeoffs for irregular applications," *J. Parallel Distrib. Comput.*, vol. 104, no. C, pp. 234–251, Jun. 2017. doi: 10.1016/j.jpdc.2016.06.006

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 339–346
DOI: 10.15439/2018F177
ISSN 2300-5963 ACSIS, Vol. 15

Analyzing energy/performance trade-offs with power capping for parallel applications on modern multi and many core processors

Adam Krzywaniak*, Jerzy Proficz†, Pawel Czarnul*
* Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Narutowicza 11/12, 80-233 Poland
Email: adam.krzywaniak@pg.edu.pl, pczarnul@eti.pg.edu.pl
† Centre of Informatics — Tricity Academic Supercomputer & networK (CI TASK)
Gdansk University of Technology
Narutowicza 11/12, 80-233 Poland
Email: j.proficz@task.gda.pl

***Abstract*—In the paper we present extensive results from analyzing energy/performance trade-offs with power capping observed on four different modern CPUs, for three different parallel applications such as 2D heat distribution, numerical integration and Fast Fourier Transform. The CPU tested represent both multi-core type CPUs such as Intel® Xeon® E5, desktop and mobile i7 as well as many-core Intel® Xeon Phi™ x200 but also server, desktop and mobile solutions used widely nowadays. We show that using enforced power caps we can find points of lower than default energy consumption but mostly for desktop and mobile solutions at the cost of increased execution time. We show with particular numbers how energy consumed, power consumption and execution time change for the point of minimum energy used versus the default configuration with no power limit, for each application and each tested CPU.**

I. Introduction

NOWADAYS the consumption of electric energy by the Information and Communication Technology (ICT) sector reaches extreme values, it is estimated as 269 TWh per year and 2% of global CO_2 emissions. An average data center, having 2 600 m^2 server rooms, causes almost 2 MW IT load [1]. Thus the energy conservation is very important for such environments as well as for mobile/IoT computations where the battery lifetime can be significantly extended by performing various procedures such as power level capping or calculation offload [2].

Considering the microprocessor devices: Central Processor Units (CPUs) and their applications, usually the actual power level used by such a device is proportional to its current workload. However, many modern CPUs are able to control their maximum power level via special API, e.g. RAPL [3]. Thus, in many cases for such CPUs, energy consumption depends not only on the workload, but also on the actual power cap, set by the managing software or directly by the developer.

For CPUs, it is important to distinguish between the power level and energy consumption — the factor is execution time, as it is presented in the following sections, sometimes the same problem is solved using a lower amount of energy (measured in J or kWh), despite the higher (average) power level (measured in W) observed in the device. There is no simple conversion between these two values, but in many cases, there is a spot where limited power causes lower energy consumption. In this paper we are going to analyze such minima — these can be exploited to trade off between energy consumption and execution time.

Our original contribution covers: (i) the presentation and analysis of execution time, power and energy consumption measurements for different power level caps of various CPU types, (ii) evaluation of trade-off between execution time and energy consumption for three representative HPC applications.

The next section describes the related works, the detailed goal of tests is presented, afterwards performed experiments are described, including the testbed applications, systems as well as the obtained results. And finally conclusions and future work are covered.

II. Related work

In the context of high performance computing (HPC), energy consumption and energy efficiency are among the most important challenges. It is important in particular for execution that is energy efficient for various levels of utilization [4]. The authors of [5] investigate software methods aimed at improving energy efficiency in parallel computing. In particular it focuses on load imbalance, mixed precision in floating-point operations. Power consumption of compute components is characterized. Energy efficiency metrics are introduced including dynamic energy improvement for n processors. The taxonomy of methods considered in this work includes power-aware

Supported partially by the Polish Ministry of Science and Higher Education. The experiments were partially performed using high-performance computing infrastructure provided by Centre of Informatics — Tricity Academic Supercomputer & networK (CI TASK) at Gdansk University of Technology.

scheduling and resource management, parallelization oriented including balancing, communication focused, approximation methods with part of computations executed with lower energy usage and slight loss of accuracy. In papers [6] and [7] we modeled energy consumption of parallel applications focusing on various communication routines establishing both functions and coefficients valid for various cluster systems.

In terms of measurement of power consumption and energy usage, several tools and techniques have been used and reported. IgProf which is an open source performance profiler is available for x86 and x86-64 as well as ARMv7 and ARMv8 platforms. The authors of [8] added a module for statistical sampling energy profiling. Measurements have been taken using the RAPL interface. The STREAM benchmark has been used to gather results that demonstrate expected correlation between execution time and energy consumption.

RAPL (Running Average Power Limit) provides counters to take measurements of energy consumption of CPUs, integrated GPUs and memory as well as to set corresponding power limits allowing to manage energy efficiency of a system. Paper [9] focuses on measurement and power limiting for main memory for server platforms. It has been shown that power limits can be enforced with minimization of performance impact of the approach. SPECCPU2000 sub-benchmarks were run for various power limits.

Paper [10] validates DRAM related results from RAPL for desktop and server environments with DDR3 and DDR4 types of memories. RAPL results were compared to actual measurements with matching in general within roughly 20%. Tests were performed with a variety of benchmarks including sleep, HPL Linpack, gcc PAPI, SmallptGPU2 ray-tracer, Kerbal Space Program. RAPL has also been validated in works [3] and [11]. In the latter, the authors concluded that RAPL power estimation is more accurate on IvyBridge than on SandyBridge generation of CPUs. In work [12] and paper [13] the authors investigate power consumption of various components including instruction decoders in x86-64 processor which was done through microbenchmarks. The authors have concluded that decoders consume between 3% and 10% of the total processor package power.

A hybrid hardware/software power capping system called PUPiL was evaluated in [14] for maximizing performance under power capping. The solution was compared to RAPL and e.g. a software-based DVFS control system and a software based decision system. PUPiL showed response time similar to hardware approaches and generally better performance than RAPL under power constraints.

The authors of [15] notice the phenomenon called PERC (Performance-Equivalent Resource Configurations) according to which applications with various configurations of resources show similar performance at various power consumption and use it for their PowerCap algorithm that selects a configuration that follows power limits. The authors claim that the algorithm requires 50% less reconfiguration and 12% more power compared to the DVFS approach.

In paper [16] authors propose an algorithm for scheduling execution of independent jobs on a system with integrated CPU-GPU with consideration of power caps. The authors have shown that throughput has been improved by between 9% and 46% over default schedules.

As an example, in [17] the author has performed detailed analysis of power consumption of Intel® i7-4820K. It should be noted, similarly to our findings for our testbed applications, that the power consumption of an application computing prime numbers reaches roughly 40W at the highest considered frequency at the TDP of the CPU equal to 130W.

The authors of [18] present empirical assessment of vendor provided power capping on a Cray XC40 system and comparison of performance with p-state control. They concluded generally better performance of the latter for many benchmarks in HPC.

III. Motivations and concept of research

Based on the aforementioned related works, we intend to perform detailed research, for a representative set of HPC applications, into energy/performance trade-offs for modern multi- and many- core CPUs using software imposed power caps.

Specifically, we are looking into such a configuration, for each application and each CPU, for which the total energy consumed during computations is minimized, compared to the default configuration without power consumption caps for a CPU i.e. full computational power. For such energy minimized configurations, we are looking into energy/performance trade-offs. It is especially interesting to analyze and observe it for various modern CPUs that differ, in terms of the target market, design and numbers of cores:

- server: Intel® Xeon Phi™ x200 (many-core CPU), Intel® Xeon® E5 (multi-core CPU) present in many workstations and cluster nodes,
- desktop: Intel® Core™ i7 desktop present in many home and office computers,
- mobile: Intel® Core™ i7 mobile present in many laptops and notebooks.

The software power caps as well as energy consumption measurements are implemented using RAPL driver available in modern Intel® CPUs. Due to technical limitations in measuring the impact of our power caps on the whole server we read the energy consumption using RAPL from the Package (CPU + DRAM) and acknowledge it representative and valuable result.

In terms of applications, we use three parallel applications, that differ in the computing paradigms and compute/synchronization overhead ratios:

- geometric single program multiple data stream: heat distribution,
- master-slave: numerical integration and FFT.

This continues our work [19] of analysis of representative parallel applications with consideration of energy usage.

IV. EXPERIMENTS

A. Testbed applications

For the testing purposes we selected three representative problems found in high performance computing (HPC) environment, and accordingly, implemented three applications, which are executed concurrently, and are horizontally scalable, i.e. speeding up with the increase of the core number; however they utilize shared memory for data exchange and synchronization, thus in this case they cannot be distributed between more compute nodes.

The application were implemented in C language v. C99, using OpenMP [20] for processing parallelization. They were compiled by the GCC v. 4.8 with maximal provided optimization (parameter -O3). They use the default OpenMP configuration (omp directive: schedule(auto)) regarding the thread number and computation partitioning, in the execution environment we did not tune up these settings. The applications use the floating point double precision for calculations (C language type: double).

The first application performs a numerical integration of a given function in a specified range. The specified partition is split between working threads and each thread calculates the sum of its range. The intermediate results are stored separately for each thread, although OpenMP is responsible for their reduction (omp directive: reduction(+:)). For testing purposes we defined the function as $f(x) = \frac{1}{1+x}$. The arguments of the application allow to specify the range and the calculation's precision as a number of subpartitions to be integrated.

The second application is a simplified version of the 2D heat distribution simulation (based on the conception proposed in [21]) over the closed square area, divide into $N \times N$ parts and containing a set of working heaters. For test purposes we set $N = 1000$ and introduced one heater located in the area corner. The input parameters cover a speed of heat propagation and a number of iteration to be simulated. The solution uses three memory buffers: (i) a constant heater map in the room, (ii) an input buffer with the current heat distribution, (iii) an output buffer with the heat distribution after performing current step. The buffers (i) and (ii) are swapped after each step of simulation: the output buffer in step i becomes the input buffer of step $i + 1$. The temperature of each square in the room can be calculated independently, thus potentially the above problem can be parallelized (omp directive: for) among threads as well as the threads do not interfere each other, each one has its own area to perform the simulation.

The third application is a parallel implementation of Fast Fourier Transform (FFT). It uses Radix-2 algorithm with Decimation-in-Time parallelization strategy [22]. At the beginning the sequence of N transformed samples (the input data) is parallelly shuffled, then the $log_2 N$ iterations are executed, where the parallel computations over complex numbers are performed: each thread has its own range of the data to process (omp directive: for). The result is placed in the array replacing the input data. For the benchmark purposes the input data is automatically generated.

B. Testbed systems

As a testbed environment we used 4 different systems. Two of them contained server dedicated processors with multi-core (Xeon® E5 v4) and many-core (Xeon Phi™ x200) architectures. Another two systems was based on Intel® Core™ i7 processors, one dedicated for desktop and one dedicated for mobile personal computers. Parameters of aforementioned systems are presented in Table I.

C. Results

Obtained results are presented for each testbed system separately. Therefore, for each of the four systems we present three figures individual for each of testbed applications and one common figure. In the individual figures for each power limit preset (bottom axis) we present execution time of the tested application (left axis) as well as total energy consumed (right axis). The common figure presents average power (left axis) for each test run against the power limit (bottom axis) that was preset for each of three tested applications.

Figure 1 presents results obtained using the testbed system with server Xeon E5 v4 for Fast Fourier Transform, simulation of heat distribution and numerical integration respectively. The most important observation is that the testbed aplications used in experiments are not able to reach the TDP of server Xeon processor. In each case for the experiment with maximum power limit we use less then 50% of available power. However, when the limit is set below 50% of TDP and close to the reference power consumption the average power consumption starts to respect the enforced limit. For this system the benefits of lowering the power consumption can be observed only for one testbed application (FFT) for which we can find the minimum of energy consumed while running calculations with different power limits. Howewer, the minimum is still saving less than 3% of energy comparing to reference run with no limits. The two other applications have the most energy efficient point at their default settings of the power limit.

Figure 2 presents the results obtained using testbed system with another server processor, Xeon Phi x200, again for Fast Fourier Transform, simulation of heat distribution and numerical integration respectively. Although both server processors present far different architectures (multi-core vs. many-core), the results of the experiments are quite similar. The main common feature of experimental results is that again our testbed applications are using less then 50% of available power. The system respects the preset limit in each case until the minimum value (85W) is reached. For two of tested applications (heat distribution and FFT) we observed the minimum of energy consumed for the value of power limit 135W. However, the energy benefits are not significant again (1-3% energy saved). For numerical integration the lowest energy consumption was obtained again when the power limit had its default value.

Figure 3 presents results of the same testbed applications for the first of non-server testbed systems with mobile PC dedicated Intel® Core™ i7 processor. Results for the last system with another Intel® Core™ i7 processor, dedicated

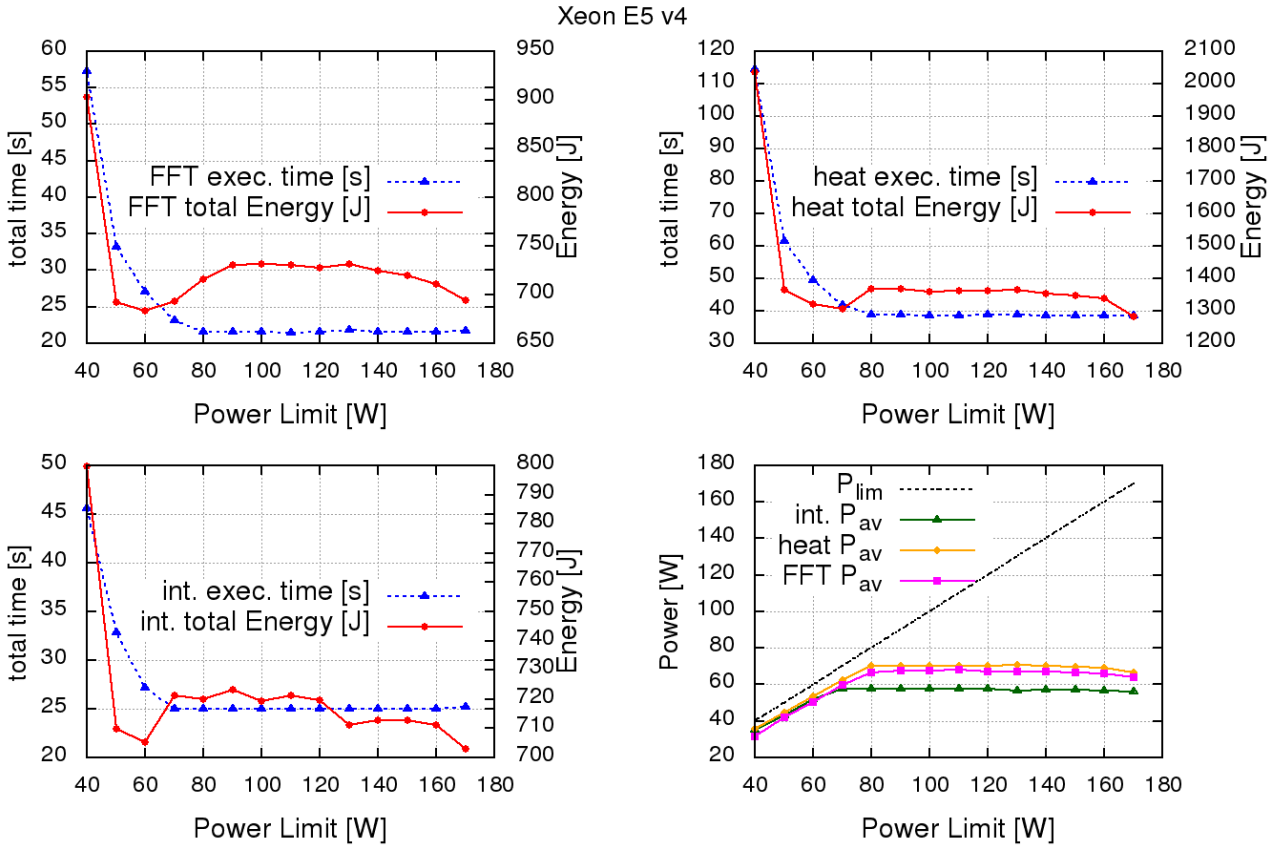


Fig. 1. Tests results for Xeon® E5 v4.

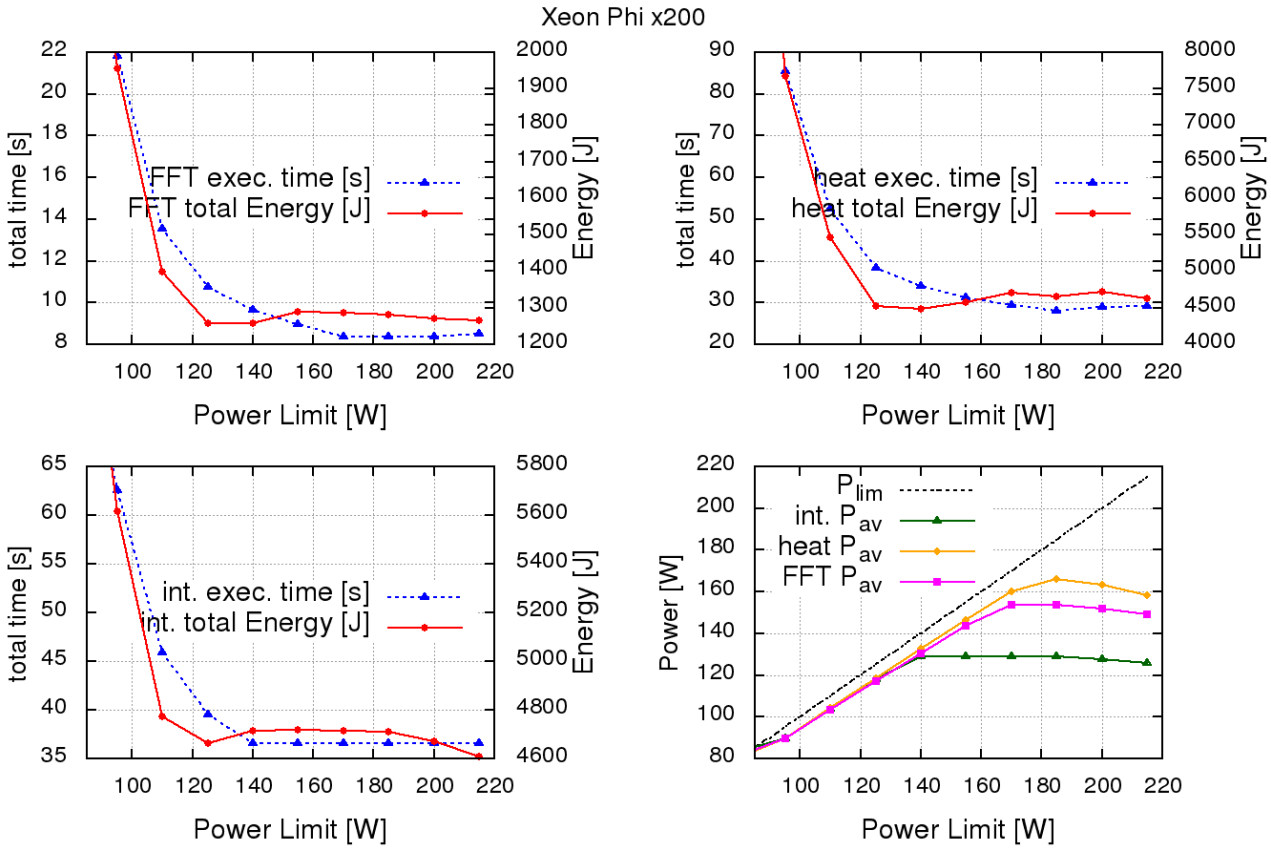


Fig. 2. Tests results for Xeon Phi™ x200.

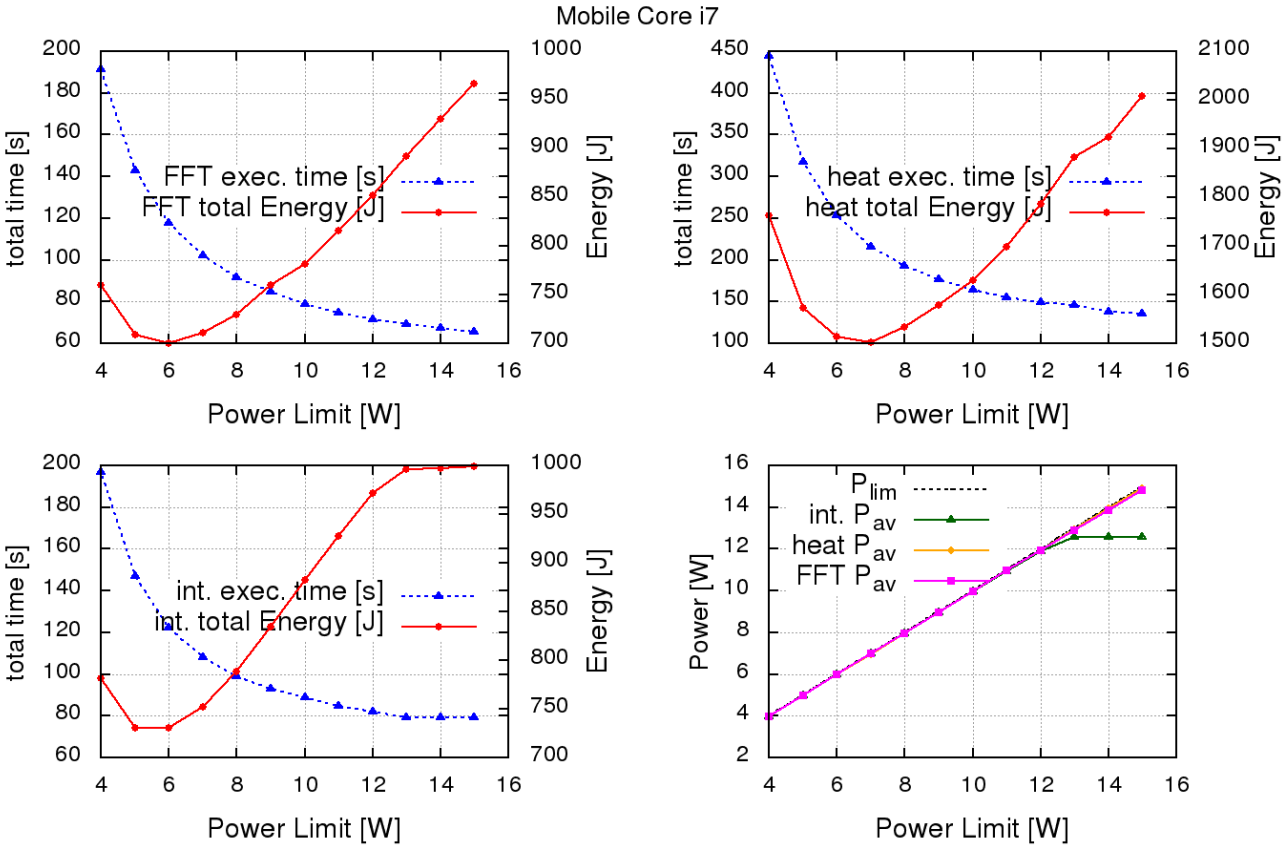


Fig. 3. Tests results for mobile Core™ i7.

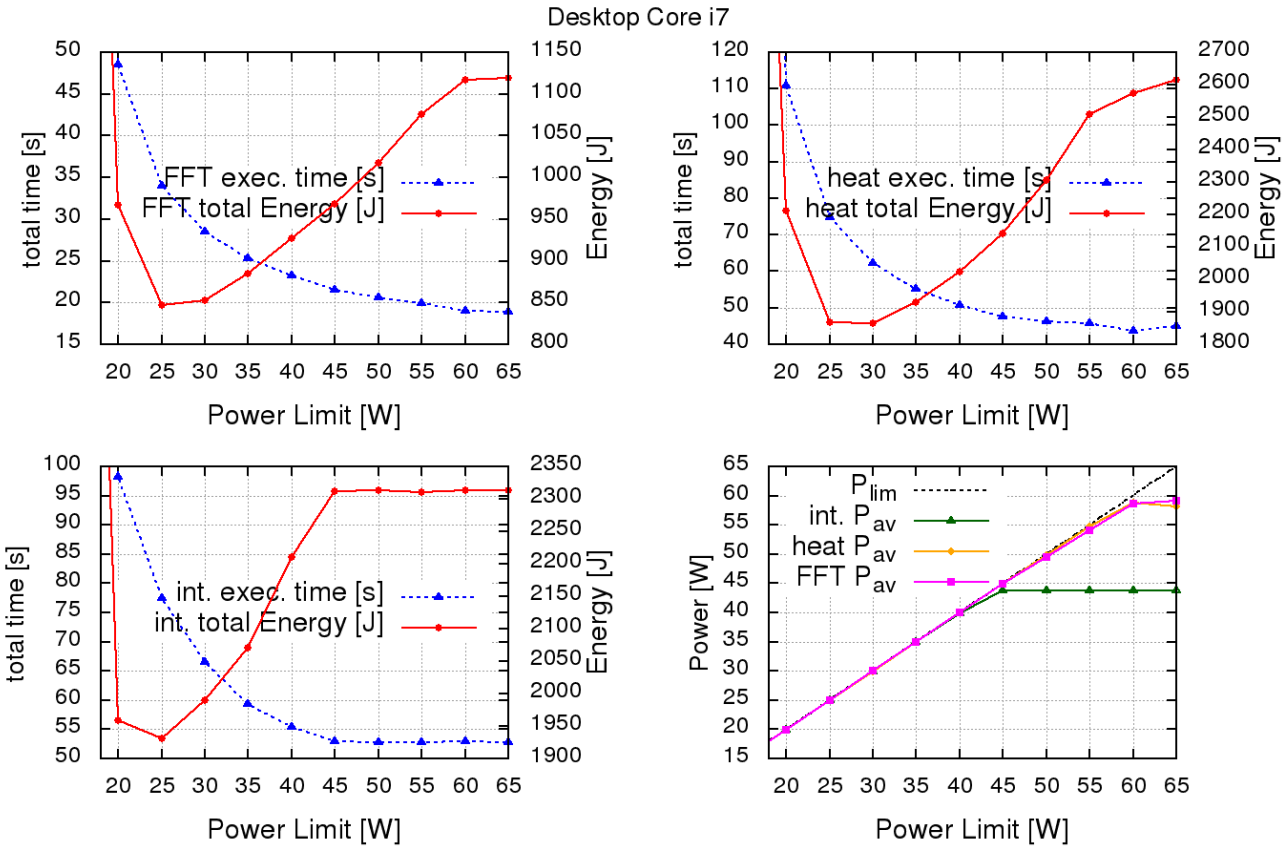


Fig. 4. Tests results for desktop Core™ i7.

TABLE I
TESTBED SYSTEMS USED IN EXPERIMENTS.

System	Processor	Base Frequency	Physical Cores	Logical Cores	Architecture	Cache	RAM
Xeon E5 v4	Intel® Xeon® E5-2620 v4	2.10 GHz	2 x 8	32	Broadwell	2 x 20 MB	128 GB
Xeon Phi x200	Intel® Xeon Phi™ 7210	1.30 GHz	64	256	Knights Landing	32 MB	256 GB
Mobile Core i7	Intel® Core™ i7-5500U	2.40 GHz	2	4	Broadwell	4 MB	16 GB
Desktop Core i7	Intel® Core™ i7-7700	3.60 GHz	4	8	Kaby Lake	8 MB	16 GB

for desktop PC are presented in the last figure, Figure 4. In both mobile and desktop systems the proposed applications seem to generate reasonable load and compared to the server testbed systems much more of available power is used. The level of power consumption is the highest for heat distribution simulation and the lowest for numerical integration. Both systems respect the preset limit well.

For the last two testbed systems we finally observed significant energy consumption benefits caused by limiting the power. The most efficient cases allow to save 25-28% of energy using the Mobile Core i7 system and 16-29% of energy using the Desktop Core i7 system. Of course, as we expected, with gain on energy savings we increase execution time. However, the time loss is much higher than the energy savings. For Mobile Core i7 system execution time increased by 59-86% and for the Desktop Core i7 system the time increase was in range of 38-80%. However, while considering only minimal energy points the time loss might suggest that power limiting is unreasonable, other low energy points with much better performance can be found. If we consider execution time against power limit we can observe that time grows non-linearly with a linear decrease of power limit and the energy consumption has several points besides the minimum in a region below e.g. 20% of energy saved. In such a situation we can search for much better performance with a power limit slightly higher and energy savings level similar to the best possible result. An examplary illustration of the proposed approach could be seen in the results of FFT tests in Figure 4 (top left) where the minimum of energy was obtained for the power limit 25W but for the 30W limit we are able to obtain as good energy savings as in the best case (around 24%) but the time loss drops from 79% to 50%.

D. Conclusions

The results of experiments with limiting the power we proposed and executed on selected testbed systems can lead to several conclusions. First of all, the RAPL driver is able to limit the average power consumption for each of testbed systems and the systems respect the enforced power limit when set between minimal and maximal value. In the experiments we focused on lowering the power consumption and measuring the performance (execution time) and energy consumption during test application runs. We selected the most energy efficient power limit settings and compared the results with the reference values with non-limited (reference) test runs. Table II collects the aforementioned data and correlates them with testbed systems and testbed applications.

The data collected together allows not only for answering the concerns that was a goal of this article but it is possible to compare the performance of the systems and energy efficiency between each other as well. For the testbed applications selected by us for the experiments the best performing system for 2 out of 3 applications was Xeon Phi x200. On the other hand the most energy efficient system was also a server dedicated processor but Xeon E5 v4. Both server systems showed that for such testbed applications the power consumption limiting gives no or unsignificant results of energy saving.

The other pair of testbed systems based on Mobile and Desktop Core i7 processors proved that power consumption limiting can result in significant energy savings but, what is expected, we have to take the loss of performance into account. For the most energy efficient settings which offer between 16% and 29% of energy savings the performance loss is between 38% and 86%. One more conclusion when looking at the power utilisation for the Mobile and Desktop Core i7 systems is that when the testbed application is able to make use of more available power when no limits are set, the better are results of lowering the energy consumption. We can assume that if we had another testbed application that would be able to exploit more of the TDP of our server testbed systems we could probably observe better energy saving results.

V. FINAL REMARKS AND FUTURE WORK

The paper presented the experiments measuring the electrical energy consumption under a set of power caps for three representative HPC applications and four different processors. For some of CPU-application pairs the result analysis shows the existence of energy minima where the power capping provides significant savings — up to 28.8% for Desktop Core i7 executing the Heat Distribution simulation (see Table II for more details).

The future works are going to cover the following issues:

- analysis of the trade-off to find out potential points where values for measures incorporating execution time and energy used would be optimal for a specific application,
- benchmarking other applications, especially those that take more power from our testbed systems,

TABLE II
SUMMARY OF RESULTS PRESENTING MINIMAL ENERGY CASE FOR EACH EXPERIMENT.

		Testbed application											
		Fast Fourier Transform				Heat Distribution				Numerical Integrate			
		E [J]	P_{lim} [W]	P_{av} [W]	t [s]	E [J]	P_{lim} [W]	P_{av} [W]	t [s]	E [J]	P_{lim} [W]	P_{av} [W]	t [s]
	Reference	694.4	170.0	64.1	21.7	1282.1	170.0	66.3	38.7	703.1	170.0	55.8	25.2
Xeon	Best case	674.7	60.0	52.1	25.9	1282.1	170.0	66.3	38.7	703.1	170.0	55.8	25.2
E5 v4	Difference	-2.8%	-64,7%	-18.6%	19.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Reference	1266.0	215.0	149.1	8.5	4623.4	215.0	158.3	29.2	4,605.9	215.0	125.9	36.6
Xeon Phi	Best case	1257.2	140.0	130.4	9.6	4482.6	140.0	132.4	33.9	4,605.9	215.0	125.9	36.6
x200	Difference	-0.7%	-34.9%	-12.5%	13.5%	-3.0%	-34.9%	-16.3%	15.8%	0.0%	0.0%	0.0%	0.0%
	Reference	966.5	15.0	14.8	65.3	2008.5	15.0	14.9	134.9	999.2	15.0	12.6	79.4
Mobile	Best case	700.5	6.0	6.0	117.4	1502.2	7.0	7.0	215.7	730.6	5.0	5.0	147.0
Core i7	Difference	-27.5%	-60.0%	-59.7%	79.9%	-25.2%	-53.3%	-53.2%	59.9%	-26.9%	-66.7%	-60.5%	85.1%
	Reference	1119.6	65.0	59.2	18.9	2616.1	65.0	58.2	44.9	2,313.9	65.0	43.8	52.9
Desktop	Best case	847.2	25.0	25.0	33.9	1863.8	30.0	29.9	62.2	1,931.4	25.0	25.0	77.4
Core i7	Difference	-24.3%	-61.5%	-57.8%	79.5%	-28.8%	-53.8%	-48.6%	38.5%	-16.5%	-61.5%	-43.0%	46.4%

- power-aware modeling of compute devices in frameworks for simulation of application runs in high performance computing environments such as MERPSYS [23],
- development of a tool for automatic detection of the optimal power settings for the aforementioned time-energy measures using historical data (e.g. via machine learning),
- proposing a new method for minimizing the electrical energy usage dynamically at runtime for various HPC/cloud workloads [24].

We assume that the expectations of the IT industry will generate a high demand for green computing methods used for exchanging time of computations into savings in the energy consumption (e.g. dedicated for off-pick hours of data centers). Thus, we hope that our work will stimulate even more research on the subject.

REFERENCES

[1] M. Avgerinou, P. Bertoldi, and L. Castellazzi, "Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency," *Energies*, vol. 10, no. 10, 2017. doi: 10.3390/en10101470. [Online]. Available: http://www.mdpi.com/1996-1073/10/10/1470

[2] H. Krawczyk, M. Nykiel, and J. Proficz, "Mobile offloading framework: Solution for optimizing mobile applications using cloud computing," in *Computer Networks*, P. Gaj, A. Kwiecień, and P. Stera, Eds. Cham: Springer International Publishing, 2015. ISBN 978-3-319-19419-6 pp. 293–305.

[3] K. N. Khan, M. Hirki, T. Niemi, J. K. Nurminen, and Z. Ou, "Rapl in action: Experiences in using rapl for power measurements," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 3, no. 2, pp. 9:1–9:26, Mar. 2018. doi: 10.1145/3177754. [Online]. Available: http://doi.acm.org/10.1145/3177754

[4] B. Subramaniam and W. Feng, "Towards energy-proportional computing using subsystem-level power management," *CoRR*, vol. abs/1501.02724, 2015. [Online]. Available: http://arxiv.org/abs/1501.02724

[5] C. Jin, B. R. de Supinski, D. Abramson, H. Poxon, L. DeRose, M. N. Dinh, M. Endrei, and E. R. Jessup, "A survey on software methods to improve the energy efficiency of parallel computing," *The International Journal of High Performance Computing Applications*, vol. 31, no. 6, pp. 517–549, 2017. doi: 10.1177/1094342016665471. [Online]. Available: https://doi.org/10.1177/1094342016665471

[6] P. Czarnul, J. Kuchta, P. Rosciszewski, and J. Proficz, "Modeling energy consumption of parallel applications," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016, pp. 855–864.

[7] J. Proficz and P. Czarnul, "Performance and Power-Aware Modeling of MPI Applications for Cluster Computing," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9574, pp. 199–209. ISBN 9783319321516. [Online]. Available: http://link.springer.com/10.1007/978-3-319-32152-3_19

[8] D. Abdurachmanov, P. Elmer, G. Eulisse, R. Knight, T. Niemi, J. K. Nurminen, F. Nyback, G. Pestana, Z. Ou, and K. Khan, "Techniques and tools for measuring energy efficiency of scientific software applications," *Journal of Physics: Conference Series*, vol. 608, no. 1, p. 012032, 2015. [Online]. Available: http://stacks.iop.org/1742-6596/608/i=1/a=012032

[9] H. David, E. Gorbatov, U. R. Hanebutte, R. Khanna, and C. Le, "Rapl: Memory power estimation and capping," in *Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '10. New York, NY, USA: ACM, 2010. doi: 10.1145/1840845.1840883. ISBN 978-1-4503-0146-6 pp. 189–194. [Online]. Available: http://doi.acm.org/10.1145/1840845.1840883

[10] S. Desrochers, C. Paradis, and V. M. Weaver, "A validation of dram rapl power measurements," in *Proceedings of the Second International Symposium on Memory Systems*, ser. MEMSYS '16. New York, NY, USA: ACM, 2016. doi: 10.1145/2989081.2989088. ISBN 978-1-4503-4305-3 pp. 455–470. [Online]. Available: http://doi.acm.org/10.1145/2989081.2989088

[11] A. Mazouz, B. Pradelle, and W. Jalby, "Statistical validation methodology of cpu power probes," in *Revised Selected Papers, Part I, of the Euro-Par 2014 International Workshops on Parallel Processing - Volume 8805*. New York, NY, USA: Springer-Verlag New York, Inc., 2014. doi: 10.1007/978-3-319-14325-5_42. ISBN 978-3-319-14324-8 pp. 487–498. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14325-5_42

[12] M. Hirki, "Energy and performance profiling of scientific computing; tieteellisen laskennan energia- ja suorituskykyprofilointi," G2 Pro gradu, diplomity, 2015. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-201512165699

[13] M. Hirki, Z. Ou, K. N. Khan, J. K. Nurminen, and T. Niemi, "Empirical study of the power consumption of the x86-64 instruction decoder," in *USENIX Workshop on Cool Topics on Sustainable Data Centers (CoolDC 16)*. Santa Clara, CA: USENIX Association, 2016. [Online]. Available: https://www.usenix.org/conference/cooldc16/workshop-program/presentation/hirki

[14] H. Zhang and H. Hoffmann, "Maximizing performance under a power cap: A comparison of hardware, software, and hybrid techniques," in *Proceedings of the Twenty-First International Conference on*

Architectural Support for Programming Languages and Operating Systems, ser. ASPLOS '16. New York, NY, USA: ACM, 2016. doi: 10.1145/2872362.2872375. ISBN 978-1-4503-4091-5 pp. 545–559. [Online]. Available: http://doi.acm.org/10.1145/2872362.2872375

[15] F. Sun, H. Li, Y. Han, G. Yan, and J. Ma, "Powercap: Leverage performance-equivalent resource configurations for power capping," in *2016 Seventh International Green and Sustainable Computing Conference (IGSC)*, Nov 2016. doi: 10.1109/IGCC.2016.7892618 pp. 1–8.

[16] Q. Zhu, B. Wu, X. Shen, L. Shen, and Z. Wang, "Co-run scheduling with power cap on integrated cpu-gpu systems," in *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, May 2017. doi: 10.1109/IPDPS.2017.124 pp. 967–977.

[17] M. Travers, "Cpu power consumption experiments and results analysis of intel i7-4820k," uSystems Research Group, School of Electrical and Electronic Engineering, Newcastle University, UK, Tech. Rep. NCL-EEE-MICRO-TR-2015-197, 2015, http://async.org.uk/tech-reports/NCL-EEE-MICRO-TR-2015-197.pdf.

[18] K. Pedretti, S. L. Olivier, K. B. Ferreira, G. Shipman, and W. Shu, "Early experiences with node-level power capping on the cray xc40 platform," in *Proceedings of the 3rd International Workshop on Energy Efficient Supercomputing*, ser. E2SC '15. New York, NY, USA: ACM, 2015. doi: 10.1145/2834800.2834801. ISBN 978-1-4503-3994-0 pp. 1:1–1:10. [Online]. Available: http://doi.acm.org/10.1145/2834800.2834801

[19] A. Krzywaniak and P. Czarnul, "Parallelization of selected algorithms on multi-core cpus, a cluster and in a hybrid cpu+xeon phi environment," in *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology - ISAT 2017 - Part I, Szklarska Poręba, Poland, September 17-19, 2017*, ser. Advances in Intelligent Systems and Computing, L. Borzemski, J. Swiatek, and Z. Wilimowska, Eds., vol. 655. Springer, 2017. doi: 10.1007/978-3-319-67220-5_27. ISBN 978-3-319-67219-9 pp. 292–301. [Online]. Available: https://doi.org/10.1007/978-3-319-67220-5_27

[20] "OpenMP home," URL: https://www.openmp.org/, accessed: 2018-05-11.

[21] J. Sanders and E. Kandrot, *CUDA by Example: An Introduction to General-Purpose GPU Programming*, 1st ed. Addison-Wesley Professional, 2010. ISBN 0131387685, 9780131387683

[22] M. Balducci, A. Choudary, and J. Hamaker, "Comparative analysis of FFT algorithms in sequential and parallel form," Tech. Rep., 1996.

[23] P. Czarnul, J. Kuchta, M. Matuszek, J. Proficz, P. Rosciszewski, M. Wojcik, and J. Szymanski, "Merpsys: An environment for simulation of parallel application execution on large scale hpc systems," *Simulation Modelling Practice and Theory*, vol. 77, pp. 124 – 140, 2017. doi: https://doi.org/10.1016/j.simpat.2017.05.009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1569190X17300916

[24] P. Orzechowski, J. Proficz, H. Krawczyk, and J. Szymanski, "Categorization of cloud workload types with clustering," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, D. K. Lobiyal, D. P. Mohapatra, A. Nagar, and M. N. Sahoo, Eds. New Delhi: Springer India, 2017. ISBN 978-81-322-3592-7 pp. 303–313.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 347–350
DOI: 10.15439/2018F60
ISSN 2300-5963 ACSIS, Vol. 15

Acceleration of 3D ECT image reconstruction in heterogeneous, multi-GPU, multi-node distributed system

Michał Majchrowicz*, Paweł Kapusta†, Lidia Jackowska-Strumiłło‡ and Dominik Sankowski§
Lodz University of Technology
Institute of Applied Computer Science
ul. Stefanowskiego 18/22, Łódź, Poland
* Email: mmajchr@iis.p.lodz.pl † Email: pawel.kapusta@p.lodz.pl ‡ Email: lidia_js@iis.p.lodz.pl
§ Email: dsan@iis.p.lodz.pl

***Abstract*—Electrical Capacitance Tomography (ECT) is an effective and non-invasive visualization technique, which is used in many industrial applications. Unfortunately, image reconstruction in 3D ECT is a complex computational task requiring operations on large size matrices. In this paper, a new approach to 3D ECT image reconstruction is proposed. A new heterogeneous, multi-GPU, multi-node distributed system has been developed, with a framework for parallel computing and a special plug-in dedicated to ECT.**

I. Introduction

ELECTRICAL Capacitance Tomography (ECT) is a measurement technique that can be used for non-invasive monitoring of industrial processes in 2D [7], 3D [1] and even 4D dynamic mode. ECT is performing the task of imaging of materials with a contrast in dielectric permittivity by measuring capacitance from a set of electrodes placed around the investigated object.

In order to achieve a high quality of 3D image, complex reconstruction algorithms performing many matrix calculations have to be applied. Therefore different solutions accelerating these calculation have been reported in the past by the Authors[8][14], especially these dealing with sparse matrices and Finite Elements Method [9] as well as neural networks approach [5][6] and even fuzzy logic [21].

In this work we propose a novel heterogeneous, multi-GPU (Graphics Processing Unit), multi-node distributed system, with a framework for parallel computing and a special plug-in dedicated to ECT. The system features and its efficiency have been compared to the previously developed distributed system based on the Xgrid platform.

A. Image reconstruction in ECT

The scheme of image synthesis in Electrical Capacitance Tomography is called image reconstruction. It is based on solving the so called inverse problem, in which the spatial distribution of electric permittivity from the measured values of capacitance C is approximated [1] [16].

Image reconstruction using deterministic methods requires execution of a large number of basic operations of linear algebra, such as transposition, multiplication, addition and subtraction [10][15]. Matrix calculations for a large number of elements is characterized by a high computational load. Matrix multiplication is a key operation in ECT imaging and therefore many researchers decided even to build a custom hardware for this purpose.

The LBP algorithm is one the most used reconstruction algorithms, even though it is characterized by low spatial resolution. Nevertheless it is not as computationally complex as other solutions. Moreover there is still active research on improving it's characteristics [17]. It is based on the following equation [3]:

$$\varepsilon = \mathbf{S}\boldsymbol{C}_m \tag{1}$$

where:
ε - electric permittivity vector (output image),
$\mathbf{S}$ - sensitivity matrix,
$\boldsymbol{C}_m$ - capacitance measurements vector.

The Landweber algorithm is based on the following iterative equation:

$$\varepsilon_{k+1} = \varepsilon_k - \alpha\mathbf{S}^T(\mathbf{S}\varepsilon_k - \boldsymbol{C}_m) \tag{2}$$

where:
ε_{k+1} - image obtained in current iteration,
ε_k - image from the previous iteration,
α - convergence factor (scalar),
$\mathbf{S}^T$ - sensitivity matrix, transposed,
$\mathbf{S}$ - sensitivity matrix,
$\boldsymbol{C}_m$ - capacitance measurements vector.

In the case of the Landweber algorithm each iteration improves the overall quality of the output image. As a result acceleration of image reconstruction process is a very important issue. Nevertheless, due to its nature it is necessary to exchange the data (ε_{k+1}) in every iteration.

II. Design assumptions

As a result of the earlier performed studies [8][13] the Authors have developed a new distributed system dedicated to ECT computations. It is specially designed to accelerate matrix computations that are a crucial part of reconstruction

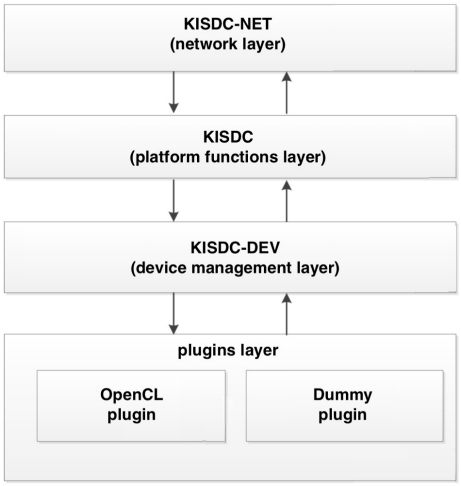


Fig. 1. Activity Diagram for performing calculations using the KISDC platform

algorithms used in ECT [9]. The earlier developed solution was based on the Xgrid platform, used as a network layer. However, the analysis of this system showed the limitations of this solution, and the main conclusion from the previous research [14] was, that the new software for the system should be developed.

A special framework was designed and built that provides software tools needed both for the system architecture expansion and new algorithms development and implementation in a distributed heterogeneous environment. The proposed approach allows for a greater flexibility of the developed solutions, provides tools for their easy testing and enables further acceleration of ECT image reconstruction.

The framework was designed to ensure an efficient use of the computing power of all the devices in which the nodes are equipped. This architecture is scalable and allows users

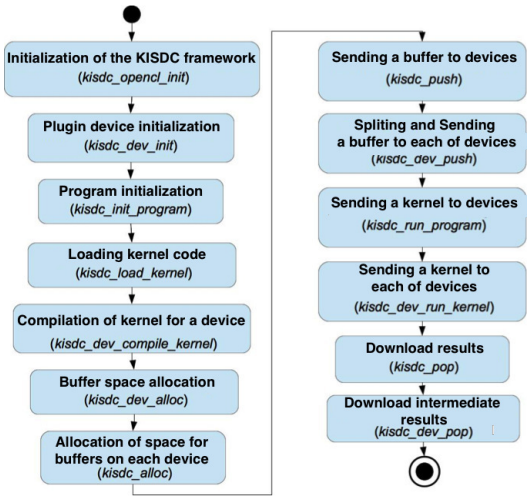


Fig. 2. Activity Diagram for performing calculations using the KISDC platform

to expand the computing power of the system by adding more nodes. The above assumptions pose many challenges in the architecture of the system itself, but their application makes it straightforward to use the environment to speed up computations in existing projects, thus testing and developing new distributed algorithms is much faster.

The system was designed as a modular, layered architecture (Fig. 1). This approach allows limiting the dependencies between the individual modules. Moreover, thanks to this architecture, it is possible to abstract the compute devices using KISDC-DEV module, that hides the type of the hardware from the user and makes all the algorithm written using the provided Application Programming Interface (APIs) hardware-agnostic

Expansion of the computing power of the system is possible through the use of "plug-in" architecture (by adding support for new devices, such as FPGAs). The basic operations of linear algebra were implemented in the system as a set of functions in the form of an API.

```
err = clGetDeviceIDs(platform, gpu ? CL_DEVICE_TYPE_GPU : CL_DEVICE_TYPE_CPU,
    numberOfDevices, &device_id, NULL);
context = clCreateContext(0, 1, &device_id, NULL, NULL, &err);

commands = clCreateCommandQueue(context, device_id, 0, &err);
program = clCreateProgramWithSource(context, 1, (const char **) & KernelSource,
    NULL, &err);

err = clBuildProgram(program, 0, NULL, NULL, NULL, NULL);

kernel = clCreateKernel(program, "matrixmul", &err);


A = clCreateBuffer(context,  CL_MEM_READ_ONLY,  sizeof(float) * count1, NULL, NULL)
    ;
B = clCreateBuffer(context,  CL_MEM_READ_ONLY,  sizeof(float) * count2, NULL, NULL)
    ;
C = clCreateBuffer(context, CL_MEM_WRITE_ONLY, sizeof(float) * count2, NULL, NULL);

err = clEnqueueWriteBuffer(commands, A, CL_TRUE, 0, sizeof(float) * count1, data, 0
    , NULL, NULL);
err = clEnqueueWriteBuffer(commands, B, CL_TRUE, 0, sizeof(float) * count2, data, 0
    , NULL, NULL);

err  = clSetKernelArg(kernel, 0, sizeof(cl_mem), &A);
err |= clSetKernelArg(kernel, 1, sizeof(cl_mem), &B);
err |= clSetKernelArg(kernel, 2, sizeof(cl_mem), &C);

err = clGetKernelWorkGroupInfo(kernel, device_id, CL_KERNEL_WORK_GROUP_SIZE, sizeof
    (local), &local, NULL);

global = count1;
err = clEnqueueNDRangeKernel(commands, kernel, 1, NULL, &global, &local, 0, NULL,
    NULL);

clFinish(commands);
err = clEnqueueReadBuffer( commands, output, CL_TRUE, 0, sizeof(float) * count3,
    results, 0, NULL, NULL );
```

Fig. 3. Simplified code fragment that multiplies the matrix using the OpenCL libraries and single GPU

The use of a heterogeneous system for distributed computing in ECT required the implementation of a series of algorithms without which the proposed system would not work properly. The most important of these are as follows: division of matrices between nodes, basic operations of linear algebra (transposition, addition, subtraction, multiplication), data transfer between nodes, planning and division of tasks, support for heterogeneous devices, support for calculations using graphics cards, supports modern multi-core processors as a set of devices and a possibility to extend existing solutions with pseudo inheritance from implemented layouts.

A. Application Programming Interface

An important aspect of the designed environment is also the API, which greatly simplifies the performance of matrix calculations in a distributed heterogeneous environment. Even for a single computer configuration, the KISDC framework makes it possible to significantly simplify the code (Fig. 4) in

comparison to amount of code required when using another solution, such as for example OpenCL (Fig. 3).

```
pKISDC *KISDC;
KISDC = kisdc_opencl_init();

KISDC->DEV = KISDC->kisdc_dev_init();
KISDC->dummyProgram = KISDC->kisdc_init_program(KISDC,kernel);

A = KISDC->kisdc_alloc(KISDC,SIZE_A_1,SIZE_A_2,1,0);
B = KISDC->kisdc_alloc(KISDC,SIZE_B_1,SIZE_B_2,1,0);
C = KISDC->kisdc_alloc(KISDC,SIZE_C_1,SIZE_C_2,1,0);

KISDC->kisdc_push(KISDC,A,Abuf);
KISDC->kisdc_push(KISDC,B,Bbuf);

paramsArray = (pMemBuf**)malloc(5*sizeof(pMemBuf*));

paramsArray[0] = A;
paramsArray[1] = B;

KISDC->dummyProgram->args = paramsArray;
KISDC->dummyProgram->argc = 2;
KISDC->dummyProgram->out = C;

KISDC->kisdc_run_program(KISDC,KISDC->dummyProgram);
Cbuf = KISDC->kisdc_pop(KISDC,C);
```

Fig. 4. A fragment of code that multiplies the matrix using the KISDC platform and multiple GPUs

The code examples listed in Fig. 4 and Fig. 3 involve a very simple scenario of using GPUs to calculate the product of two matrices (for example in the LBP algorithm). In the case of the algorithms of a higher complexity, the difference in the code sizes for both solutions will be larger and the KISDC implementation will have even a bigger advantage over OpenCL.

Moreover, if instead of one, three or more GPUs are used for computations, a significant changes in the code must be done in each case, when using the CUDA [11] or OpenCL library [19]. Whereas, with the KISDC framework, no changes are necessary. However, this small number of lines written by the programmer corresponds to hundreds of lines of code within the KISDC system.

The KISDC architecture simplifies also performance tests of the developed image reconstruction algorithms in Electrical Capacitance Tomography in various hardware configurations and to implement these algorithms in a distributed system. As shown in the activity diagram (Fig. 2), the process of

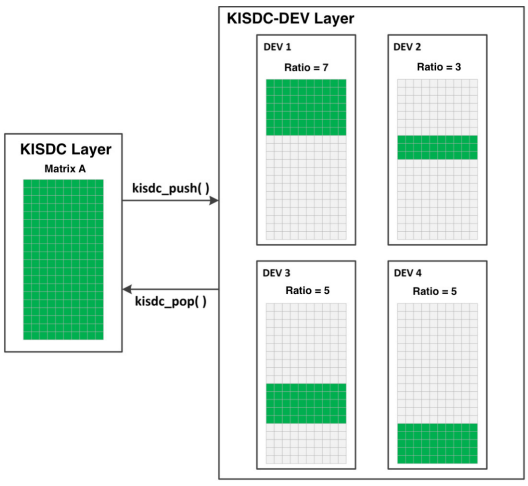


Fig. 5. Data distribution using KISDC platform

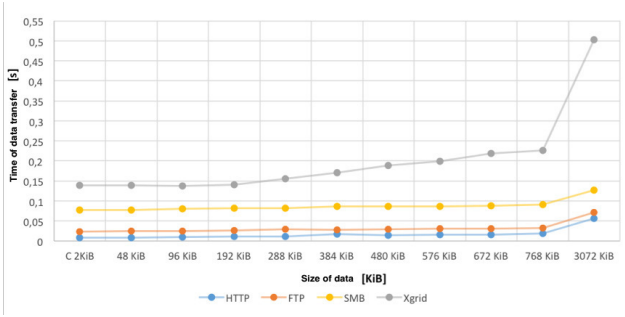


Fig. 6. Comparison of average data transmission times for different network protocols

performing calculations using the KISDC system allows for much more flexibility in the number and type of devices used. This was possible by adopting significant flexibility in the data distribution layer (Fig.5).

III. Test Results

In the previously built distributed system a ready Xgrid platform was applied as a network layer [14]. In the current work the author's KISDC system with KISDC-NET network layer was designed and implemented. While designing the KISDC-NET layer, the existing network protocols were applied and tested in advance in order to choose the best solution.

The network characteristics of the previously developed solution based on the Xgrid system was compared with the new system using other data distribution protocols: HTTP (Hypertext Transfer Protocol), FTP (File Transfer Protocol) and SMB (Server Message Block). On the Basis of the obtained results (see Fig. 6) HTTP protocol has been selected as the best one for the KISDC-NET.

Both distributed systems: the one based on the Xgrid platform and the KISDC have been extensively tested and compared. In both cases the hardware was identical, consisting of two nodes of high computing power, both using 8 thread Intel i7 930 CPUs and Nvidia GPUs (Tesla S1070 + Tesla

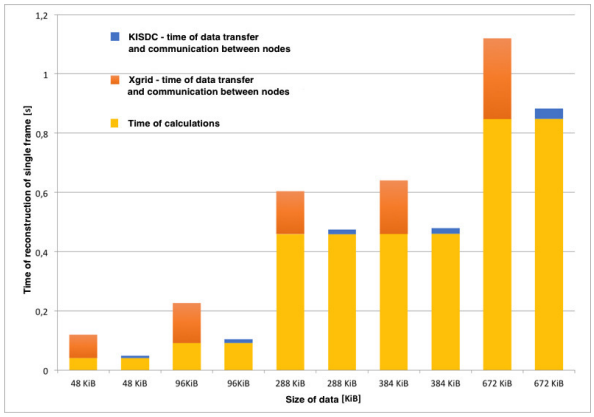


Fig. 7. Comparison of image reconstruction time in Xgrid and KISDC systems for dual computer configuration

C2070 compute devices in the first node and dual GTX 570 in the second).

The comparison of times of a single frame reconstruction in the two node system are shown in Figure 7. Yellow color represents calculation time (the same for the both systems), blue color is related to data transfer time for the KISDC system, and orange color denotes data transfer time for the Xgrid system. For each of the analyzed data sizes, the speed up of image reconstruction expressed in the number of reconstructed frames per second was noted. The most significant acceleration was achieved for 48 KiB and 96 KiB image vectors.

IV. Conclusions

A flexible, distributed computing system for tomographic image reconstruction called KISDC has been designed and developed. The system's framework allows to accelerate any kind of computation dealing with a basic linear algebra operations. However, it should be noted that the KISDC is highly scalable and can be easily extended either by specific OpenCL kernel or by a plug-in providing support for a special kind of calculations.

The work described in this paper was focused on improvement of data management in the distributed system and on reducing delays in the data transmission over the computer network. The comparison of times of data transfer and communication between the nodes shows very clearly that the use of the new developed system with HTTP protocol ensures much better results than with the Xgrid platform. It is also evident that the KISDC system allowed for a significant reduction of the total time of a single frame reconstruction and a major speed up in implementations of both the LBP and the Landweber algorithms.

Acknowledgment

This work was financed by the Lodz University of Technology, Faculty of Electrical, Electronic, Computer and Control Engineering as a part of statutory activity (project no. 501/12-24-1-5418)

References

[1] Banasiak, R., Wajman, R., Fidos, H., Jaworski, T., Fiderek, P., Kapusta, P., Majchrowicz, M., Sankowski, D., "Fusion of three-dimensional electrical capacitance tomography and fuzzy logic inference for phases fraction and flow structures identification in horizontal and vertical gas-liquid flow pipelines.," *7th World Congress on Industrial Process Tomography,* Kraków, 2013, pp. 818–827.

[2] Barrachina, S., Castillo, M., Igual, F.D., Mayo, R., Quintana-Ortí, E.S., Quintana-Ortí, G., "Exploiting the capabilities of modern GPUs for dense matrix computations.," *Concurrency and Computation: Practice and Experience* 2009.

[3] Cui, Z., Wang, Q., Xue, Q., Fan, W., Zhang, L., Cao, Z., Sun, B., Wang, H., Yang, W., "A review on image reconstruction algorithms for electrical capacitance/resistance tomography," *Sensor Review, Vol. 36 Issue: 4,* 2016, pp. 429–445.

[4] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T. (1999): Hypertext Transfer Protocol – HTTP/1.1. Request For Comments (RFC), 1999, RFC Editor.

[5] Garbaa, H., Jackowska-Strumiłło, L, Grudzień, K., Romanowski, A., "Neural network approach to ECT inverse problem solving for estimation of gravitational solids flow," *2014 Federated Conf. on Computer Science and Inf. Systems (FedCSIS), IEEE Xplore Digital Library*, 2014, pp. 19–26.

[6] Garbaa, H., Jackowska-Strumillo, L., Grudzien, K., Romanowski, A., "Application of electrical capacitance tomography and artificial neural networks to rapid estimation of cylindrical shape parameters of industrial flow structure", Archives of Electrical Engineering, Vol. 65(4), 2016, pp. 657–669.

[7] Grudzień K., Chaniecki Z., Romanowski A., Niedostatkiewicz, M., Sankowski, D. "ECT Image Analysis Methods for Shear Zone Measurements during Silo Discharging Process," *Chinese Journal Of Chemical Engineering, Vol. 20 (2)*, 2012, pp. 337-345.

[8] Kapusta, P., Majchrowicz, M., Sankowski, D., Jackowska-Strumiłło, L., Banasiak, R., "Distributed multi-node, multi-GPU, heterogeneous system for 3D image reconstruction in Electrical Capacitance Tomography - network performance and application analysis," *Przegląd Elektrotechniczny, 89 (2 B)*, 2013, pp. 339-342.

[9] Kapusta, P., Majchrowicz, M., Sankowski, D., Jackowska-Strumiłło, L., "Acceleration of image reconstruction in 3D Electrical Capacitance Tomography in heterogeneous, multi-GPU system using sparse matrix computations and Finite Element Method," *2016 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE Xplore Digital Library*, 2016, pp. 679-683,

[10] Kapusta P., Duch P., Jaworski T., Kucharski J., Ślot K., "Generative network input shaping for controlling visual object rendition in Adversarial Networks", Proceedings of International Interdisciplinary PhD Workshop 2017, Lodz, 9-11 Sep. 2017, pp 380–385.

[11] Kirk, D.B., Hwu, W.-M.W., "Programming Massively Parallel Processors," *Morgan Kaufmann* 2010.

[12] Krüger, J., Westermann, R., "Linear algebra operators for GPU implementation of numerical algorithms," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH* 2003.

[13] Majchrowicz, M., Kapusta, P., Wąs, Ł., Wiak, S, "Application of General-Purpose Computing on Graphics Processing Units for Acceleration of Basic Linear Algebra Operations and Principal Components Analysis Method," *Man-Machine Interactions 3, Advances in Intelligent Systems and Computing Volume 242,* Springer International Publishing, 2014, pp. 519–527.

[14] Majchrowicz M., Kapusta P., Jackowska-Strumiłło L., Sankowski, D., "Analysis of Application of Distributed Multi-Node, Multi-GPU Heterogeneous System for Acceleration of Image Reconstruction in Electrical Capacitance Tomography," *Image Processing & Communications, vol. 20, Issue 3*, 2015, pp. 5–14.

[15] Majchrowicz M., Kapusta P., L. Jackowska-Strumiłło, D. Sankowski, "Acceleration of image reconstruction in 3D Electrical Capacitance Tomography in heterogeneous, multi-GPU, multi-node distributed system", Proceedings of International Interdisciplinary PhD Workshop 2017, Lodz, 9-11 Sep. 2017, pp. 164–169.

[16] Sankowski, D., Grudzień, K., Chaniecki, Z., Banasiak, R., Wajman, R., Romanowski, A., "Process tomograhy development at Technical University of Lodz", *Electrical Capacitance Tomography Theoretical Basis and Applications, edited by Dominik Sankowski and Jan Sikora,* Warszawa, 2010, pp. 70-95.

[17] Sun B., Yue S., Cui Z., Wang H., "A new linear back projection algorithm to electrical tomography based on measuring data decomposition", *Measurement Science and Technology, Vol. 26, Number 12*, 2015, pp. 1270-1283

[18] Soleimani, M., Mitchell, C.N., Banasiak, R., Wajman, R., Adler, A., "Four-dimensional electrical capacitance tomography imaging using experimental data," *Progress In Electromagnetics Research PIER, 90,* Hong Kong, 2009, EMW Publishing, pp. 171–186.

[19] Tsuchiyama, R., Nakamura, T., Iizuka, T., Asahara, A., Miki, S., Tagawa, S. (2010): The OpenCL Programming Book. 1st ed. Fixstars Corporation.

[20] Tchorzewski, P., Rymarczyk, T., Sikora, J. " (2016). Using Topological Algorithms to Solve Inverse Problem in Electrical Impedance Tomography.," In: Mikulka, J. (Ed) Proc. International Interdisciplinary PhD Workshop Location: Brno, Czech Republic, 12-15 Sep. 2016, pp. 46-50.

[21] Fiderek, P., Kucharski, J., Wajman, R. (2017) Fuzzy inference for two-phase gas-liquid flow type evaluation based on raw 3D ECT measurement data. Flow Measurement and Instrumentation. Vol. 54, pp. 88-96.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 351–359
DOI: 10.15439/2018F166
ISSN 2300-5963 ACSIS, Vol. 15

An Experimental Analysis on Scalable Implementations of the Alternating Least Squares Algorithm

Dânia Meira
Data Science Retreat/EyeEm
Berlin, Germany
Email: meira.dania@gmail.com

José Viterbo
Institute of Computing
Fluminense Federal University
Niterói, RJ, Brazil
Email: viterbo@ic.uff.br

Flavia Bernardini
Institute of Computing
Fluminense Federal University
Niterói, RJ, Brazil
Email: fcbernardini@ic.uff.br

***Abstract*—The use of the latent factor models technique overcomes two major problems of most collaborative filtering approaches: scalability and sparseness of the user's profile matrix. The most successful realizations of latent factor models are based on matrix factorization. Among the algorithms for matrix factorization, alternating least squares (ALS) stands out due to its easily parallelizable computations. In this work we propose a methodology for comparing the performance of two parallel implementations of the ALS algorithm, one executed with MapReduce in Apache Hadoop framework and another executed in Apache Spark framework. We performed experiments to evaluate the accuracy of generated recommendations and the execution time of both algorithms, using publicly available datasets with different sizes and from different recommendation domains. Experimental results show that running the recommendation algorithm on Spark framework is in fact more efficient, once it provides in-memory processing, in contrast to Hadoop's two-stage disk-based MapReduce paradigm.**

I. Introduction

AMONG the many techniques used to implement recommender systems, collaborative filtering, which is based on comparing the profile of preferences of the users, is a very popular technique in e-commerce applications, due to its good results [1]. Neighborhood-based approaches present scalability problems, given that the algorithm has to process all the data to compute a single prediction. Hence, if there is a large number of users and items, such approaches may not be appropriate for online systems which recommend in real time. Furthermore, these algorithms are more sensitive than model-based to some common problems of recommender systems. One common problem is the sparsity of the matrix that stores the ratings that represent the users' preferences about the available items. This refers to a situation in which transactional or feedback data is sparse and insufficient to identify similarities in users' interests making it difficult and unreliable to predict which consumers are similar [2]. Another recurrent problem in generating recommendations happens when we wish to recommend items that no one in the community has yet rated or interacted with. This is known as the cold-start problem and pure collaborative filtering cannot help in a cold start setting, since no user preference information is available to form any basis for recommendations [3].

Nevertheless, there are models that can help bridge the gap from existing items to new items, by inferring similarities among them. Model-based approaches, instead of directly using the ratings stored, as the neighborhood-based systems, use ratings to learn a predictive model. The model building process is performed by different machine learning algorithms such as Bayesian networks [4], neural networks [5], and Singular Value Decomposition [6]. These approaches tend to be faster in prediction time than the neighborhood-based approaches. However, the construction of the model is a complex task that demands the estimation of a multitude of parameters, and usually requires a considerable amount of time [1].

These problems become more evident when trying to construct recommender systems associated with Websites that have a large number of users and items and, thus, associated with huge databases. Online systems demand high availability and short response time, as they must integrate and quickly process incoming streams of data from all users' activities, in order to generate the recommendations. All this process need to occur with a latency of seconds, as the most promising items selected by the recommendation algorithms have to be showed to the users while they are still browsing the Website. The greater the number of users to serve and items from which to recommend, the greater is the amount of processing required, which increases the time it takes to generate each recommendation. The digital music Spotify platform [7] is a practical example of an online recommender system with high demand: their music personalization service has more than 50 million active users, 30 million cataloged songs and around 20 thousand new songs added per day [8]. Amazon [9] generates recommendations from a database with 253 million products [10] for users of 270 million active customer accounts [11]. An efficient approach is essential in all those cases. Nowadays, to tackle such performance challenges, online recommender systems have combined two strategies: (i) efficient algorithms, that avoid the computational complexity of calculating each of the entries of the high

dimensional and sparse matrix; and (ii) optimized data storage and processing. This means processing real-time information to build a predictive model and present its output in seconds.

In order to solve this problem, some authors have developed a class of model-based collaborative filtering algorithms that are fast and easy to calculate, called latent factor models [12]. They attempt to identify relevant features (latent factors) that explain observed ratings. These features can be interpreted as the preference of the users and the characteristics of the items being recommended. Using these latent factors, it is possible to infer the user's preference and make a recommendation of the better items for him/her. The most successful techniques to perform latent factors modeling are based on matrix factorization [13]. They have become popular recently because they combine scalability and predictive accuracy, and, besides, they offer flexibility for modeling different real situations, being superior to the neighborhood-based methods for producing recommendations because they allow the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels [14]. Recent works suggest modeling only the observed ratings, while avoiding overfitting, through an adequate regularized model [15].

Some parallel algorithms for latent factor models with regularization have been designed aiming at improving the modelling performance. Among them, two can be highlighted: (i) the low-rank matrix factorization with Alternating Least Squares (ALS), which uses a series of broadcast-joins [16], built on top of the open source MapReduce implementation Hadoop [17], and its ecosystem, which we call HadoopMR-Mahout; and (ii) the Alternating Least Squares with Weighted-λ-Regularization (ALS-WR) [18] which has been implemented in Apache Spark's Machine Learning library, MLlib, which we call Spark-MLLib [19]. Scalability and performance are key issues for recommender systems, since computational complexity increases with the number of users and items, but the performance gain for these implementations has not yet been systematically evaluated in any comparative study.

Although ALS Matrix Factorization algorithms are not new, some recent works shows that evaluating solutions that can be faster in specific situations, such as memory restrictions and some other high processing situations that may occur, still needs some attention. Authors of [20] propose some techniques for finding efficient and portable ALS Matrix Factorization for Recommender Systems. They apply thread batching technique and three architecture-specific optimizations for a new solution, and they implement an ALS solver in OpenCL so that it can run on various platforms (CPUs, GPUs, and MICs). Authors of [21] propose a new software solution to improve the performance of Recommender Systems, relying heavily on Apache Spark technology to speed up the computation of recommendation algorithms.

This work aims to conduct an experimental analysis to compare two different scalable implementations of the Alternating Least Squares algorithms (Spark-MLLib and HadoopMR-Mahout) for collaborative filtering recommendation. We performed experiments to evaluate the accuracy of generated recommendations and the execution time of both algorithms, using publicly available datasets with different sizes and from different recommendation domains.

This work is organized as follows. In the next section, we explain the fundamental concepts about model-based approaches for implementing collaborative filtering. In Section 3, we discuss matrix factorization implementations, explaining how parallel implementations improve efficiency of recommender algorithms. In Section 4, we describe the methodology used for the comparative study between the two different implementations of the ALS algorithm, on different recommendation domains and dataset sizes, and present our experimental results for three assessed dimensions: accuracy, efficiency and scalability. In Section 5, we discuss the contributions and limitations of the proposed study, presenting also some topics for future work.

II. Model-Based Collaborative Filtering

The fundamental assumption of CF is that if users X and Y rate n items similarly, or have similar behaviors (e.g., buying, watching, listening), hence they will rate or act on other items similarly. CF techniques use a database of preferences for items by users to predict additional topics or products a new user might like. The problem space can be formulated as a matrix of users versus items, with each cell representing a user's rating on a specific item. This matrix will be referred as ratings matrix from now on.

Under this formulation, the problem is to predict the values for specific empty cells. In collaborative filtering, this matrix is usually very sparse, since each user only rates a small percentage of the total available items. To fill in the missing entries of the ratings matrix, models are learnt by fitting the previously observed ratings. Once the goal is to generalize these observed ratings in a way that allows us to predict future, unknown ratings, caution should be exercised to avoid overfitting the observed data. This can be achieved by modeling the latent factors of the ratings matrix, that is, finding a small set of latent features that explain observed ratings and describe the general characteristics of users and items. The most successful techniques to model latent factors are based on matrix factorization, because they combine scalability and predictive accuracy.

A. Matrix Factorization

Matrix factorization models map both users and items to a joint latent factor space, such that user-item interactions are modeled as inner products in that space. The latent space tries to explain ratings by characterizing both items and users on the same set of factors, which are the characteristics inferred from the observed ratings [18]. The intuition of this method is that it can be equivalent to a summarization. It boils down the world of user preferences for individual items to a world of user preferences for more general and less numerous features (like genre). This is, potentially, a much smaller set of data.

Although this process loses some information, it can sometimes improve recommendation results because this process

smooths the input in useful ways when it generalizes the features that describe the items, making similar what appeared to be distinct at first, thus avoiding overfitting the observed ratings. For example, imagine two car enthusiasts. One loves Corvette, and the other loves Camaro, and they want car recommendations. These enthusiasts have similar tastes: both love a Chevrolet sports car. But in a typical data model for this problem, these two cars would be different items. Without any overlap in their preferences, these two users would be deemed unrelated. However, a matrix factorization based recommender would perhaps find the similarity. The matrix factorization output may contain features that correspond to concepts like Chevrolet or sports car, with which both users would be associated. From the overlap in features, a similarity could be computed. These features correspond to the latent factors, or singular values of the ratings matrix and their correspondence to concepts are not explicit. Also, the exact number of singular values describing a matrix is not previously known, so there is a need to experiment to find the appropriate number of singular values that best summarizes the concepts for a given domain.

Consider a recommender system with m users and n items. Let $R = [r_{ui}]$ be the ratings matrix, where $r_{ui} \in \mathbb{R}^{(m \times n)}$. Matrix factorization models map both users and items to a joint latent factor space of dimensionality k, that is, $\widehat{R}$ is a rank-k approximation of the ratings matrix R. Let $P = [p_u]$ be the user feature matrix, where $p_u \in \mathbb{R}^k$, and $Q = [q_i]$ be the item feature matrix, where $q_i \in \mathbb{R}^k$. So, user-item interactions are modeled as inner products:

$$\widehat{r}_{ui} = q_i^T \times p_u \tag{1}$$

An example of matrix factorization computation is found next on Figure 1. On a system with five users (represented by the upper matrix in the figure) and six items (represented by the left matrix in the figure), the estimation of the rating value of item 3 given by user 4, $\widehat{r}_{43}$, is given by the inner product of the vectors representing item 3, i.e., q_3^T, and user 4 , i.e., p_4.

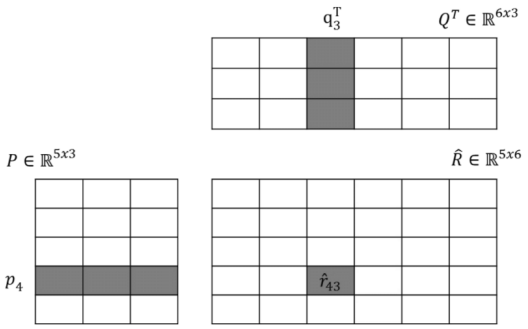


Fig. 1. Sinusoid

The major challenge is to compute the mapping of each item and user to latent factor vectors $q_i, p_u \in \mathbb{R}^k$. The traditional implementation to learn latent factors is the singular value decomposition (SVD), but it suffers from the high portion of missing values in the ratings matrix, because, in general, users have rated only a small set of items [14]. The SVD can be computed one column at a time, whereas for the partially specified case, no such recursive formulation holds [18]. Also, addressing only few known ratings is highly likely to model overfitting [12]. Earlier works relied on imputation [2], [22], which fill in missing ratings and make the ratings matrix dense. However, the data may be considerably distorted by inaccurate imputation and also computing the SVD becomes very expensive after imputation, as it significantly increases the size of the matrices.

More recent works suggested modeling directly only the observed ratings, while avoiding overfitting through an adequate regularized model [15]. This model minimizes the regularized squared error on the set of observed ratings, as shown in Eq. 2, where k is the set of (u, i) pairs for which r_{ui} is known (the training set). The system learns the model by fitting the previously observed ratings. However, solving Eq. 2 with many parameters, when k is relatively large, from a sparse dataset usually overfits the data. The overfitting is avoided by regularizing the learning parameters, whose magnitudes are penalized by the λ constant [23]. This is also known as the Tikhonov regularization [24]. Eq. 2 is solved using a learning algorithm such as the alternating least squares (ALS) [18], which is the focus of this work.

$$min(p, q) \sum_{(u,i) \in k} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2) \tag{2}$$

B. Alternating Least Squares (ALS)

Eq. 2 is not convex when both q_i and p_u are unknown. However, fixing one of them turns the optimization into a quadratic problem that can be solved. So, ALS technique rotates between fixing the q_i's and p_u's. When all p_u's are fixed, the system recomputes the q_i's by solving a least-squares problem, and vice versa. This ensures that each step decreases Eq. 2 until convergence.

Although it is computationally more expensive than Stochastic Gradient Descent (SGD), ALS implementation is favorable in at least two cases. The first is when dealing with densely filled matrices, as such in systems centered on implicit data. Because the training set cannot be considered sparse, looping over each single training case (as in the case of SGD) would not be practical. The second case is when the system can use parallelization. The algorithm computes each q_i independently of the other item factors and computes each p_u independently of the other user factors, which allows for massive parallelization of the implementation [18].

When re-computing the user feature matrix P for example, p_i, the i-th row of P, can be re-computed by solving a least squares problem only including r_i, the i-th row of R, which holds user i's interactions, and all the columns q_j of Q that correspond to non-zero entries in q_i. This re-computation of p_i is independent from the re-computation of all other rows of P and therefore, the re-computation of P is easy to parallelize if efficient data access to the rows of R and the corresponding

columns from Q is effectively managed. The sequence of re-computing of P followed by re-computing Q is referred to as a single iteration in ALS. Algorithm 1 summarizes the steps of the ALS algorithm.

From a data processing perspective, computing ALS means that a parallel join occurs between the interaction data R and Q (the item features) in order to re-compute the rows of P. Analogously, a parallel join is conducted between R and P (the user features) to re-compute Q. Finding an efficient execution strategy for these joins is crucial to the performance of any parallel solution, since the required amount of inter-machine communication, as network bandwidth is the scarcest resource in a cluster.

III. Matrix Factorization Implementations

In this section, we discuss previous works that propose matrix factorization implementations to solve the recommendation problem. We describe each of the proposed implementations and the performance results obtained.

A. Traditional Implementations

The Netflix Prize, a competition that began in October 2006, has motivated the progress in the field of collaborative filtering. The nature of the competition has encouraged rapid development, where innovators built on each generation of techniques to improve prediction accuracy. In September 2009 the prize was awarded to the BellKor's Pragmatic Chaos team that managed to achieve the winning RMSE of 0.8567 on the test subset, which represents a 10.06% improvement over Cinematch, Netflix's own recommendation algorithm.

The recommendation strategy used by the winning solution was an ensemble of more than 100 different predictor sets, the majority of which are factorization models, learned by stochastic gradient descent (SGD), applied directly on the raw data.

For single machine implementations, SGD is the preferred technique to compute a low-rank matrix factorization, because it is easy to implement and computationally less expensive then ALS. Unfortunately, SGD is inherently sequential, because it updates the model parameters after each processed interaction. Techniques for parallel SGD have been proposed, yet they are either hard to implement, exhibit slow convergence or require shared-memory.

The SGD implementation used in this solution is described by [25] as possible to be executed to factorize the 17,000 x 500,000 matrix with 40 latent factors on 2G of RAM, a C compiler, and good programming habits. But in the paper describing the winning solution, he did not specify the environment nor the performance of the algorithm, as this was not important for the prize. The algorithms could run for as many as long as needed, since the only evaluated metric was the RMSE.

Finally, in 2012 Netflix announced that they did not implement the Netflix Prize solution algorithm, and they gave two reasons for that. The first reason is that the new methods were evaluated off-line but the additional accuracy gains measured did not seem to justify the engineering effort needed to bring them to a production environment. Also, their focus on improving personalization had shifted since 2007, just a year after the beginning of the competition, when Netflix streaming service was launched. From DVDs to an online streaming service, Netflix as a whole changed dramatically, not only the way the users interact with the service but also the types of data available to use in the algorithms.

As of 2012, Netflix reported having more than 23 million subscribers in 47 countries. Those subscribers streamed 2 billion hours from hundreds of different devices in the last quarter of 2011. Every day they add 2 million movies and TV shows to the queue and generate 4 million ratings. They have adapted their recommendation algorithm to this new scenario, and 75% of what people watch is from some sort of recommendation. This new strategy still runs the learning algorithm in batch, as briefly discussed in the Large-Scale Recommendation Systems Workshop on the ACM Conference Series on Recommender Systems in 2013, held at Hong Kong.

B. Parallel Implementations

Another team participating in the Netflix Prize proposed, in 2008, a parallel implementation of matrix factorization, called the Alternating-Least-Squares with Weighted-λ-Regularization (ALS-WR) [18]. This solution was motivated by two main reasons: the size of the dataset, which was 100 times larger than previous benchmark datasets, resulting in much longer model training time and much larger system requirements; and the fact that the observed ratings corresponded to only about 1% of the complete ratings matrix, which means dealing with a very sparse matrix. Since this implementation was motivated by the Netflix data, it is dealing with observed ratings, or explicit feedback. Thus, it solves the matrix factorization problem with ALS using only the observed ratings. Rewriting Eq. 2, Eq. 3 is obtained, where n_{m_i} and n_{m_u} are the number of observed ratings for the item i, and for the user u respectively. Let I_u denote the set of items i that user u has rated, then n_{m_u} is the cardinality of I_u; similarly I_i denotes the set of users who rated item i, and n_{m_i} is the cardinality of I_i.

$$\begin{aligned} min(p,q) & \sum_{(u,i)\in k} (r_{ui} - q_i^T p_u)^2 \\ & + \lambda \left(\sum_i n_{m_i} ||q_i||^2 + \sum_u n_{m_u} ||p_u||^2 \right) \end{aligned} \tag{3}$$

The solution for Eq. 3 follows the steps demonstrated in Section II-B, but, instead of initializing the matrix Q to random values on Step 1 in Alg. 1, it suggests assigning the average rating for that item as the first row, and small random numbers for the remaining entries. The stopping criterion used is based on the observed RMSE on the validation dataset. After one round of updating both Q and P, if the difference between the observed RMSEs is less than 0.0001, the iteration stops and the obtained P, Q are used to make final predictions on the test dataset.

Algorithm 1 ALS algorithm

```
1: procedure ALS(P, Q)        ▷ Matrices representing user feature matrix and item feature matrix, respectively
2:     Initialize matrix Q with random values
3:     repeat
4:         Fix Q, solve P by minimizing the objective function (the sum of squared errors)
5:         Fix P, solve Q by minimizing the objective function similarly
6:     until Stop criteria is satisfied
7:     return P, Q
8: end procedure
```

In this cited approach, a version that allows for parallel computation of Matlab was used. It creates several separate copies of Matlab, each with its own private workspace, and each running on its own hardware platform, collaborate and communicate to solve problems. Each such running copy of Matlab is referred to as a "lab", with its own identifier (labindex) and with a static variable (numlabs) telling how many labs there are. Matrices can be private (each lab has its own copy, and their values differ), replicated (private, but with the same value on all labs) or distributed (there is one matrix, but with rows, or columns, partitioned among the labs).

Because all of the steps use R, two distributed copies of it were used: one distributed by rows (i.e., by users) and the other by columns (i.e., by items). Both P and Q matrices were distributed computed and updated. While computing P, it is required a replicated version of Q, and vice versa. Thus, the labs communicate to make the replicated versions of these matrices from the distributed versions that are first computed. Matlab's "gather" function performs the inter-lab communication needed for this.

To update Q, it is required a replicated copy of P, local to each lab. The ratings data distributed by columns (items) is used. The data is distributed by blocks of equal numbers of items. The lab that stores the ratings of item i will, naturally, be the one that updates the corresponding column of Q, which is items i's feature vector. Each lab computes q_i for all items in the corresponding item group, in parallel. These values are then "gathered" so that every node has all of Q, in a replicated array. To update P, similarly all users are partitioned into equal-size user groups and each lab just updates user vectors in the corresponding user group, using the ratings data partitioned by rows.

The broadcast step is the only communication cost due to using a distributed, as opposed to a shared-memory, algorithm. This method reported taking up less than 5% of the total run time. The algorithm achieves a nearly linear speedup; for $k = 100$, it takes 2.5 hours to update P and Q once with a single processor, as opposed to 5 minutes with 30 processors.

This first work implemented ALS in parallel Matlab and executed on a Linux cluster, with 30 Xeon 2.8GHz processors and every four processors shared 6 GB of RAM. When applied to the Netflix dataset with 100 latent factors and 30 iterations was computed in 2.5 hours and obtained a RMSE of 0.8985 which is a performance improvement of 5.91% over Netflix's Cinematch system.

After the popularization of the Hadoop platform, the parallelization of the ALS algorithm was revisited with a new proposal for a parallel implementation using a series of broadcast-joins that can be efficiently executed with MapReduce [16]. This implementation has partially contributed to Apache Mahout, the open source machine learning library that runs on top of Apache Hadoop framework, and is publicly available. The evaluation setup was a cluster of 26 machines, each with two 8-core Opteron CPU and 32GB of RAM. The experiments showed that on the Netflix dataset, which consists of more than a million ratings given to 17,700 movies by 480,189 users, it was possible to run 37 to 47 iterations of the algorithm, and it typically converges after 15 iterations [18].

This approach is limited to use-cases where neither Q nor P need to be partitioned, meaning they individually fit into the memory of a single machine of the cluster. A rough estimate of the required memory for the re-computation steps in ALS is $max(|M|, |N|) \times k \times 8byte$, as alternatively, a single dense double precision representation of the matrices Q or P has to be stored in memory on each machine. Even for 10 million users or items and a rank $k = 100$, the estimated required memory would be less than 8 GB, which can easily be handled by today's commodity hardware. Experiment results show that, despite this limitation, this implementation is able to handle datasets with billions of data points.

In such a setting, an efficient way to implement the necessary joins for ALS in MapReduce is to use a parallel broadcast-join. The smaller dataset (Q or P) is replicated to every machine of the cluster. Because all of the steps use R, each machine already holds a local partition of R which is stored in the DFS. Then the join between the local partition of R and the replicated copy of P (and analogously between the local partition of R and Q) can be executed by a map operator. This operator can additionally implement the logic to re-compute the feature vectors from the join result, which means that it is possible to execute a whole re-computation of Q or P with a single map operator.

Figure 2 illustrates the parallel join for re-computing P using three machines. First, the broadcast of Q is done to all participating machines, which create a hashtable for its contents, the item feature vectors. R is stored in the DFS partitioned by its rows and forms the input for the map operator, where e.g., $R(1)$ refers to partition 1 of R. The map operator reads a row r_i of R (the interaction history of user i) and selects all the item feature vectors q_j from the

hashtable holding Q that correspond to non-zero entries j in r_i. Next, the map operator solves a linear system created from the interactions and item feature vectors and writes back its result, the re-computed feature vector p_i for user i. The re-computation of Q works analogously, with the only difference that P is broadcasted and R is stored with partitioning done by its columns (the interactions per item) in the DFS.

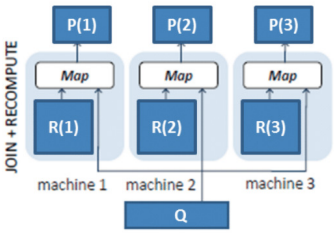


Fig. 2. Parallel recomputation of user features by a broadcast join [16]

This proposed approach is able to avoid some of the drawbacks of MapReduce and the Hadoop implementation described in Section III-A. It uses only map jobs that are easier to schedule than jobs containing map and reduce operators. Additionally, the costly shuffle-phase is avoided, in which all data would be sorted and sent over the network, once the join and the re-computation are done in a single job, which also spare to materialize the join result in the HDFS. This implementation contains multithreaded mappers that leverage all cores of the worker machines for the re-computation of the feature matrices and uses JBlas for solving the dense linear systems present in ALS. The broadcast of the feature matrix is conducted via Hadoop's distributed cache in the initialization phase of each re-computation. Furthermore, Hadoop is configured to reuse the VMs on the worker machines and cache the feature matrices in memory to avoid that later scheduled mappers have to reread the data. The main drawback of a broadcast approach is that every additional machine in the cluster requires another copy of the feature matrix to be sent over the network.

The implementation was also validated on a synthetic dataset called Bigflix, generated from the Netflix dataset and containing 25 million users and more than 5 billion ratings. The performed scale-out test measured the average runtime per job during 5 iterations with 10 latent factors on clusters of 5, 10, 15, 20 and 15 machines. With 5 machines iteration takes about 19 minutes and with 25 machines it was 6 minutes faster.

IV. Experimental Analysis

In this section, we give details of the ALS implementations on Mahout and MLlib libraries that will be executed on Hadoop and Spark respectively. We describe the datasets and the ambient on which the implementations are evaluated, and present the experimental evaluation on the parallel implementations.

A. Datasets

The datasets chosen to run the experiments are from the movies domain (MovieLens) and jokes domain (Jester). Due to copyright problems, Netflix dataset is not available for download anymore. So, to perform the recommendation evaluation on the movies domain, the MovieLens data is frequently used. The MovieLens dataset consists of anonymous ratings of movies and contains approximately 10 million ratings from 71,567 users on 10,681 movies. Ratings are made on a 5-star scale (whole-star ratings only) and each user has at least 20 ratings. The dataset was collected and made available by GroupLens Research, which currently operates a movie recommender based on collaborative filtering, at their webpage. This dataset was previously used to evaluate matrix factorization based methods with neighbor based correction technique, and achieved a best RMSE score of 0.8275 [13].

The Jester dataset consists of anonymous ratings of jokes collected between November 2006 and May 2009. Thus, this data is in a humor domain. It was firstly used to test the Eigentaste recommender and now is freely available for research use. The full data set contains 1,761,439 ratings from 59,132 users on 140 jokes. The ratings are real values ranging from -10.00 to +10.00. Ten percent of the jokes (called the gauge set, which users were asked to rate) are densely rated, others, more sparsely. Two thirds of the users have rated at least 36 jokes. The remaining users have rated between 15 and 35 jokes. The average number of ratings per user is 46, so it is a particularly dense data set compared to Netflix Prize and MovieLens. This dataset was previously used to evaluate matrix factorization based methods with neighbor based correction technique, and achieved a best RMSE score of 4.1229 [13].

B. Mahout ALS Implementation

Mahout 0.9, which was used for this evaluation, presents a MapReduce implementation of ALS that is composed of two jobs: a parallel matrix factorization job, which contains training phase of the ALS algorithm, and a recommendation job that outputs a list of recommended item ids for each user.

Given the ratings matrix (R), the matrix factorization job computes the two intermediate matrices: user-to-feature (P) and item-to-feature (Q). This implementation follows the strategy described in Section III-B, the parallel broadcast-join [16]. Firstly, the smaller dataset (Q or P) is replicated to every machine of the cluster. Also, the ratings matrix is partitioned, and each partition sent to a machine on the cluster, which stores it in the local HDFS. The join between the local partition of R and the replicated copy of P (and analogously between the local partition of R and Q) can be executed by a map operator. This operator can additionally implement the logic to re-compute the feature vectors from the join result, which means that it is possible to execute a whole re-computation of Q or P with a single map operator.

The recommendation job processes the user-to-feature matrix and item-to-feature matrix calculated from the factorization job to compute the top-N recommendations per user. The

```
val Rb = spark.broadcast(R)
for (i <-1 to ITERATIONS){
   P = spark.paralellize(0 until n)
            .map(j => updateUser(j, Rb, Q))
            .collect()
   Q = spark.paralellize(0 until m)
            .map(j => updateUser(j, Rb, P))
            .collect()
}
```

Fig. 3. ALS Spark implementation

predicted rating between user and item is a dot product of the user's feature vector and the item's feature vector.

C. MLlib ALS Implementation

This is a blocked implementation of the ALS factorization algorithm that groups the two sets of factors (referred to as "users" and "items") into blocks and reduces communication by only sending one copy of each user vector to each item block on each iteration, and only for the item blocks that need that user's feature vector. This is achieved by precomputing some information about the ratings matrix to determine the "out-links" of each user (which blocks of items it will contribute to) and "in-link" information for each item (which of the feature vectors it receives from each user block it will depend on). This allows the implementation to send only an array of feature vectors between each user block and item block, and have the item block find the users' ratings and update the items based on these messages.

Because all of the steps use the ratings matrix R, it is helpful to make it a broadcast variable so that it does not get re-sent to each node on each step. Figure 3 shows the ALS Spark implementation. Note in Lines 3 to 5 that collection 0 until u are parallelized and collected to update each array [26]. The ALS recommender accepts as input an RDD (Resilient Distributed Datasets) of ratings (user: Int, product: Int, rating: Double).

D. Experimental Results

The experiments were developed with Python 2.6 and firstly executed in local single machine mode for testing. Then, the final experiments were executed at Amazon Web Services (AWS).

The clusters used for the evaluation consists of t2.small EC2 instances running Ubuntu 64-bit OS with Oracle Java (JDK) 7, Apache Hadoop 1.2.1 and Apache Spark 1.1.1. Each t2.small instance has a 3.3GHz core processor, 2GB of RAM and 15GB of SSD storage. The accuracy and efficiency experiments were conducted on a cluster of 4 machines.

To evaluate these algorithms, the datasets were randomly divided into three non-overlapping subsets, named: training (60%), test (20%), and validation (20%). These datasets are saved on two datanodes of the HDFS, since this is the smaller cluster configuration for scalability experiment.

These two datanodes are accessible for all the workers through the experiments, since Spark is running in the same Hadoop cluster through Spark's standalone mode, that is, by simply placing a compiled version of Spark on each node on the cluster.

1) Accuracy and Efficiency Experiment: To evaluate the quality of the recommendations produced by each of the two implementations, multiple models are trained based on the training set, and that which achieves the smallest root-mean-square error (RMSE), given by Eq. 4, on the validation set after running 20 iterations of the algorithm is chosen as the best fit ALS model [18]. Finally, this model is evaluated on the test set.

$$RMSE = \sqrt{\frac{1}{|S_{val}|} \sum_{(m,n) \in S_{val}} (r_{ui} - \widehat{r}_{ui})^2} \qquad (4)$$

The parameters tested to find the best fit ALS model are combinations resulting from the cross product of the dimensionality of the latent factor space, $k = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]$ and the regularization parameter $\lambda = [0.10, 0.25, 0.50, 0.75, 1.00]$.

Analyzing the convergence of the Spark-MLlib ALS on the Jester validation set for the number of latent factors (k), we can see that the recommendation quality usually improves when increasing λ until the optimal value of 0.5. Then, the increment worsens the recommendation accuracy. For the HadoopMR-Mahout ALS on the Jester validation set, we observed the same behavior presented by the Spark-MLlib implementation: the recommendation quality usually improves when increasing λ until the optimal value of 0.5 but beyond that, the recommendation is worse.

The best fit Spark-MLlib ALS for the Jester dataset has RMSE on the test set of 4.1339 and 4.1395 for the HadoopMR-Mahout. Both models have the same value for the parameters k and λ, but the MLlib implementation achieves a result 0.13% better, with a training execution time that is more than 10 times faster, in a cluster with 4 t2.small instances, as shown on Table I.

TABLE I
BEST FIT ALS MODEL RESULTS FOR JESTER DATASET

	Spark-MLlib	HadoopMR-Mahout
$k_{BestfitALS}$	20	20
$\lambda_{BestfitALS}$	0.5	0.5
RMSE(Validation Set)	4.1378	4.1385
RMSE(Test Set)	4.1339	4.1395
Execution time (sec)	61.4	671.4

For the MovieLens dataset, the best fit Spark-MLlib ALS is trained with $k = 20$, $\lambda = 0.5$ and RMSE = 0.8099 on the validation set. We observed that the model converges on each value of $\lambda \geq 0.5$ regardless of the feature space size. The RMSE on the test set is 0.8091, which means that the model does not overfit the observed ratings. For the HadoopMR-Mahout ALS modelling results, the best fit is trained with

$k = 20$, $\lambda = 0.5$, and RMSE = 0.8196 on the validation set, the same parameters found for the Spark-MLlib implementation. The convergence behavior found before repeats itself here, for each $\lambda \geq 0.5$ regardless of the feature space size. Also, the RMSE on both implementations for the same regularization parameter is very close: the largest difference is of only 0.0002 or 0.001% of the rating score, represented on a scale of -10.0 to +10.0.

Comparing the best fit ALS models achieved by both implementations, again the Spark-MLlib solution has a better performance: more accurate, with a RMSE on the test set 1.4% smaller than the HadoopMR-Mahout implementation, and more efficient, with execution more than 5 times faster to run (Fig. 4). The results for the MovieLens dataset are summarized on Table II.

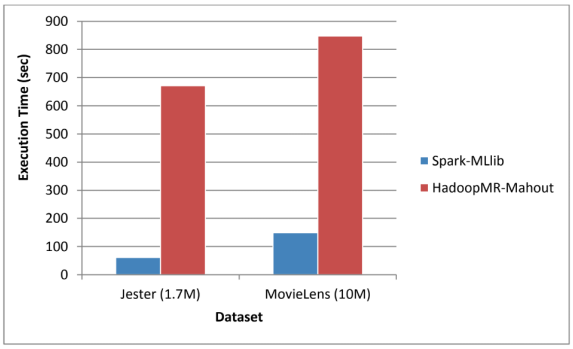


Fig. 4. Execution time for the best fit ALS model on a cluster with 4 machines

TABLE II
BEST FIT ALS MODEL RESULTS FOR MOVIELENS DATASET

	Spark-MLlib	HadoopMR-Mahout
$k_{BestfitALS}$	20	20
$\lambda_{BestfitALS}$	0.1	0.1
RMSE(Validation Set)	0.8099	0.8196
RMSE(Test Set)	0.8091	0.8202
Execution time (sec)	149.1	847.9

2) Scalability Experiment: To test the scalability of these recommender systems, we measure the walltime of 20 iterations of the best fit ALS model on each of the datasets on different cluster sizes, consisting of 2, 4 and 6 AWS EC2 t2.small instances. We observe that the computation speedup does not linearly scale with the number of machines, which is an expected behavior since both implementations have a broadcast of the ratings matrix so every additional machine causes another copy of it to be sent over the network. Comparing the speedup values for the two implementations, shown on Fig. 5, we find that, when training the HadoopMR-Mahout ALS model with 6 machines, it shows an improvement of 1.60x on the Jester dataset and 1.45x on the MovieLens dataset over the execution with 2 machines, and for the Spark-MLlib implementation, executing with 6 machines provides an improvement of 1.86x on the Jester dataset and 2.39x on the MovieLens dataset over the execution with 2 machines.

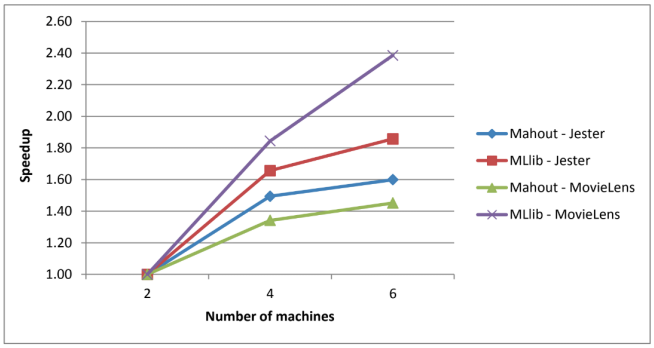


Fig. 5. Speedup for both implementations

As seen in Table III, the distributed and parallel ALS implementation on MLlib executed on the Spark cluster with 6 machines achieved the faster training time for both datasets: 54.7 seconds for Jester that contains about 1.7 million joke ratings, and 115.3 seconds for MovieLens that contains about 10 million movie ratings. By extrapolating these results, we find that a recommender system with a dataset with 100 million ratings input, which is 10 times bigger than the MovieLens dataset, would take about 415 seconds to be trained on a cluster with 6 machines with the t2.small EC2 configuration. If we wished to put such a system into production, we could either utilize more of these general purpose instances or choose machines with more RAM, such as the M3 instances or the R3 memory optimized instances, which suggests that the Spark implementation is suitable for real world use cases.

TABLE III
SUMMARY OF RECOMMENDATION TIME

Dataset	size (# of ratings)	Recommendation time (in sec)
Jester:	1.7 mi	54.7
MovieLens:	10 mi	115.3
100 mi	415	

V. CONCLUSIONS

Alternating Least Squares (ALS) algorithm is an efficient approach in situations where generating online recommendations and processing large datasets is required. In this work, we described two scalable parallel implementations of the ALS algorithm, the Mahout ALS and MLlib ALS. Each one uses a different framework for distributed processing on clusters of commodity hardware, respectively, Hadoop MapReduce and Spark.

We performed an experimental analysis comparing the different implementations of the ALS algorithm for collaborative filtering recommender systems, using datasets from two different domains: MovieLens, from the movies domain, and Jester, from the jokes domain. First we found the best fit ALS model for each of the datasets. Using the optimized parameters to train the ALS models, we performed the evaluation of the implementations in terms of execution time and accuracy results on the test set.

The experimental results showed that Spark-MLlib solution has a better performance than the Mahout ALS in terms of accuracy and efficiency for both recommendation domains. For the Jester dataset, the RMSE on the test set with Spark-MLlib was 0.13% better than with HadoopMR-Mahout, and the training was more than 10 times faster in a cluster with 4 machines. For the MovieLens dataset, the RMSE on the test set was 1.4% smaller on the Spark-MLlib implementation, and the modeling was 5 times faster. This study also featured a scalability experiment, running the best fit ALS model on clusters of 2, 4 and 6 machines. Again, the results were favorable to Spark, since it has a more expressive computational speedup: training time on a cluster with 6 machines was 86% faster on the Jester dataset and 139% on the MovieLens dataset when comparing to execution time on a cluster with 2 machines.

Deploying a recommender system on six t2.small instances available from EC2 took 115.3s for a dataset containing about 10 million ratings, and, by extrapolation, it would take about 415s for a dataset with 100 million ratings. The results suggest that a cluster with at least six t2.small instances or fewer and more potent machines, like M3 or R3 memory optimized instances available on EC2, would run a user's full recommendations measures in a few seconds, which is a suitable time frame for production settings. Future works are desirable in order to keep comparing the recommendation algorithms implementations available in the newer releases of MLlib and Mahout, as well as newer technologies, since both engines for large-scale data processing are rapidly evolving.

REFERENCES

[1] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso, "Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems," *ACM Trans. Web*, vol. 5, no. 1, pp. 2:1–2:33, Feb. 2011. doi: 10.1145/1921591.1921593. [Online]. Available: http://doi.acm.org/10.1145/1921591.1921593

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, ser. WWW '01. New York, NY, USA: ACM, 2001. doi: 10.1145/371920.372071. ISBN 1-58113-348-0 pp. 285–295. [Online]. Available: http://doi.acm.org/10.1145/371920.372071

[3] Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 116–142, Jan. 2004. doi: 10.1145/963770.963775. [Online]. Available: http://doi.acm.org/10.1145/963770.963775

[4] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. ISBN 1-55860-555-X pp. 43–52. [Online]. Available: http://dl.acm.org/citation.cfm?id=2074094.2074100

[5] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007. doi: 10.1145/1273496.1273596. ISBN 978-1-59593-793-3 pp. 791–798. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273596

[6] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl, "Application of dimensionality reduction in recommender system – a case study," in *WebKDD Workshop, held in conjunction with the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2000)*, 2000.

[7] Spotify, "Spotify," n.d., available at https://www.spotify.com/.

[8] C. Johnson, "Scala data pipelines for music recommendations," 2015, available at http://www.slideshare.net/MrChrisJohnson/scala-data-pipelines-for-music-recommendations.

[9] Amazon.com, "Amazon.com," n.d., available at http://www.amazon.com/.

[10] ExportX, "How many (more) products does amazon sell?" 2014, available at http://export-x.com/2014/08/14/many-products-amazon-sell-2.

[11] Statista, "Number of worldwide active amazon customer accounts from 1997 to 2014 (in millions)," 2014, available at http://www.statista.com/statistics/237810/number-of-active-amazon-customer-accounts-worldwide/.

[12] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. Kantor, Eds. Springer, 2011, pp. 33–48.

[13] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Using visual representations of data to enhance sensemaking in data exploration tasks." *Journal of Machine Learning Research*, vol. 10, pp. 623–656, 2009.

[14] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009. doi: 10.1109/MC.2009.263. [Online]. Available: http://dx.doi.org/10.1109/MC.2009.263

[15] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," in *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '07. New York, NY, USA: ACM, 2007. doi: 10.1145/1281192.1281206. ISBN 978-1-59593-609-7 pp. 95–104. [Online]. Available: http://doi.acm.org/10.1145/1281192.1281206

[16] S. Schelter, C. Boden, M. Schenck, A. Alexandrov, and V. Markl, "Distributed matrix factorization with mapreduce using a series of broadcast-joins," in *Proceedings of the 7th ACM Conference on Recommender Systems*, ser. RecSys '13. New York, NY, USA: ACM, 2013. doi: 10.1145/2507157.2507195. ISBN 978-1-4503-2409-0 pp. 281–284. [Online]. Available: http://doi.acm.org/10.1145/2507157.2507195

[17] The Apache Software Foundation, "Apache hadoop," n.d., available at http://hadoop.apache.org/.

[18] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *Algorithmic Aspects in Information and Management. AAIM 2008. LNCS, vol 5034*, R. Fleischer and J. Xu, Eds. Springer, 2008.

[19] The Apache Software Foundation, "MLlib," n.d., available at http://spark.apache.org/mllib/.

[20] J. Chen, J. Fang, W. Liu, T. Tang, X. Chen, and C. Yang, "Efficient and portable als matrix factorization for recommender systems," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, May 2017. doi: 10.1109/IPDPSW.2017.91 pp. 409–418.

[21] C. Enrique, T. Alexander, C. Héctor, J. G. Francisco, G. Felipe, B. Belén, and A. Diego, "In-memory distributed software solution to improve the performance of recommender systems," *Software: Practice and Experience*, vol. 47, no. 6, pp. 867–889, 2017. doi: 10.1002/spe.2467. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2467

[22] D. Kim and B.-J. Yum, "Collaborative filtering based on iterative principal component analysis," *Expert Syst. Appl.*, vol. 28, no. 4, pp. 823–830, May 2005. doi: 10.1016/j.eswa.2004.12.037. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2004.12.037

[23] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, ser. NIPS'07. USA: Curran Associates Inc., 2007. ISBN 978-1-60560-352-0 pp. 1257–1264. [Online]. Available: http://dl.acm.org/citation.cfm?id=2981562.2981720

[24] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics*, vol. 4, p. 1035–1038, 1963.

[25] S. Funk, "Netflix update: Try this at home," 2006, available at http://sifter.org/~simon/journal/20061211.html.

[26] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 10–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1863103.1863113

5th International Conference on Cryptography and Security Systems

CRYPTOGRAPHY and security systems are two fields of security research that strongly interact and complement each other. The International Conference on Cryptography and Security Systems (CSS) is a forum of presentation of theoretical, applied research papers, case studies, implementation experiences as well as work-in-progress results in these two disciplines.

TOPICS

The main topics of interests include:

- network security
- cryptography and data protection
- peer-to-peer security
- security of wireless sensor networks
- security of cyber physical systems
- security of Internet of Things solutions
- heterogeneous networks security
- privacy-enhancing methods
- covert channels
- steganography and watermarking for security applications
- cryptographic protocols
- security as quality of service, quality of protection
- data and application security, software security
- security models, evaluation, and verification
- formal methods in security
- trust and reputation models
- reputation systems for security applications
- intrusion tolerance
- system surveillance and enhanced security
- cybercrime: threats and countermeasures
- 5G Security
- DDoS attacks: detection and mitigation
- Security of Smart Grid systems

EVENT CHAIRS

- **Kotulski, Zbigniew,** Warsaw University of Technology, Faculty of Electronics and Information Technology, Institute of Telecommunications, Department of Cybersecurity, Poland
- **Ksiezopolski, Bogdan,** Maria Curie-Sklodowska Univerity, Faculty of Mathematics, Physics and Computer Science, Institute of Computer Science, Department of Cybersecurity and Polish-Japanese Academy of Information Technology

PROGRAM COMMITTEE

- **Cabaj, Krzysztof,** Institute of Computer Science, Warsaw University of Technology, Poland
- **Caviglione, Luca,** National Research Council (CNR), Italy
- **Chan-Tin, Eric,** Oklahoma State University, United States
- **Dittmann, Jana,** Otto-von-Guericke Universität Magdeburg, Germany
- **Domingos, Maria Dulce Pedroso,** Universidade de Lisboa, Portugal
- **Gajewski, Piotr,** Military University of Technology, Poland
- **Górski, Janusz,** Gdańsk University of Technology, Poland
- **Gutierrez, Jaime,** Universidad de Cantabria, Spain
- **Johnson, Thomas,** Oklahoma State University, United States
- **Kotenko, Igor,** St.Petersburg Institute for Informatics and Automation, Russia
- **Kula, Mieczysław,** University of Silesia, Poland
- **Lafourcade, Pascal,** Clermont University, France
- **Leprévost, Franck,** University of Luxembourg, Luxembourg
- **Mauw, Sjouke,** University of Luxembourg, Luxembourg
- **Mazurczyk, Wojciech,** Warsaw University of Technology, Poland
- **Pejaś, Jerzy,** West Pomeranian University of Technology, Poland
- **Pieprzyk, Josef,** Queensland University of Technology, Australia
- **Piotrowski, Zbigniew,** Military University of Technology, Poland
- **Respício, Anna,** Universidade de Lisboa, Portugal
- **Ryan, Peter Y A,** University of Luxembourg, Luxembourg
- **Seredyński, Franciszek,** Cardinal Wyszyński University in Warsaw, Poland
- **Stokłosa, Janusz,** WSB University in Poznan, Poland
- **Sulema, Yevgeniya,** National Technical University of Ukraine, Ukraine
- **Szałachowski, Paweł,** SUTD, Singapore
- **Tiplea, Ferucio Laurentiu,** Alexandru Ioan Cuza University of Iasi, Romania
- **Ustimenko, Vasyl,** Marie Curie-Sklodowska University, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 363–368
DOI: 10.15439/2018F197
ISSN 2300-5963 ACSIS, Vol. 15

An Improved Architecture of a Hardware Accelerator for Factoring Integers with Elliptic Curve Method

Michał Andrzejczak
Wojskowa Akademia Techniczna
ul. Urbanowicza 2, 01-489 Warszawa, Poland
Email: michal.andrzejczak@wat.edu.pl

***Abstract*—Elliptic Curve Method (ECM) is a well-known method for factoring integers, which is usually used in the Number Field Sieve algorithm as a subroutine for factoring smaller integers than the targeted one. ECM is called many times and can be executed in parallel for different inputs. This method mainly consist of simple operations on elliptic curves. Thus, ECM is suitable for hardware implementations that can efficiently reduce computational time. This work describes a new, improved FPGA-based hardware accelerator for ECM, designed for large scale computations. Our accelerator can operate with an on board ARM processor or with an external host computer. This design can factor several numbers at once and can be easily ported to various FPGA boards. Different methods for improving results (e.g. the use of DSP blocks, cache-registers, reorganizing instruction order) are described and their performance is analyzed. As a result, one of the fastest hardware ECM units is achieved.**

I. Introduction

FACTORIZATION is one of the main hard problems used in construction of cryptosystems. One of the most known and the most popular cryptosystem based on factorization problem is RSA. To the present day, several algorithms for factoring integers have been developed and used in various cases. The best algorithm for factoring integers with large factors used in RSA is Generalized Number Field Sieve (GNFS) [2]. This method require to factor a lot of smaller numbers in one of the main steps of the algorithm and the Elliptic Curve Method can be efficiently used for this. ECM performs many operations on a small data, requiring little memory and can be run many times in parallel with the same probability of factoring chosen number. Thus, special purpose hardware can efficiently improve overall factorization time. In this paper an improved architecture of a hardware accelerator for factoring integers with Elliptic Curve Method (ECM) is presented. Analysis of previous architecture is included and detected weaknesses are described with potential improvements. At the end, influences of several changes in initial design are compared, with the best result more than three times better than previously reported in literature.

II. Elliptic Curve Method

The ECM was proposed by H.W. Lenstra [1] (called Phase 1) in late 80's and its principles are based on Pollard (p-1) method. Later, ECM was extended and improved by Brent [3] and Montgomery [4] (called Phase 2).

Let choose a field K with characteristic different from 2 and 3. The elliptic curve $E_{A,B}$ is the set of points $(X, Y) \in K$ such that

$$Y^2 = X^3 + AX + B$$

where $A, B \in K$ and $4A^3 + 27B^2 \neq 0$ with a special point $O_E = (0 : 1 : 0)$ called a "point at infinity".

For more efficient computer implementation, Montgomery's form of elliptic curve is recommended due to lack of number inversion computation. Montgomery's form can be obtained from Weierstrass form presented above by following change of the variables $X \to (3x + a)/(3b)$, $Y \to y/b$, $A \to (3 - a^2)/(3b^2)$, $B \to (2a^3 - 9a)/(27b^3)$. Homogeneous form of this curve is:

$$by^2z = x^3 + ax^2z + xz^2$$

with the triple $(x : y : z)$ represents the point $(x/z : y/z)$ in affine coordinates. Projective coordinates of curve in Montgomery's form allow all intermediate computations to be performed using only x and z coordinate. The y coordinate can be retrived from two others coordinates, but is not necessary in ECM algorithm.

Let q be an unknown factor of N - the number being factorized. The ECM starts with randomly selecting an elliptic curve $E_{a,b}$ and a random point on it. Computations are performed modulo the number N, as if $\mathbb{Z}/n\mathbb{Z}$ was a field. First step of computations can be done just once. In this step, product of all prime numbers and its powers is computed. Most time consuming operation is done in second step, where scalar multiplication of chosen point by computed product is performed. In the last step, greatest common divisor of resulted z coordinate and a factorized N is computed. Pseudocode for ECM is shown in Listing 1.

Algorithm 1 ECM algorithm, phase 1

Require: a composite number N, random point P_0 on random elliptic curve E, integer bound B_1
Ensure: a factor of N or $fail$
$k \leftarrow \prod_{p \leq B_1} p^{\log_p B_1}$
$Q_0 \leftarrow kP_0$;
$q \leftarrow gcd(z_Q, N)$;
if $q > 1$ **then**
  return q
else
  return $fail$
end if

A. Complexity of the ECM

The complexity of ECM is sub-exponential and is described as:

$$\mathcal{O}(n) = e^{\sqrt{\log p \log \log p}\sqrt{2}+o(1)} M(\log n)$$

where $M(\log n)$ is the complexity of multiplication $\mod n$. First part depends only on factors of the chosen integer. The only way to speed up computations is to execute point multiplication as fast as possible, so basically what this paper is about.

III. Initial Design

The initial design was proposed in [8]. The main idea of that hardware accelerator is to spread as many as possible autonomous ECM units in one FPGA chip. The ECM units have Harvard architecture with separable instruction list and data memory. Design can be described in three levels. First of them, the top level, describe FPGA device and interconnections between main modules and external components. Lower level is about design of ECM unit, interconnections between memory, controllers and arithmetic modules. The last level describes architecture and algorithms used in modules building the ECM unit (modular multipliers, adders, controller).

A. Top level

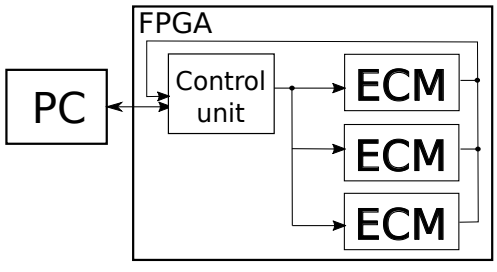


Fig. 1. Top level

In Fig. 1 the top level is shown. FPGA chip is filled by as many ECM units as possible with one global control unit for all of them, responsible for Montgommery Ladder execution, communication with external PC and managing the work units. Connected PC is used for random curves sampling and for last stage of the algorithm, the gcd computation. It is done in this way to maximize logic usage and simplify chip design. Many independent ECM units allow better clock signal distribution. FPGA is responsible only for point multiplication over elliptic curve.

B. ECM level

Every ECM unit is equipped with internal memory, memory controller, microprocessor and 4 arithmetic units (two for modular multiplication, one for addition and subtraction) as shown in 2. During initialization, every ECM unit need random elliptic curve and random point over this curve. Provided curves should be in Montgomery form and every coordinate should be converted to Montgomery domain [5].

The memory controller is responsible for communicating with two way memory bank. Loaded data words are concatenated and put into bus registers. This controller has also internal semaphore table for preventing data override during parallel execution and additional table for storing result address of computed data.

The main controller has ROM memory for instruction and can execute simple commands. Every instruction takes two memory addresses for data input and one address for writing result. There are 5 instructions:

- **ADD** - instruction used for addition
- **SUB** - used for subtraction
- **MULA**- multiplication by first unit
- **MULB**- multiplication by second unit
- **LOADN**- modulus read from memory

These instructions can be used to replace computation path with computations over Edwards curves by simply reprogramming the ROM table.

The ECM unit is equipped with two modular multiplication unit which allow faster point multiplication. This idea was taken from [6].The computation flow for point doubling and addition in Montgomery form ([5]) is shown in Table I) and uses two multipliers in parallel.

C. Module level

Modular multiplication is the most time consuming operation. During every point doubling/addition it is performed 11 times and this number can be reduced to 10. To obtain the lower number of multiplications one coordinate of P_0 must be chosen arbitrarily to simplify computations by selecting $z_{P_0} = z_{P-Q} = 1$. Thus, use of two modular multipliers can increase total throughput. For multiplication, logic based algorithm [7] was used. The aim of that was to have design capable to be deployed on low cost devices without enough DSP modules. This also save logic required for routing to these modules in designs with high percentage of logic usage. Implemented algorithm perform modular multiplication of n - bit numbers in Montgomery form in n clock cycles. Modular reduction in Montgomery's domain is based on efficient hardware bit shift operation and was chosen due to very good performance.

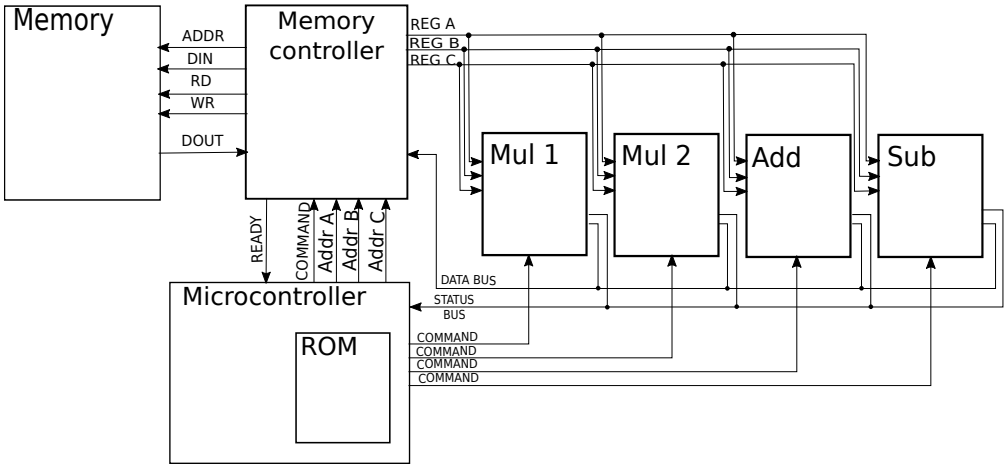


Fig. 2. ECM level

TABLE I
ONE STEP OF SCALAR MULTIPLICATION IN CASE OF $z_{P-Q} = 1$

Adder	Substractor	Multiplier 1	Multiplier 2
$a_1 = x_P + z_P$	$s_1 = x_P - z_P$		
$a_2 = x_Q + z_Q$	$s_2 = x_Q - z_Q$	$m_1 = s_1^2$	$m_2 = a_1^2$
	$s_3 = m_2 - m_1$	$m_3 = s_1 \cdot a_2$	$m_4 = s_2 \cdot a_1$
$a_3 = m_3 + m_4$	$s_4 = m_3 - m_4$	$x_{2P} = m_1 \cdot m_2$	$m_6 = s_3 \cdot a_{24}$
$a_4 = m_1 + m_6$		$x_{P+Q} = a_3^2$	$m_8 = s_4^2$
		$z_{P+Q} = m_8 \cdot x_{P-Q}$	$z_{2P} = s_3 \cdot a_4$

D. Comparision

Basic parameters of this design are shown in Table II in comparison with other results reported earlier in literature. The design was compiled for low cost Altera DE1-SOC board equipped with Cyclone VCSEMA5F31C6 FPGA and for high end Stratix IV targeted for high performance computing.

IV. ANALYSIS AND IMPROVEMENTS

The reported results for initial design are very competitive. However, deep analysis of proposed solution indicates several bottlenecks which may be improved to achieve much better performance.

Fig. 3 shows data dependency graph for point multiplication. Every arrow represents memory read operation and circles are the arithmetic operations with result write to memory. Memory controller is capable only to read from memory to one register at once which result in doubling the same operation. Moreover, data is loaded almost immediately after being write to memory in several cases.

A. Memory improvements

Analysis of the simulation diagrams proved that memory operations give one of the biggest slowdown on design. Every arithmetic operation needs two load operations of operands (which is done by sequential memory loads, concatenated at the end) and one write operation of result, done in similar manner. For 192-bit length numbers and 32-bit size memory, communication overhead takes around 20 clock cycles for one operation.

Simple solution for this problem is to increase memory base size to decrease this overhead. Increasing memory base size from 32-bits to 128-bits can decrease the number of memory calls from 21 to 6 clock cycles. Size of the design increase slightly with this improvements, further called **opt1**.

On the other hand, situation when one variable is loaded twice in a row or written and read in next step is very common. The first issue can be solved by expanding instruction set by load instruction for two operands at once. Several cases when the same data is loaded for two arithmetic units can be improved by splitting arithmetic instructions for more atomic operations. Instead one arithmetic instruction, where attributes are addresses in memory, we use load to register instructions and execute arithmetic operation instruction (without any arguments). .

The second issue needs additional cache registers for temporary results. Adding these registers slightly increases design size, but offered overall performance by FPGA chip is improved. With this change it is possible to replace order list. New orders can be more atomic. With atomic instructions the program size increase, but there is no need in memory controller to be responsible for parallel data access. 4 temporary registers were added to design. With these registers and with direct result to input operation, memory usage is limited only

TABLE II
RESULTS OF THE IMPLEMENTATION COMPARED WITH OTHERS REPORTED IN LITERATURE

Author:	**Gaj [6]**	**Gaj [6]**	**Zimmermann [11]**	**Zimmermann [11]**	**de Meulenaer [10]**	**Andrzejczak [8]**
Device:	S35000	V4LX200	V4SX35	XC4VSX35	XCV4SX25	SGX530
Number length:	198 - bit	198 - bit %	202 - bit	134 - bit	135 - bit	192 - bit
Max. number of modules	13	24	24	24	1	96
Max. clock freq.	80 MHz	104 MHz	200 MHz	200 MHz	220 MHz	150 MHz
Clock cycles in phase 1	1 666 500	1 666 500	1 473 596	797 288	13 750	2 101 400
Time for phase 1	21 ms	16 ms	7.37 ms	3.99 ms	63 s	14 ms
Curves/sec:	624	1 448	3 240	6000	16 000	6822

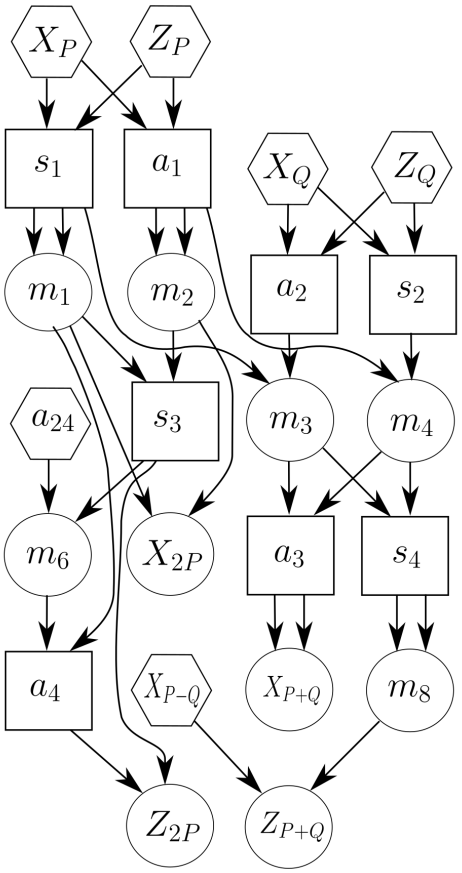


Fig. 3. Data computation graph for point multiplication

to load first coordinates for first bit of ladder and load one coordinate (a_{24}) for every bit. The intermediate results can be stored in these 4 additional temporary registers or can be directly passed to input of another arithmetic unit. The saved memory can be used in Phase 2 to store more pre computation results and improvements from **opt1** are less significant in overall result and are not included in this design, called **opt2**

Instructions have no longer the same format. Arithmetic instructions are not taking any arguments, they operate on data provided to special input registers loaded earlier. Extended instruction set with description is shown in Table III

TABLE III
EXTENDED INSTRUCTION SET WITH FORMAT DESCRIPTION

Name	**Description**
RLOAD	Load data from one register to another one. Can be used to load data for two registers at once
MLOAD	Memory load to one or two registers
MULA	Start multiplication in unit A
MULB	Start multiplication in unit B
ADD	Start addition
SUB	Start subtraction
WAITFOR	Waits for end of computation in selected module

After memory operation optimization and with the new instruction set, data computation graph is changed. Fig. 4 presents improved data computation graph. All coordinates loaded from memory are marked by gray color and correspond to first improvement. Second improvement is presented with red circles marking data loaded for different modules in one load operation. Double loads for integer squaring or doubling are marked by one pointer. Values stored in temporary registers are marked with dots.

B. Multiplication unit replacement

The other way to increase number of checked elliptic curves is to speed-up multiplication computation. The modular multiplication based on logic gates takes n clock cycles and to decrease this number DSP multiplication algorithms should be used. Algorithm for modular Montgomery multiplication proposed by Itoh [9] was selected. Multiplier is parametrized by radix and multiplication is performed in n^2 steps, where:

$$n = \frac{\text{number length}}{radix}$$

Optimal selection of radix is crucial for overall performance. Bigger radix needs more DSP (Digital Signal Processing) blocks used for integer multiplication. The size of multiplier (in Logic Elements) increase as increase the number of DSP blocks needed, because of longer paths used to route signals to these blocks. The best results have been achieved for radix 32, requiring only 3 DSP blocks per multiplication and executing

TABLE IV
IMPROVEMENTS COMPARISION

Parameter	Initial	opt1	opt2	opt3	Stratix IV opt3
Logic:	30 060	31 394	30 982	25 566	372 951
Logic usage(%):	97 %	98 %	97 %	80 %	92 %
DSP:	-	-	-	66	654
Max. clock freq.	88 MHz	88 MHz	85 MHz	90.1 MHz	151 MHz
Num. of Units	10	10	9	11	109
Clock cycles per bit	1580	1374	1215	481	481
Clock cycles in phase 1:	2 101 400	1 827 420	1 615 950	639 730	639 730
Curves/sec:	418	480	473	1546	25 557
Improvement factor:	**1**	**1.14**	**1.13**	**3.69**	**3.72**

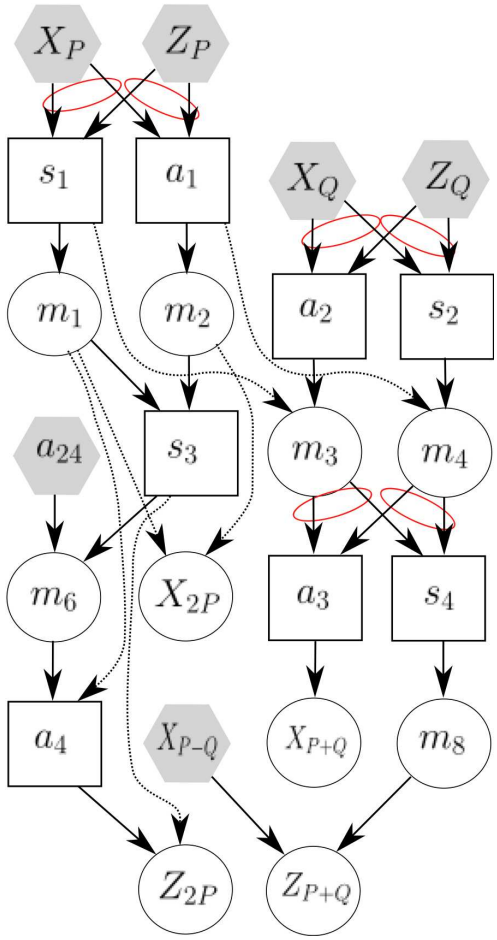


Fig. 4. Data computation graph for improved point multiplication

in 54 clock cycles. Bigger radix significantly decrease the maximum clock frequency, due to longer propagation paths and require more DSP block, which lead to higher logic utilization. Variant with shorter radix executes in more clock cycles and leave many DSP blocks unused in global view.

The replacement of multiplication algorithm saved around 150 clock cycles per each call of this function. Two modules are used at once, so in general around 750 clock cycles are saved compared to previous used logic based algorithm. Moreover, DSP-based module requires less logic element and Quartus Prime compiler was able to fit one more ECM unit in targeted Cyclone V device. For 11 ECM units the compiler needs 25 566 ALM (Adaptive Logic Modules) which is 80% of all available resources. Adding one more ECM unit is not possible. For 12 ECM units compiler can not place all of them close to hardware multipliers, so longer routing path are needed. For targeted Cyclone V, 33 091 ALM is required and it is above 100% of available resources. Multiplier replacement is called **opt3**

Improvements were implemented incrementally. Every next design contains improvement from previous versions. Table IV compares results of constructed modules. The last column contains compilation data for Stratix IV GX530, used to check throughput of initial design. This one is much bigger than the low cost Cyclone V and can be used in practice to factorize numbers in GNFS. Achieved results are 3.72 times better than at the beginning. The total number of sieved curves is the highest one from reported in literature.

V. CONCLUSIONS

An improved hardware architecture for factoring integers has been presented. Careful analysis of algorithm and hardware design lead to architectural changes resulting 3.72 times faster device. The most efficient was the change of multiplication algorithm, which reduced almost half of the total number of computations. Additional temporary registers may be more significant for total computation time if the base of memory was not increased and memory operations will still take almost four times more. With all improvements combined, the fastest architecture is obtained.

Recently Intel introduced new more powerful devices called Stratix 10. Further works will adapt described design for new devices. Preliminary simulations shows around 180 thousands curves per second for one of the biggest devices from the new family.

REFERENCES

[1] H. W. Lenstra, "Factoring Integers with Elliptic Curves" *Annals of Mathematics,* vol. 126, no.2, pp. 694–673, 1985.

[2] A. K. Lenstra and H. W. Lenstra, "The Development of the Number Field Sieve" *Lecture Notes in Math,* Volume 1554, 1993.
[3] R. P. Brent, "Some integer factorization algorithms using elliptic curves," *Australian Computer Science Communications,* vol. 8, pp. 148-163, 1986.
[4] P. L. Montgomery, "Speeding the Pollard and elliptic curve methods of factorization," *Mathematics of Computation,* vol. 48, pp. 243-264, 1987.
[5] P. L. Montgomery, "Modular Multiplication Without Trial Division," *Mathematics of Computation,*, vol. 44, pp. 519-519, 1985.
[6] K. Gaj et al., "Area-time efficient implementation of the elliptic curve method of factoring in reconfigurable hardware for application in the number field sieve," *IEEE Transactions on Computers,* vol. 59, pp. 1264-1280, 2010/9
[7] K. Gaj, M. Huang, S. Kwon, T. A. El-Ghazawi, "An Optimized Hardware Architecture for the Montgomery Multiplication Algorithm," *Public Key Cryptography,* , 2008
[8] M. Andrzejczak, "Koprocesor kryptograficzny wspierajacy faktoryzacje liczb metoda krzywych eliptycznych," [Konferencja mlodych naukowcow wiwat 2017, Falenty, Polska, 2017], in press.
[9] K. Itoh, M. Takenaka, N. Torii, S. Temma, Y. Kurihara "Fast Implementation of Public-Key Cryptography on a DSP", [Cryptographic Hardware and Embedded System], 2002.
[10] G. de Meulenaer, F. Gosset, G. de Dormale, J. Quisquater, "Integer Factorization Based on Elliptic Curve Method: Towards Better Exploitation of Reconfigurable Hardware", [15th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM 2007)], Napa, CA, 2007, pp. 197-206.
[11] R. Zimmermann, "Optimized Implementation of the Elliptic Curve Factorization Method on a Highly Parallelized Hardware Cluster", Master Thesis, TU Braunschweig, 2009 r.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 369–379
DOI: 10.15439/2018F322
ISSN 2300-5963 ACSIS, Vol. 15

Graph-based quantitative description of networks' slices isolation

Zbigniew Kotulski, Tomasz Wojciech Nowak, Mariusz Sepczuk, and Marcin Alan Tunia
Faculty of Electronics and Information Technology of the Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665, Warsaw
Email: z.kotulski@tele.pw.edu.pl, T.Nowak@tele.pw.edu.pl, msepczuk@tele.pw.edu.pl, M.Tunia@tele.pw.edu.pl

***Abstract*—5G networks are expected to be a set of slices which are virtual subnets designed for specific applications. A crucial requirement for providing proper functioning of the network and its security is proper isolation of slices. The aim of this paper is to propose a new method of determination of the isolation level of a slice. We propose a Graph-based model of the sliced network, which has a layered structure. In each layer, the appropriate network elements have their own isolation level. The lowest layer of the Graph-based model represents virtual network elements with isolation traits used for calculating their isolation level. Climbing to the top of the stack of layers one can calculate, successively, isolation level for a network's physical element, a link, subnetworks and, the End-to-End slice's isolation level. We present numerical examples, where suitable traits are specified and the isolation level in each layer is calculated.**

I. Introduction

NETWORK slicing is a key technology for 5G network [1]–[3]. In 5G, transmission quality, network performance and services' reliability are expected to be on extremely high-level (e.g., the bandwidth over 300 Mbps, very small latency of few milliseconds and support up to 200,000 devices/km^2 with 99.999% reliability level, see [4]–[6]). Unfortunately, some quality parameters are impossible to be satisfied simultaneously. Therefore, the network is divided into slices, where each slice is designed for services with required values of network parameters. In such a case, a crucial problem is secure isolation of slices to prevent inter-slice harmful interaction or even attacks and to provide sufficient Quality of Service in each slice, see [7], [8]. Providing proper slices' isolation is now extensively studied, both, from experimental and theoretical points of view.

Even before the concept of 5G network have been rigorously formulated, experimental research related to future networks caused a need of isolation of functions and processes. Different kinds of testbeds have taken into the isolation property. For instance, the COMCON (COntrol and Management of COexisting Networks) project [9] has been created to design novel control and management mechanisms for supporting the coexistence of networks in the Future Internet. It has considered several use cases to evaluate a reference architecture providing some isolation of specialized networks with certain functionalities to provide their dependable and predictable work. Another testbed, described in [10], was specialized for network experiments with disconnected mobile nodes. Here, the isolation property has been required for precise measurements of network's properties. The paper [11] presents a scheme (possibly inside OMF, the wireless testbeds managerial framework) that exploits wireless testbeds functionality by introducing spectrum slicing of the testbed resources. Since in wireless testbeds slicing there are inter-dependencies among the resources, the isolation of experiments is there a hard task. The paper [12] gives an approach allowing virtualization of testbeds to realize several services like environment control, virtual radio control, slice feedback, and a virtual radio isolation. At least two of them provide the isolation of slices: the environment control is responsible for maintaining control and performance isolation across different environments while virtual radio isolation service is required for isolating the radio resources used by each slice due to the inherent nature of the wireless medium.

Practical experiments concerning isolation are presented in paper [13]. The authors consider OpenVZ and User Mode Linux (UML) for virtualization of the ORBIT wireless testbed and evaluate their relative merit. Their results show that the operating system level virtualization mechanism outperforms UML in terms of system overheads and performance isolation. To measure isolation, they propose two performance measurement metrics: transient response and cross coupling between experiments. The transient response is the instantaneous change in throughput of an experiment running on one slice caused due to time varying change in offered load on another slice, while the cross coupling is the difference in throughput with virtualization as a percentage of the throughput without virtualization. Both measures are estimated in experiments. Another experimental testing of isolation can be found in [14], where the authors compare the container-based approach and general virtual machines (Xen). They show that both approaches give comparable isolation features (with respect to fault isolation, resource isolation, and security isolation), while the container-based approach gives better efficiency expressed in terms of overall performance (throughput, latency, etc.) and/or scalability (measured in number of concurrent VMs), what reflects better performance isolation.

Some approach to numerical evaluation of isolation loss using, both, experimental and theoretical results, has been proposed in the paper [15]. Usually in VM environments, the performance isolation is calculated based on performance loss ratio. For containers, we can consider misbehavior and orchestration and management, so the measurements that only

take performance loss into consideration are not sufficient. In the paper [15] the authors propose a performance isolation measurement model that combines the performance loss and resource shrinkage of containers. They also validate their model experimentally using the open- source container project Docker.

In contradiction to widely applied container-based virtualization [16] as a solution for isolating resources of users or slices, the authors of the paper [17] propose an alternative to enable the isolation, based on commodity OS, utilizing existing features in commodity OS. Assigning every user-id in the OS a dedicated and isolated network the address and the routing table, this method enhances the commodity OS with the property of network name-space isolation.

Except of nodes-located information processing, isolation is also required in network processing. For instance, paper [18] proposes a method to share the host's global IP address for all the guest slivers on a node and isolate their network usage in port-space. In programmable networks all VMs can be configured with the same IP and MAC addresses as the host so that any Ethernet frames from outside can be received by a VM or host. To isolate the packets of different VMs, each VM is assigned with a range of port numbers. The port range of each VM can be got from the database of PLC node. As a result, the corresponding flow entries (forwarding rules) are installed after a VM is launched. Each VM can only listen on the ports that assigned to it. The OVS switches packet based on the destination port number. without address translation. The paper [19] combines the programmable switch OpenFlow with network virtualization and design the INP platform OFIAS, i.e., OpenFlow In A Slice. With the flexibility of OpenFlow and the scalable multiplexing of virtualization, OFIAS can smoothly support multi-party INP with good isolation performance.

Next to purely experimental investigations made in testbeds, more theoretical approaches to isolation can be found in the literature. For instance, in the paper [20] the authors present a framework that improves current infrastructure by extending link virtualization with a new component which they call Multi-Hop Virtual Link. In their proposal this component may be implemented as a tunnel which traverses multi-hop physical nodes. For more virtual links, a Link Switch Engine is applied for strong isolation of switch ability offered to different virtual links. The authors propose to allocate separate hardware to each virtual link providing strong isolation among multiple virtual links.

A next step in making a concept of isolation very practical tool for networks is proposing not only qualitative description and experimental validation of isolation, but also its formal modeling and quantitative representation of its level (estimated of calculated from the model). Such a need has been observed, both, in practical investigations and more theoretical research. For instance, the CONFINE IP Project, Community Networks Testbed for the Future Internet [21], has considered the sliver isolation, covering two different aspects for all resources (nodes and a network): the resources isolation and performance isolation. The resource isolation means that a slice does not interfere with the operation of other slices and is completely separated from the others: it cannot access the data of other slices, cannot kill their processes, and cannot access the core management system (it is secure). Performance isolation means providing mechanisms to guarantee performance on a predictable level with sufficient (up to a certain point) amount of available resources. To make such a concept functional one should describe these intuitive properties and expectation with some measurable parameters.

A milestone of formal modeling of slices is the paper [22], where the authors have presented an abstraction that supports programming isolated slices of the network. The proposed semantics of slices ensures that the processing of packets on a slice is independent of all other slices. Further, in the paper they formally define slices and propose algorithms for compiling slices. Finally, the authors describe a tool for automatic verification of formal isolation properties on a level of network packets processing.

Among experimental papers, a more complete pattern of slices isolation gives the paper [23]. The authors consider several resultant parameters to estimate slices isolation (all for several container-based virtualization implementations), which are: Computing Performance, Memory Performance, Disk Performance, Network Performance, Performance Overhead in HPC Applications, all of them measured according to their own methodology. Finally, they use an Isolation Benchmark Suite (IBS) [24], [25], which includes six different stress tests: CPU intensive test, memory intensive test, a fork bomb, disk intensive test and two network intensive tests (send and receive). Such parameters are suitable for estimating performance of isolated systems in different virtualization environments.

The recent trend of isolation modeling is calculation of overall parameters of the sliced network, like its performance properties, e.g., end-to-end (E2E) delay for a slice [26], using detailed transmission or nodes' parameters. In this paper we propose a new Graph-based model suitable for isolation modeling of slices. It uses several isolation parameters (proposed in our earlier paper [27] and makes possible to establish a common level of isolation for an E2E slice in the 5G network.

The rest of the paper is organized as follows. In Section 2 we introduce a Graph-based model of slices and isolation. It uses hierarchical graphs and makes possible to calculate isolation level on a given level of abstraction. Section 3 is an overview of parameters and properties suitable to model isolation of network's elements. In Section 4 we give a mathematical background for calculating isolation level presenting suitable methods and formulas. In Section 5 we illustrate the theoretical results of previous Sections with two numerical examples, while Section 6 concludes the paper and outlines future work.

II. Graph-based model of slices and isolation

All communication networks can be considered as a set of interconnected layers, including their different elements and roles. The same approach can be also applied to 5G networks

(see Figure 1). The highest layer represents a high-level view on 5G network which includes two main subnetworks: RAN and CN and the gateway between them. The lower layer refers to all resources located in a selected subnet. The next layer contains a physical resource which can be described by some properties and parameters that define it. The required virtual resources can be located in the last layer. The virtualization was made based on the mentioned earlier properties.

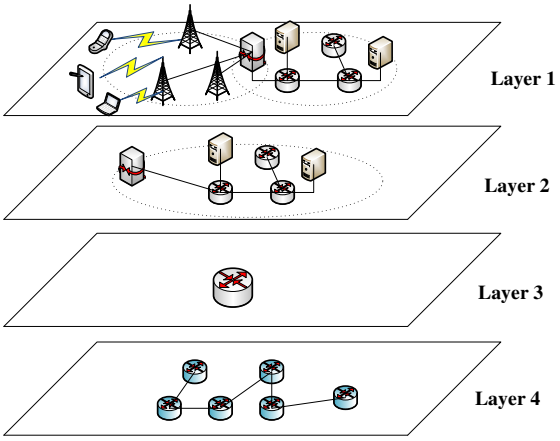


Fig. 1. The 5G network layer decomposition.

Of course, such a general decomposition must be appropriately adapted and transformed to represent a slice-based 5G network with an isolation assurance. For that reason, a new Graph-based layered model of the network has been created.

A. Model assumptions

The described model has been created based on a few assumption, such as:

- RAN, CN and every resource (e.g., router, link, switch, server, etc.) are represented as hypergraphs;
- properties of resources are represented as graph vertices;
- one property can be divided into several virtual properties used in a slice's structure and a value can be assigned to each of them; based on these values it is possible to create requirements how a slice can be created or validate if it is possible to create a slice with a defined set of properties;
- the stratification enables to consider isolation on many levels (e.g., in Layer 5 isolation exists between properties), so it is easy to show a slice as a path with vertices which represent virtual properties.

B. Graph-based model

Our proposed Graph-based architecture is shown in Figure 2. The architecture includes 5 layers. Each of them represents different aspect of slicing:

- Layer 1: the layer consists of hypergraphs with all resources of RAN and CN. Moreover, both areas have common part (common resources which match RAN slice and CN slice, e.g. gateway connecting RAN and CN);
- Layer 2: the layer is represented by selected hypergraph of one of area: RAN or CN with sub-hypergraphs (or a second level hypergraph) The sub-hypergraphs apply to all physical resources in this area;
- Layer 3: the sub-hypergraph in this layer refers to dedicated resource which has several properties (in isolation sense), the graph vertices;
- Layer 4: in this layer a property is a graph vertex which can be virtualized;
- Layer 5: this layer includes all virtual properties, created from a vertex in Layer 4, which are the foundations of slices.

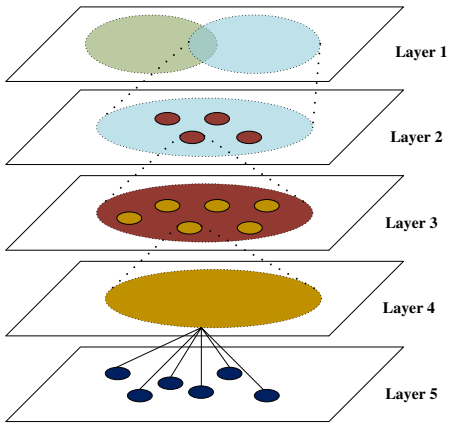


Fig. 2. The 5G Graph-based model.

III. Properties and Parameters of Isolation

In the paper [27] network has been described in terms of *properties* and *parameters*, called in general the *traits*. Both should be normalized; in the paper [27] has been proposed an example method for this operation. The method assumes that each trait's value is preprocessed with the *normalization function* $g : \Lambda \rightarrow \Omega$, where Λ is the trait's domain and Ω is a continuous subsection of the real line $\mathbb{R}$; further in this paper we will assume that $\Omega = [0, 1]$. Different types of traits could have different normalization functions, however the Ω should be common for all the traits. The normalization function for the one trait could change between vertices as well, see [27]. The value $\alpha = \inf_{x \in \Omega} x$ will be assigned to the *worst value* of the trait (from the isolation point of view). The value $\omega = \sup_{x \in \Omega} x$ will be assigned to the *best value* of the trait (from isolation point of view). In the paper [27] there were defined the following trait families:

- *raising trait*: higher trait's value is better (e.g., available link's throughput in $Mbit/s$);
- *falling trait*: lower trait's value is better (e.g., link's BER);

- *Gaussian trait*: trait is the raising trait for $x < \gamma$ and the falling trait for $x > \gamma$, where $\gamma = const$ (e.g., jitter in the packets stream).

The normalization function should satisfy the following assumptions:

- $g(x) \leq g(y)$ iff y is a better trait's value than x;
- $g(\beta) = C$, where C is a constant dependent on the Ω set and the β was defined in [27] as the trait's *typical value*. In the paper [27] was proposed $C = \frac{\alpha+\omega}{2} = 0.5$.

The Table I contains the example of normalization functions for typical values' domains. Some functions could be parametrized by additional parameters independent from functions' arguments, e.g. β, γ, q, r parameters from the Table I.

IV. Calculating the isolation level

At the beginning, we need to define types of vertices. Let us assume that the subset of vertices from the Layer $\mathcal{L}$ is indicated as $\{V_1, V_2, \ldots, V_n\}$. These vertices can be classified as:

1) *similar vertices*, when all of them are described by a common set of traits $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$;
2) *non-similar vertices*, otherwise.

In fact, the second type of vertices is generalization of the first vertices' type. This observation will be used in a further part of this Section.

A. Isolation-merging function

1) Calculation for similar vertices: For the vertex $V_k, k = 1, 2, ..., n$ we will define a vector $\mathbf{I_k}$ of isolation traits as:

$$\mathbf{I_k} \stackrel{\text{def}}{=} \left(p_{k,1}, p_{k,2}, \ldots, p_{k,m}\right)^T. \tag{1}$$

Now we can propose a formula for calculating the isolation level $\mathbf{I}$ for the Layer $\mathcal{L}$ as:

$$\mathbf{I}(\{V_1, V_2, \ldots, V_n\}) \stackrel{\text{def}}{=} F(\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I_n}), \tag{2}$$

where the function $F : (\Omega^m)^n \to \Omega^m$ is a *general merging function*. In this paper we will assume that traits are independent, so the function F can be defined as:

$$F(\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I_n}) = \begin{pmatrix} f_1(p_{1,1}, p_{2,1}, \ldots, p_{n,1}) \\ f_2(p_{1,2}, p_{2,2}, \ldots, p_{n,2}) \\ \vdots \\ f_m(p_{1,m}, p_{2,m}, \ldots, p_{n,m}) \end{pmatrix}, \tag{3}$$

where $f_i : \Omega^n \to \Omega$, for $i = 1, 2, \ldots, m$, is a *merging function*. This model is a first level of an approximation, where each trait could be changed independently and one's trait's value does not affect other trait's value.

Now let consider the merging function $f(x_1, x_2, \ldots, x_n)$ and mark as $x_{min} = \min\{x_1, x_2, \ldots, x_n\}$ and $x_{max} = \max\{x_1, x_2, \ldots, x_n\}$. Let assume that the function satisfies the following assumptions:

$$(\forall x \in \Omega) \quad f(x, x, \ldots, x) = x, \tag{4}$$

$$\begin{gathered}(\forall 1 \leq k \leq n)(\forall x_k \leq y_k) \\ f(x_1, x_2, \ldots, x_k, \ldots, x_n) \leq f(x_1, x_2, \ldots, y_k, \ldots, x_n).\end{gathered} \tag{5}$$

Those assumptions define our view over the isolation - if the system is built from the components with the same trait's value, then the system has the same trait value as those components (4), and the isolation will not decrease when a component of a system has been enhanced (5). From these two assumptions one can deduce the following inequalities:

$$\begin{gathered}f(x_1, x_2, \ldots, x_n) \leq f(x_{max}, x_2, \ldots, x_n) \leq \\ \leq f(x_{max}, x_{max}, \ldots, x_n) \leq \cdots \leq f(x_{max}, x_{max}, \ldots, x_{max}) = \\ = x_{max},\end{gathered} \tag{6}$$

$$\begin{gathered}f(x_1, x_2, \ldots, x_n) \geq f(x_{min}, x_2, \ldots, x_n) \geq \\ \geq f(x_{min}, x_{min}, \ldots, x_n) \geq \cdots \geq f(x_{min}, x_{min}, \ldots, x_{min}) = \\ = x_{min}.\end{gathered} \tag{7}$$

The function which satisfies the inequalities:

$$x_{min} \leq f(x_1, x_2, \ldots, x_n) \leq x_{max} \tag{8}$$

will be called *the mean function*, see [28]. An example of a mean function is the generalized weighted mean (the power mean):

$$\omega_q(x_1, x_2, \ldots, x_n) = \left(\frac{\sum_{i=1}^n w_i x_i^q}{\sum_{i=1}^n w_i}\right)^{\frac{1}{q}}, \tag{9}$$

where $\sum_{i=1}^n w_i > 0$ and $w_i \geq 0$. The parameter q is a real number; there are the following important border-case formulas:

$$\begin{cases}\omega_{-\infty}(x_1, x_2, \ldots, x_n) = \min\{x_1, x_2, \ldots, x_n\}, \\ \omega_0(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 x_2 \ldots x_n}, \\ \omega_{+\infty}(x_1, x_2, \ldots, x_n) = \max\{x_1, x_2, \ldots, x_n\}.\end{cases} \tag{10}$$

2) Calculations for non-similar vertices: More common situation is when not all considered vertices are described only by one set of traits. Let us assume that the vertex V_k is described by the set of traits $\Pi_k = \{\pi_{k,1}, \pi_{k,2}, \ldots, \pi_{k,m_k}\}$ and let us define $\Pi = \cup_{k=1}^m \Pi_k = \{\pi_1, \pi_2, \ldots, \pi_{|\Pi|}\}\}$. The goal of the reasoning is to represent the merged isolation in a space common for all vertices. Now, let us introduce the set $\Omega^* = \Omega \cup \theta$, $\theta \notin \Omega$, which has the following properties:

$$\begin{cases}(\forall x \in \Omega) & x \not> \theta, \\ (\forall x \in \Omega) & x \not< \theta, \\ (\forall x \in \Omega) & x \neq \theta.\end{cases} \tag{11}$$

The element θ is a special element, e.g. the imaginary unit i satisfies (11) when $\Omega = [0; 1]$. The aim of this value is to indicate that for considered trait and vertex the trait's value does not exists (it is undefined). Let us define the function $T_k : \Omega^{m_k} \to (\Omega^*)^{|\Pi|}$:

$$T_k \begin{pmatrix} p_{k,1} \\ p_{k,2} \\ \vdots \\ p_{k,m_k} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} r_{k,1} \\ r_{k,2} \\ \vdots \\ r_{k,|\Pi|} \end{pmatrix}, \tag{12}$$

TABLE I
EXAMPLES OF TYPICAL NORMALIZATION FUNCTIONS, BASED ON [27]

Λ	**Parameter family**	**The worst value**	**The best value**	**Typical value**	**Function** $g(x)$
$\mathbb{R}$	raising\falling	0	$\pm\infty$	β	$g(x) = 1 - 2^{-\frac{x}{\beta}}$
$\mathbb{R}$	falling\rising	$\pm\infty$	0	β	$g(x) = 2^{-\frac{x}{\beta}}$
$\mathbb{R}$	raising	$-\infty$	$+\infty$	$\beta;\beta \neq 0$	$g(x) = \left(1 + e^{\frac{\beta - x}{\lvert\beta\rvert}}\right)^{-1}$
$\mathbb{R}$	raising	$-\infty$	$+\infty$	0	$g(x) = (1 + e^{-x})^{-1}$
$\mathbb{R}$	falling	$+\infty$	$-\infty$	$\beta;\beta \neq 0$	$g(x) = \left(1 + e^{\frac{x - \beta}{\lvert\beta\rvert}}\right)^{-1}$
$\mathbb{R}$	falling	$-\infty$	$+\infty$	0	$g(x) = (1 + e^{x})^{-1}$
$[q;r] \in \mathbb{R}_{\geq 0}$	raising	q	r	$\frac{q+r}{2}$	$g(x) = \frac{x-q}{r-q}$
$[q;r] \in \mathbb{R}_{\geq 0}$	falling	r	q	$\frac{q+r}{2}$	$g(x) = \frac{r-x}{r-q}$
$\mathbb{R}$	Gaussian	$\pm\infty$	γ	β	$g(x) = 2^{-\left\lvert\frac{x-\gamma}{\beta-\gamma}\right\rvert}$

where

$$r_{k,j} = \begin{cases} \text{value of } \pi_j \text{ for } V_k, & \text{when } \pi_j \in \Pi_k \\ \theta, & \text{otherwise.} \end{cases} \tag{13}$$

Now, one can define a general merging function F^* for non-similar vertices as:

$$F^* : (\Omega^*)^{n|\Pi|} \to (\Omega^*)^{|\Pi|}. \tag{14}$$

According to equation (3), we can write:

$$F^*(\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I_n}) = \begin{pmatrix} f_1^*(r_{1,1}, r_{2,1}, \ldots, r_{n,1}) \\ f_2^*(r_{1,2}, r_{2,2}, \ldots, r_{n,2}) \\ \vdots \\ f_{|\Pi|}^*(r_{1,|\Pi|}, r_{2,|\Pi|}, \ldots, r_{n,|\Pi|}) \end{pmatrix}, \tag{15}$$

where the function $f^* : (\Omega^*)^n \to \Omega^*$ is a non-similar vertex's version of the merging function f. According to equations (4,5), the following assumptions upon the f^* could be made:

$$(\forall x \in \Omega^*) \quad f^*(x, x, \ldots, x) = x, \tag{16}$$

$$\begin{gathered} (\forall 1 \leq k \leq n)(\forall x_k \leq y_k) \\ f^*(x_1, x_2, \ldots, x_k, \ldots, x_n) \leq f^*(x_1, x_2, \ldots, y_k, \ldots, x_n), \end{gathered} \tag{17}$$

$$f^*(x_1, x_2, \ldots, x_n) = \theta \iff x_1 = x_2 = \cdots = x_n = \theta. \tag{18}$$

The equation (17) requires that the values x_k and y_k must be comparable. The θ element does not satisfy this condition for any other value from the Ω^* set, according to the (11). Consequently, when one defines trait's value for an already existing vertex, the merged trait's value with isolation merging function could be higher, lower or stay at the same point.

The $f^*(x_1, x_2, \ldots, x_n)$ function could be calculated in the following way: let the $Z = (z_1, z_2, \ldots, z_{n'})$ be a string of elements from $X = (x_1, x_2, \ldots, x_n)$ created by selecting all elements except the elements equal to θ. Then, we can use e.g. (9) formula for calculations, using only values from Z. The weights should be the same for the element $z_k; 1 \leq k \leq n'$ and the corresponding element from X.

3) Choosing the merging function: The following aspects should be considered for choosing an appropriate merging function.

- Interpretation of the merged trait, e.g. available throughput for a path of vertices is upper-bounded by a minimal value, so the merging function could be defined as $f(x_1, x_2, \ldots, x_n) = \min\{x_1, x_2, \ldots, x_n\}$.
- Implementation constraints: the integer-valued weights and function's parameters could result in more accurate and faster calculations; the operations upon integers are faster than on typical IEEE 754 [29] double precision numbers. The *money-like* types which allow exact operations are slower than hardware supported types.
- Precision constraints: each operation on non-integer number suffers from the finite precision problem, which causes losing the information on the less-important part of a number. Using large number of operations (multiplication, adding, power) leads to very uncertain results.
- The expected value of a merging function: the normalization function defines the central element $C = g(\beta)$ of Ω (i.e., 0.5 for the set $\Omega = [0, 1]$) as a typical value which should be close to the expected value of the merging function. This value could depend on the number of merged properties and the function's internal parameters, like the parameter q for the generalized mean. The Figure 3 shows how the mean value changes for the generalized mean function. Only the arithmetical mean ($q = 1$) from this functions family satisfies, for all $n > 1$, the equation:

$$\mathbb{E}\left(f_q(x_1, x_2, \ldots, x_n)\right) = g(\beta). \tag{19}$$

- The shape (e.g. convexness) of a merging function: traits are interpreted with some logic, e.g.:

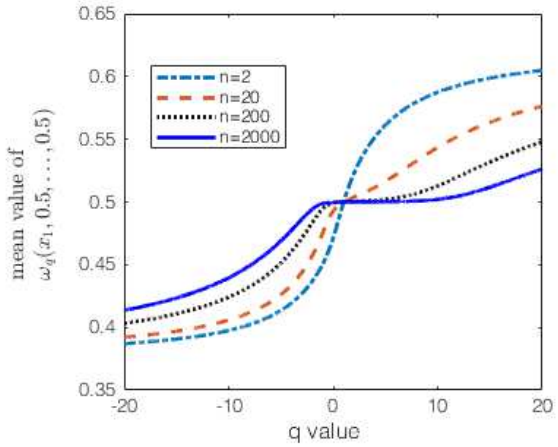


Fig. 3. Mean value of ω_q function for various number of traits n and the parameter q.

- *The OR logic*: the merged trait's value is more strongly dependent on large values than on small ones. This kind of traits should be merged with convex functions. An example trait is *number of CVE vulnerabilities* for a network's node.
- *The AND logic*: the merged trait's value is more strongly dependent on small values than on large ones. This kind of traits should be merged with concave functions. An example trait is *encryption strength* in number of key's bits for a network's node.
- *The neutral logic*: the merged trait's value depends on the small and large values equally. This kind of traits should be merged with function which satisfies the following equation:

$$\frac{\partial^2}{\partial x_k^2} f(x_1, x_2, \ldots, x_k, \ldots, x_n) = 0. \tag{20}$$

B. Calculating the isolation vector

The isolation vector for a network could be calculated recursively by calculating isolation vectors for subnetworks and a group of vertices from a layer. This *bottom-up* method allows adding special information, which should be included into the calculation process, and which cannot be defined for vertices in lower layers.

C. The comparison problem for vectors

From practical point of view, it is very important to compare calculated isolation vectors which are defined over a common set of traits. Two vectors, which have only one different trait's value, are easy to compare. When multiple traits' values are different, the situation is much more complicated. In mathematical terms the aim is to define the linear order for the set $(\Omega)^{|\Pi|}$; the θ value could be omitted, because each trait should be defined for at least one vertex in a graph, so the final vector has all non-θ values. Such an order could be defined by assigning to each of isolation vectors a number, which can be compared. This assignment is provided by *the extracting function* $\Phi : (\Omega)^{|\Pi|} \to \Omega$. This function should satisfy following inequalities for each $x_k \in \Omega$:

$$\frac{\partial}{\partial x_k} \Phi(x_1, x_2, \ldots, x_k, \ldots, x_n) \geq 0, \tag{21}$$

$$\frac{\partial^2}{\partial x_k^2} \Phi(x_1, x_2, \ldots, x_k, \ldots, x_n) \leq 0. \tag{22}$$

The inequality (21) means that the extracting function is raising for each of its parameters. The inequality (22) describes the assumption that the function is concave for each parameter.

D. Extracting the single value

The following example function family, which satisfies the assumptions (21, 22) could be used for extracting the single value for an isolation vector:

$$\Phi_q(x_1, x_2, \ldots, x_k, \ldots, x_n) = 1 - \left(\frac{1}{n} \sum_{i=1}^{n} (1 - x_i)^q \right)^{\frac{1}{q}}. \tag{23}$$

This function has the following partial derivatives:

$$\begin{aligned} &\frac{\partial}{\partial x_k} \Phi_q(x_1, x_2, \ldots, x_k, \ldots, x_n) = \\ &= \left(\sum_{i=1}^{n} (1 - x_i)^q \right)^{\frac{1}{q}-1} \frac{(1 - x_k)^{q-1}}{n} \geq 0, \end{aligned} \tag{24}$$

$$\begin{aligned} &\frac{\partial^2}{\partial x_k^2} \Phi_q(x_1, x_2, \ldots, x_k, \ldots, x_n) = \\ &= \frac{1 - q}{n^2} \left(\sum_{i=1}^{n} (1 - x_i)^q \right)^{\frac{1}{q}-2} (1 - x_k)^{q-2} \sum_{i=1; i \neq k}^{n} (1 - x_i)^q . \end{aligned} \tag{25}$$

If $q \geq 1$, then the second derivative calculated in the equation (25) is non-positive, so the assumption defined in equation (22) is satisfied.

V. Examples

In this section we consider two examples: calculating the isolation of a single node in a network and calculating the isolation over an E2E path for a single slice. The second example contains a list of steps which are included in the isolation assessment process.

A. Example 1. Single node

Let us consider a single network element with traits, traits' values and normalization functions defined in the Table II. The normalization functions are fitted to expected traits' domains, which is the main reason during selecting the normalization function. We assumed for the trait *symmetric encryption algorithm's strength* that at this moment the largest available (and practically used) key size is 256 bits. Theoretically, the key could have any length (e.g. one-time keys for stream ciphers), but very large key size is impractical as well, so in such typical case will not be considered as an option. Since the domain is constrained to the range of integer numbers, the linear normalization function for this trait was chosen.

TABLE II
SET OF NORMALIZATION FUNCTIONS AND NORMALIZED VALUES FOR A SINGLE NODE

Parameter	Value	Normalization function	Normalized value
Symmetric encryption algorithm's strength $\begin{cases} a = 0 \\ b = 256bits \end{cases}$	160 bits	$g(x) = \frac{x-a}{b-a}$	0.625
Average time between vulnerabilities assessments $\beta = 4h$	$8h$	$g(x) = 2^{-\frac{x}{\beta}}$	0.25
Amount of electromagnetic radiation $\beta = 25dB\mu V/m$	$55dB\mu V/m$	$g(x) = \left(1 + e^{\frac{x-\beta}{\lvert\beta\rvert}}\right)^{-1}$	0.2315

The isolation vector for this vertex (which is a representation of this single node scenario) is $(0.625, 0.25, 0.2315)$. We can *extract the single value* from this vector using the function $\Phi_2(x_1, x_2, x_3)$:

$$I_{final} = 1 - \sqrt{\frac{\sum_{i=1}^{3}(1-x_i)^2}{3}} = \\ = 1 - \sqrt{\frac{(1-0.625)^2 + (1-0.25)^2 + (1-0.2315)^2}{3}} = \\ = 0.3433. \tag{26}$$

B. Example 2. A simple end-to-end slice

The process of isolation analysis is defined as follows:

1) definition of use-case to be modeled;
2) definition of all network resources to be modeled (nodes and links);
3) definition of each resource affiliation to RAN and/or CN;
4) for each resource definition of relevant isolation parameters and properties to be determined or measured;
5) for CN, RAN and E2E definition of relevant isolation parameters and properties to be determined or measured;
6) choosing a set of functions to normalize isolation parameters and properties;
7) choosing a set of functions to calculate isolation from parameters and properties;
8) choosing a set of functions to compare two or more isolation tuples;
9) definition of a slices spanned across previously defined resources;
10) for each layer, calculation of slices' isolation, with a previously chosen set of functions;
11) performing comparison, if needed, with a previously chosen set of comparison functions.

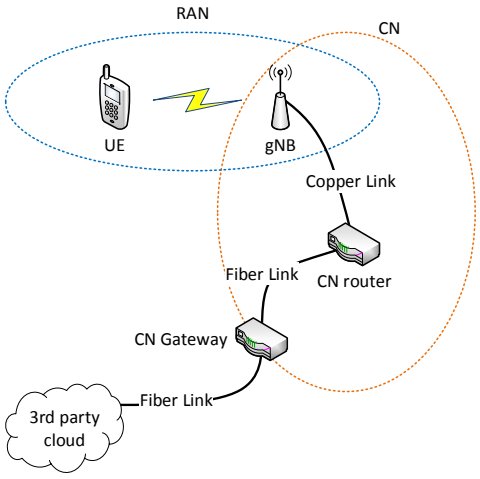


Fig. 4. The example of a network scenario for the purpose of isolation analysis.

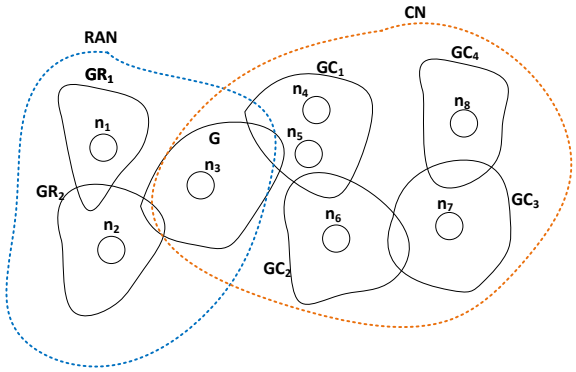


Fig. 5. The representation of the selected piece of network in the Graph model.

As an example, the simple network is analyzed in context of E2E calculation of isolation (see Figure 4). RAN network consists of a single User Equipment (UE), which connects with gNB node using Radio Link. CN consists of three types of equipment: gNB (which is common for RAN and CN), CN Router and CN Gateway. gNB is connected with CN router using Copper Link and CN Router is connected with CN Gateway using Fiber Link. CN Gateway is connected with a 3rd-party-vendor cloud using Fiber Link, which is considered outside of RAN and CN.

The described piece of network can be represented as a graph structure using the Graph model (see Figure 5)

RAN and CN are the hypergraphs which include sub-hypergraphs: GR_x, GC_x and G (GR_x - RAN sub-hypergraph of the element x, GC_x - CN sub-hypergraph of the element x, G sub-hypergraph of the gateway). Each sub-hypergraph has several nodes (n_i) which present the properties. Every property can be virtualized: v_1n_1, v_2n_1, v_3n_1, etc. (v_in_j means a virtual property "i" of a physical property "j").

According to the process of E2E isolation analysis, each step may be described as follows:

1) **Use case definition:** in a network depicted in the Figure 4 a slice is being configured from UE to CN gateway. There is a need to assess isolation level of the slice based on the equipment and media used to serve this slice. There are two types of routers in CN which can be chosen in the network's implementation (it is assumed that each of them has assessed the same chosen parameters and properties). There is a need to assess the isolation of the slice while the first and the second types of routers are chosen. The isolation in these two cases should be compared to choose a solution with better isolation for a slice. The analysis should be performed only on Layers 1, 2 and 3 of the model.
2) **Network resources to be modeled:** in the analysis there should be modeled the following resources:
   - User Equipment in RAN;
   - Radio Link in RAN;
   - gNB as a gateway between RAN and CN (affiliated to RAN and CN);
   - Copper Link in CN;
   - CN Router in CN;
   - Fiber Link in CN;
   - CN Gateway as a gateway between CN and 3rd-party-cloud (affiliated to CN).

   Figure 5 represents the network for the isolation analysis in a graph form.
3) **RAN-CN affiliation:** the affiliation was described in the previous point.
4) **Parameters and properties for elements:** the following isolation parameters and properties were identified for each resource:
   - User Equipment in RAN (GR_1):
     - slicing application: programming language used (e.g., Java), enumeration (n_1).
   - Radio Link in RAN (GR_2):
     - symmetric encryption algorithm strength: the number of bits, a nonnegative integer (n_2).
   - gNB as a gateway between RAN and CN (G):
     - average time between vulnerabilities assessments: hours, a nonnegative real number (n_3).
   - Copper Link in CN (GC_1):
     - symmetric encryption algorithm strength: the number of bits, a nonnegative integer (n_4);
     - amount of electromagnetic radiation: i.e., dBV/m, a real number (n_5).
   - CN Router in CN (GC_2):
     - average time between vulnerabilities assessments: hours, a nonnegative real number (n_6).
   - Fiber Link in CN (GC_3):
     - symmetric encryption algorithm strength: the number of bits, a nonnegative integer (n_7).
   - CN Gateway as a gateway between CN and 3rd-party-cloud (GC_4):
     - average time between vulnerabilities assessments: hours, a nonnegative real number (n_8).
5) **Parameters and properties for CN, RAN and E2E:** the following isolation parameters and properties were identified:
   - Core Network (CN):
     - symmetric encryption algorithm strength;
     - amount of electromagnetic radiation;
     - average time between vulnerabilities assessment.
   - Radio Access Network (RAN):
     - symmetric encryption algorithm strength;
     - programming language used;
     - average time between vulnerabilities assessment.
   - End-to-End (E2E):
     - symmetric encryption algorithm strength;
     - programming language used;
     - amount of electromagnetic radiation;
     - average time between vulnerabilities assessment;
     - (produced by extracting the single value from other traits) isolation level.
6) **Choose a set of functions to normalize isolation parameters and properties:** These functions are defined in the Table III.
7) **Choose a set of functions to calculate isolation from parameters and properties:** We will use as *the isolation-merging function* the following formula:

$$\omega_{-1}(x_1, x_2, \ldots, x_n). \tag{27}$$

   for calculating the isolation inside the RAN or CN part of the network and the weighted version of this mean for calculating isolation E2E for these two parts of the network. The value $q = -1$ is used, because it is very fast to implement, it is a merging function with the *AND logic*. The Figure 3 shows that this function has the mean value very close to 0.5 (the central element of the Ω set), which is advisable.
   We use the weighted function for merging isolation between CN and RAN for flattening an impact of each graph's vertex in results. The weights are defined in the column *Weights* in the Table V. Since the gNB node (and in consequence its vertex in the Graph model) belongs to RAN and CN as well, its impact is doubled. To avoid this excessive influence, the subnetworks on the path should be separated.
8) **Choose a set of functions to compare two or more isolation tuples:** We will use the proposed method for calculating the final isolation and the calculated final isolations (the I_{final} values) will be used for comparisons of the slices' isolation levels.
9) **Define a slice spanned across previously defined resources:** We assume in this scenario that the slice is from UE to the 3rd-party-cloud and contains all devices and links between these nodes.
10) **For each layer, calculate slices isolation with a previously chosen set of functions:** The results of

TABLE III
SET OF NORMALIZATION FUNCTIONS AND NORMALIZED VALUES

Part	Element	Parameter	Value	Normalization function	Normalized value	Typical values / margin values
RAN	UE in RAN	Programming language used	C++ (0.75)	N/A	0.75	N/A
	Radio Link in RAN	Symmetric encryption algorithm strength	160 bits	$g(x) = \frac{x-a}{b-a}$	0.625	$\begin{cases} a = 0 \\ b = 256bits \end{cases}$
gNB	gNB as a gateway between RAN and CN	Average time between vulnerabilities assessments	8h	$g(x) = 2^{-\frac{x}{\beta}}$	0.25	$\beta = 4h$
CN	Copper Link in CN	Symmetric encryption algorithm strength	128 bits	$g(x) = \frac{x-a}{b-a}$	0.5	$\begin{cases} a = 0 \\ b = 256bits \end{cases}$
		Amount of electromagnetic radiation	$55dB\mu V/m$	$g(x) = \left(1 + e^{\frac{x-\beta}{\lvert\beta\rvert}}\right)^{-1}$	0.2315	$\beta = 25dB\mu V/m$
	CN Router in CN	Average time between vulnerabilities assessments	12h	$g(x) = 2^{-\frac{x}{\beta}}$	0.125	$\beta = 4h$
	Fiber Link in CN	Symmetric encryption algorithm strength	256 bits	$g(x) = \frac{x-a}{b-a}$	1	$\begin{cases} a = 0 \\ b = 256bits \end{cases}$
	CN Gateway as a gateway between CN and 3rd party cloud	Average time between vulnerabilities assessments	4h	$g(x) = 2^{-\frac{x}{\beta}}$	0.5	$\beta = 4h$

TABLE IV
ISOLATION INSIDE RAN AND CN

Part	Parameter	Values	Merged values
RAN	Average time between vulnerabilities assessments	0.25	0.25
	Encryption algorithm strength	0.625	0.625
	Programming language used	0.75	0.75
CN	Encryption algorithm strength	0.5, 1	$\frac{2}{\frac{1}{0.5}+\frac{1}{1}} = 0.6667$
	Amount of electromagnetic radiation	0.2315	0.2315
	Average time between vulnerabilities assessments	0.25, 0.125, 0.5	$\frac{3}{\frac{1}{0.25}+\frac{1}{0.125}+\frac{1}{0.5}} = 0.2143$

TABLE V
ISOLATION VECTOR VALUES

Parameter	Values	Weights	Isolation vector's values
Programming language used	0.75	1	0.75
Encryption algorithm strength	0.625, 0.6667	1/3, 2/3	$\frac{1}{\frac{1/3}{0.625}+\frac{2/3}{0.6667}} = 0.6522$
Average time between vulnerabilities assessments	0.25, 0.2143	1/4, 3/4	$\frac{1}{\frac{1/4}{0.25}+\frac{3/4}{0.2143}} = 0.2222$
Amount of electromagnetic radiation	0.2315	1	0.2315

calculations are in Tables IV and V. From the obtained results presented in the Table V, we can build the *isolation vector* $(0.75, 0.6522, 0.2222, 0.2315)$ for this example network. One can *extract the single value* from this vector using the function $\Phi_2(x_1, x_2, x_3, x_4)$:

$$I_{final} = 1 - \sqrt{\frac{\sum_{i=1}^{4}(1-x_i)^2}{4}} = 0.4128. \qquad (28)$$

11) **Perform comparison, if needed, with a previously chosen set of comparison functions:** The defined slice has I_{final} in a medium level and it could be improved. From the Table V we can choose the worst isolation trait (by its value): **Average time between vulnerabilities assessments**. In this scenario this trait is merged from RAN and CN with different weights and with different values. The vertex to improve could be determined by exhaustive search where all traits, except current vertex's traits, has origin values and the current vertex's trait's value is set to 1 (the ω value). The result of this search is summarized in the Table VI. The *CN Router in CN*'s trait's value should be improved, because enhancement of this element could make the biggest effort on the I_{final} value. Let us assume now, that we improved this trait's value to $4h$ (0.5 after normalization). After this operation, the value of this trait in the isolation vector is 0.3333 and the $I_{final} = 0.4481$.

C. Discussion of the results

The I_{final} values are below the 0.5, which should be expected, because the traits' values in both examples are generally low or medium. In this scenario the trait's values are higher than 0, so the ω_{-1} function could be used for merging isolation. If zero values for the traits are expected, the parameter q should be greater than 0. Such a situation

TABLE VI
SEARCHING FOR THE BEST VERTEX TO IMPROVE THE WORST TRAIT: THE RESULTS

Vertex	**Origin trait's value**	**Max. available merged value in RAN**	**Max. available merged value in CN**	**Max. trait's value in the isolation vector**
G - gNB as a gateway between RAN and CN	0.25	1	0.2727	0.3333
GC_2 - CN Router in CN	0.125	0.25	0.4286	**0.3636**
GC_4 - CN Gateway as a gateway between CN and 3rd party cloud	0.5	0.25	0.2308	0.2353

could happen if linear functions are used for normalization or one of the properties' values is zero.

A change in the Step 11 for the Example 2 is small, because in the isolation vector there exists a parameter with the value 0.2315, which has a significant impact on the I_{final} and which should be improved.

The procedure for finding the best vertex and its trait to improve the isolation is defined in a heuristic way. There is a space for further improvements and research. In this case we assumed that each trait has the same merging function, but it could be trait-dependent, and it should be considered by the algorithm executed in the Step 11. This algorithm also could consider the cost of each trait's improvement in a vertex.

VI. CONCLUSIONS AND FUTURE WORK

In this paper the Graph-based quantitative description of networks' slices isolation has been proposed. We have presented a model which can be used to represent in a transparent way a slice in 5G Network, allowing its detailed analysis and supporting additional calculations. As an example of such calculations we have considered the problem of estimating the isolation level of the end-to-end slice and also isolation level of each network's element, both, physical and virtual. Moreover, we have proposed a general framework and the mathematical rules defining how the isolation of that slice can be calculated. Finally, in the paper we have included examples of isolation calculation for a single node and for the end-to-end scenario of a single slice. The numerical results proved to be promising, indicating possibilities of application of our approach in slices management and optimization.

The paper presents a research on its initial state. In our opinion the future work on the presented topic should be continued. Among others, it should include the following issues:

- Validation and verification of the presented Graph model by modeling different scenarios;
- Development of methods for comparing different types of slices;
- Proof of concept application development;
- Development of a set of parameters and properties for devices and links;
- Development of better algorithm for selecting a trait and graph's vertex to improve the isolation;
- Research of integration of the presented model with 5G MANO systems.

REFERENCES

[1] L. Peterson, T. Anderson, D. Culler, and T. Roscoe, "A blueprint for introducing disruptive technology into the Internet," *ACM SIGCOMM Computer Communication Review*, vol.33, iss.1, pp.59-64, 2003, https://doi.org/10.1145/774763.774772.

[2] C. Chapman, S. Ward, "Description of Network Slicing Concept", NGMN Alliance, January, 2016, https://www.ngmn.org/publications/all-downloads/article/description-of-network-slicing-concept.html.

[3] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end Network Slicing for 5G Mobile Networks," *Journal of Information Processing*, vol.25, pp.153-163, Feb. 2017, https://doi.org/10.2197/ipsjjip.25.153.

[4] Nokia "Dynamic end-to-end network slicing for 5G," (White Paper), 2016.

[5] T. Shimojo, Y. Takano, A. Khan, S. Kaptchouang, M. Tamura, and S. Iwashina, "Future mobile core network for efficient service operation," in *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, London, 2015, pp.1-6, https://doi.org/10.1109/NETSOFT.2015.7116190.

[6] U. Herzog, A. Georgakopoulos, I.-P. Belikaidis, M. Fitch, K. Briggs, S. Diaz, O. Carrasco, K. Moessner, B. Miscopein, S. Mumtaz, and P. Demestichas, "Quality of service provision and capacity expansion through extended-DSA for 5G," *Transaction of Emerging Telecommunications Technologies*, vol.27, iss.9, pp.1250-1261, September 2016, https://doi.org/10.1002/ett.3061.

[7] Z. Kotulski, T. Nowak, M. Sepczuk, M. Tunia, R. Artych, K. Bocianiak, T. Osko, J.-P. Wary, "On end-to-end approach for slice isolation in 5G networks. Fundamental challenges", *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, in: M. Ganzha, L. Maciaszek, M. Paprzycki (Eds). ACSIS, vol.11, pp.783-792, 2017, https://doi.org/10.15439/2017F228.

[8] Z. Kotulski, T. Nowak, M. Sepczuk, M. Tunia, R. Artych, K. Bocianiak, T. Osko, J.-P. Wary, "Towards constructive approach to end-to-end slice isolation in 5G networks", *EURASIP Journal of Information Security*, vol. 2018:2, pp.1-16, 2018, https://doi.org/10.1186/s13635-018-0072-0.

[9] D. Schlosser, M. Hoffmann, T. Hoßfeld, M. Jarschel, A. Kirstaedter, W. Kellerer, S. Kohler, "COMCON: Use Cases for Virtual Future Networks", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.584-586, 2011, https://doi.org/10.1007/978-3-642-17851-1_48.

[10] J. White, G. Jourjon, T. Rakatoarivelo, M. Ott, "Measurement Architectures for Network Experiments with Disconnected Mobile Nodes", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.315-330, 2011, https://doi.org/10.1007/978-3-642-17851-1_26.

[11] A.-Ch. Anadiotis, A. Apostolaras, D. Syrivelis, T. Korakis, L. Tassiulas, L. Rodriguez, I. Seskar, M. Ott, "Towards Maximizing Wireless Testbed Utilization Using Spectrum Slicing", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.299-314, 2011, https://doi.org/10.1007/978-3-642-17851-1_25.

[12] G. Bhanage, I. Seskar, D. Raychaudhuri, "A Service Oriented Experimentation Framework for Virtualized WiMAX Systems", in: T. Korakis et al. (Eds.): *TridentCom 2011*, LNICST 90, pp.152-161, 2012, https://doi.org/10.1007/978-3-642-29273-6_12.

[13] G. Bhanage, I. Seskar, Y. Zhang, D. Raychaudhuri, S. Jain, "Experimental Evaluation of OpenVZ from a Testbed Deployment Perspective", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.103-112, 2011, https://doi.org/10.1007/978-3-642-17851-1_7.

[14] S. Soltesz, H. Potzl, M.E. Fiuczynski, A. Bavier, L. Peterson, "Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors", *Proceeding of EuroSys'07 Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems*, pp.275-287, 2007. https://doi.org/10.1145/1272996.1273025.

[15] C. Zhao, Y. Wu, Z. Ren, W. Shi, Y. Ren, J. Wan, "Quantifying the Isolation Characteristics in Container Environments". in: X. Shi et al. (Eds.): *Network and Parallel Computing. NPC 2017*. Lecture Notes in Computer Science, vol.10578, Springer, https://doi.org/10.1007/978-3-319-68210-5_17.

[16] Sh. Ma, J. Jiang, B. Li, B. Li, "Maximizing Container-based Network Isolation in Parallel Computing Clusters", *IEEE 24th International Conference on Network Protocols (ICNP)*, 2016, https://doi.org/10.1109/ICNP.2016.7784434.

[17] M. Chen, A. Nakao, "Feather-Weight Network Namespace Isolation Based on User-Specific Addressing and Routing in Commodity OS", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.53-68, 2011, https://doi.org/10.1007/978-3-642-17851-1_4.

[18] P. Du, M. Chen, A. Nakao, "Port-Space Isolation for Multiplexing a Single IP Address through Open vSwitch", in: T. Magedanz et al. (Eds.): *TridentCom 2010*, LNICST 46, pp.113-122, 2011, https://doi.org/10.1007/978-3-642-17851-1_8.

[19] P. Du, M. Chen, A. Nakao, "OFIAS: A Platform for Exploring In-Network Processing", in: T. Korakis et al. (Eds.): *TridentCom 2011*, LNICST 90, pp.142-151, 2012, https://doi.org/10.1007/978-3-642-29273-6_11.

[20] Sh. Ma, B. Wang, X. Zhang, T. Li, "An Evolving Architecture for Network Virtualization", in: V.C.M. Leung et al. (Eds.): *TridentCom 2014*, LNICST 137, pp.379-386, 2014. , https://doi.org/10.1007/978-3-319-13326-3_36.

[21] CONFINE Project, Community Networks Testbed for the Future Internet, http://confine-project.eu/

[22] S. Gutz, A. Story, C. Schlesinger, N. Foster, "Splendid Isolation: A Slice Abstraction for Software-Defined Networks", *HotSDN'12*, August 13, 2012, Helsinki, Finland, https://doi.org/10.1145/2342441.2342458.

[23] Miguel G. Xavier, Marcelo V. Neves, Fabio D. Rossi, Tiago C. Ferreto, Timoteo Lange, Cesar A. F. De Rose, "Performance Evaluation of Container-based Virtualization for High Performance Computing Environments", *21st Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 2013, https://doi.org/10.1109/PDP.2013.41.

[24] "Isolation Benchmark Suite", 2012. [Online]. Available: http://web2.clarkson.edu/class/cs644/isolation.

[25] J.N. Matthews, W. Hu, M. Hapuarachchi, T. Deshane, D. Dimatos, G. Hamilton, M. McCabe, J. Owens, "Quantifying the Performance Isolation Properties of Virtualization Systems", *ExpCS'07*, 13-14 June, 2007, San Diego, CA, https://doi.org/10.1145/1281700.1281706.

[26] D. Sattar, A. Matrawy, "Optimal Slice Allocation in 5G Core Networks", *arXiv:1802.04655 [cs.NI]*, 2018.

[27] Z. Kotulski, T. Nowak, M. Sepczuk, M. Tunia, "5G networks: types of isolation and their parameters in RAN and CN slices", submitted.

[28] P.P Korovkin, *Inequalities*. Moscow : Mir Publishers, 1975.

[29] IEEE. *IEEE Standard for Floating-Point* - IEEE Std 754-2008. 2008.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 381–385
DOI: 10.15439/2018F234
ISSN 2300-5963 ACSIS, Vol. 15

Improving pseudorandom generator on cellular automata with bent functions

Mukhamedjanov Daniyar*, Ryaskin Gleb†, Levina Alla‡ and Kaplun Dmitrii§
*ITMO University
49 Kronverksky pr., St. Petersburg, 197101, Russia Email: danmd.info@gmail.com
†ITMO University
Email: ryaskingleb20@gmail.com
‡ITMO University
Email: alla_levina@mail.ru
§Saint-Petersburg Electrotechnical University "LETI"
Email: dikaplun@etu.ru

***Abstract*—Nowadays the practice of researching pseudorandom number generators (PRNG) becomes more scalable because of its spreading in many spheres of computer science and, especially, cybersecurity. The problem is that existing generators are still have many disadvantages in terms of velocity, complexity or flexibility. Thus, the area of researching new algorithms of generating pseudorandom sequences is more than just applicable method, but the target for multiplying cybersecurity from the hardware to application level.**

This leads to make the set of available and useful PRNG larger and better by their features, like velocity, performance, simplicity in realization. These features match PRNG, based on cellular automata (CA), but not all rules, used in CA are appropriate for their transition functions. Bent functions are perfectly complement statistical weakness of some rules because of their non-linearity without loss of other features.

I. Introduction

MODERN society is vulnerable enough for hacking attacks, which are very large-spectered, from hardware attacks (on microchips) to web hacking (or application level hacking). Since, the first thing for computer science society and cybersecurity itself, is to protect personal data of users, which is the main aim for hackers. When we say "protect personal data", it comes cryptography methods in the first sight, like ciphers or lightweight cryptography, or coding theory. The basement of majority of cryptographical methods is generating random numbers, which can be both physically processed and mathematically. This paper is about second one, to be more exact, about pseudorandom number generator, based on homogenous structures using bent functions.

A. Homogenous structures overview

Lets take the set of k-dimensional vectors Z^k, set of 1 and $0 - E_n$, ordered set $V = (\alpha_1, \alpha_{(h-1)})$, where α_i is k dimensional vector from Z^k. Besides, lets determine a function $\phi = \phi(x_0, x_1, , x_{(h-1)})$, $\phi : (E_n)^h \rightarrow E_n$, additionally $(\phi(0, 0, ..., 0) = 0)$. As the result we'll get "four" $\sigma = (Z^k, E_n, V, \phi)$. That will be formal determinition of Homogenous structures (HS) [7], where E_n set of states of one cell in ϕ - local transition function.

Generally, HS are usually represented by ordered set of many Moore automata [13], which have states of other automata as input. To understand which of automata can influence on another one special scheme or neighborhood template can be used. There are several schemes, that are usually used in HS, like Moore's scheme (2 dimensional scheme, where target cell, surrounded by square of cells 3x3 if radius $r = 1$, or 5x5 if radius $r = 2$) or Neuman's one (2 dimensional scheme, where non diagonal cells surrounding the target one)

HS can be determined as dynamical discrete systems, where time and space are discrete. Changing of states of cells is conditioned by function ϕ, which is also included in rule. The rule is like a finite automaton with input, transition function and output.

Without loss of formalization and main idea, here and further term HS will be replaced by Cellular automata (CA), as this term is more widespread. Sharing CAs by complexity feature, it can be seen that there are two types of them: uniform CA (when all the grid have one rule and only one neighborhood template) or non-uniform CA (several neighborhood templates or(and) several rules). From this, perfomance features of both CA types are almost the same, besides memory (non-uniform needs more memory space for keeping rules and neighborhood templates). To save boards of the grid of CA, we will consider, that 2-dimensional grid is about toroidal form (without distortion in sizes).

Time spacing is a method of producing sequences with periodically interruptions in reading bits for one or several evolutions (iterations). That means, that only several iterations will influence on resulting bit sequence.

Site spacing is a method of producing sequences with periodically skipping of some bits in the grid in every iteration.

Widely spread so-called Wolfram's notation [17] for the rules. Firstly, lets call configuration the set of ones and zeros (two-state CA, where the sequence is considered as random number) at particular discrete moment of time. Further rule numbers are also suggested by Wolfram. Wolfram's rule 30 as an example:

How Wolfram's rules can be encoded goes further. For

example, $f(111) = 0, f(110) = 0, f(101) = 0, f(100) = 1, f(011) = 1, f(010) = 1, f(001) = 1, f(000) = 0$, is denoted rule 30.

B. Pseudorandom number generators (PRNG) overview

Random numbers are needed in different applications, i.e. in the cryptography area and coding theory. A number of algorithms need repeatable random numbers, based on deterministic algorithms, it is more correct to call such numbers pseudorandom, since they differ from the true random sequences obtained as a result of natural physical process.

Random number generators should have a number of properties if they are to be successfully applied in long stochastic models, like those used in computational physics. The most important properties from this point of view are good results in standard statistical tests for randomness, computational efficiency, a long period (the minimum number between repetitions) and the reproducibility of the sequence, e.g. NIST Test Suite. Considering several examples of PRNG, let's pay attention to Linear Congruential method and LFSR. They are quite simple, but perfectly describe the main idea of pseudorandom numbers.

C. Bent function

Bent functions are boolean functions with an extreme value nonlinearity. The measure of nonlinearity is an important characteristic boolean functions in cryptography. Linearity and properties close to it testify to the simple structure of this function and, as a rule, represent a large source of information about many other of its properties.

The nonlinearity of a function f is the distance from f to a class of affine functions. We denote the nonlinearity of the function f in terms of N_f :

$$N_f = d(f, A(n)) = \min_{g \in A(n)} d(f, g)$$

where $A(n)$ is class of affine functions.

The formula for calculating N_f by the Walsh-Hadamard transform:

$$d(\langle a, x\rangle, f) = \hat{f}(a) = 2^{n-1} - \frac{1}{2}\max_{a \in Z_2^n} |\hat{f}(a)|$$

Let $f \in P_2(n)$, write for it the Parseval equality:

$$\Sigma_{a \in Z_2^n} \hat{f}^2(a) \geq 2^n$$

We have 2^n non-negative summands whose sum is 2^{2n}. Consequently $\max\limits_{a \in Z_2^n} \hat{f}(a) \geq 2^n$, from which it follows that $\max\limits_{a \in Z_2^n} |\hat{f}(a)| \geq 2^{\frac{n}{2}}$. Therefore

$$N_f = 2^{n-1} - \frac{1}{2}\max_{a \in Z_2^n} |\hat{f}(a)| \leq 2^{n-1} - 2^{\frac{n}{2}-1}$$

The function $f \in P_2(n)$ is called maximally nonlinear if $N_f = 2^{n-1} - 2^{\frac{n}{2}-1}$

Definition: A bent function is a Boolean function with an even number of variables for which the Hamming distance from the set of affine Boolean functions with the same number of variables is maximal.

The properties of bent functions:

1) bent functions exist only for even n;
2) bent functions depend statistically on all their arguments;
3) let f be a bent function, and h belong to the class of linear functions. Then $f \oplus h$ belongs to the class of bent functions;
4) Let $(f \in P_2(n), g \in P_2(m)$ - be functions of disjoint sets of variables. Then $f \oplus h$ is a bent function if and only if f and g are bent functions.

We give examples of bent functions of a different number of variables.

for $n = 4$:

$$f(x_0, x_1, x_2, x_3) = x_0x_1 + x_2x_3$$
$$f(x_0, x_1, x_2, x_3) = x_0x_1 + x_2x_3 + x_0 + x_1$$

for $n = 6$:

$$f(x_0, x_1, x_2, x_3, x_4, x_5) = x_0x_1 + x_2x_3 + x_4x_5$$
$$f(x_0, x_1, x_2, x_3, x_4, x_5) = x_0x_1x_2 + x_1x_3x_4 + x_0x_1 + x_0x_3 + x_1x_5 + x_2x_4 + x_3x_4$$

for $n = 8$:

$$f(x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7) = x_0x_1x_2 + x_1x_3x_4 + x_2x_3x_5 + x_0x_3x_6 + x_2x_4 + x_1x_6 + x_0x_4 + x_0x_5 + x_3x_7$$

From the point of view of cryptography, the important criteria that a Boolean function f of n variables must satisfy are the following :

- equilibrium - the function f takes values 0 and 1 equally often;
- the propagation criterion $PC(k)$ of order k - for any nonzero vector $y \in Z_2^n$ weight at most k, the function $f(x + y) + f(x)$ is balanced;
- the maximum nonlinearity - the function f is such that the value of its nonlinearity N_f is maximal;

II. Algorithm of generating pseudorandom numbers on CA using Bent functions

The main idea of algorithm is in using CA Rules and bent functions to generate pseudorandom sequences of bits by usage of simple XOR operation to improve statistical features of CA sequences. Unique features of bent functions, like non-linearity and simplicity allows to generate quite random sequences in the grid of CA. Also, in terms of hardware or software implementation bent functions are simple enough to realize. As a result, we will get deterministic bent function output with n inputs of CA cells with Wolfram's notation rules.

A. Grid

From the point of view of geometry grid in our algorithm is represented by 2-dimensional parallelogram with sizes p and q, divided by equal cells, which contains only one of two possible states 0 or 1. Actually, the grid can be chosen with random sizes, but it is strongly recommended to create the grid, where p and q are prime numbers. It can improve periodical feature of output sequence. Let's divide this grid into two blocks, where

one block is for CA rule and another one is for bent function sequence. Every block b_i is 2 dimensional parallelogram with sizes l_{b_i} and w_{b_i}. Obviously, each block consists of $l_{b_i} * w_{b_i}$ cells. But result output sequence will be only from block with CA rule.

The matter is the fact, that the grid should be filled with initial states to produce next evolutions. For this reason there are a number of ways to do it. As we need deterministic algorithm and minimum of memory usage, it must be simple way to fill the grid. The method, describing below is called NESW (NESW- North, East, South, West). Formally, we take a result of size multiplication $l_{b_i} * w_{b_i}$ of each block b_i and start filling it cyclically from the South-West corner of block with the bits of received number, moving to the North till the edge of block or almost filled cell, then moving to the East, South and West, repeating conditions. This method also reminds spiral moving to the center of the grid of the block. The whole process is demonstrated on the Fig. 1

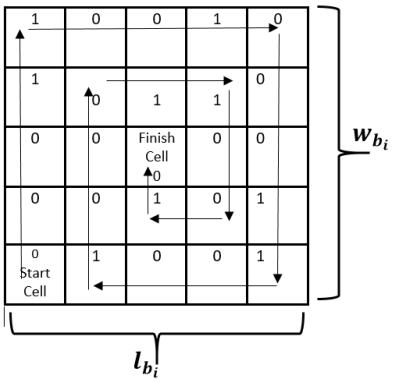


Fig. 1: b_i block grid view with NESW scheme of filling

B. Bent functions

As it mentioned before, traditional pseudorandom generators on CA use rules by Wolfram's notation, e.g. $Rule30$ or $Rule90$ and $Rule150$ together for better results. Another option is using some Boolean functions for changing states of cells in the grid, but these functions were linear. In this paper we research possiblity of using non-linear functions (bent functions as denoted before). But bent functions exist only for even number of arguments. And standard templates of neighborhood don't match for this condition. Fortunately, we can choose neighborhood template for the CA. Actually, we can choose as many templates as a number of blocks is and use different bent functions with only one restriction - possible lack of memory. Other sides of this method is flexible enough for realization.

So, we have m blocks in the grid and can choose $t_n \leq m$ templates of neighborhood and $t_f = t_n$ number of bent functions. We define target cell as a cell, which next state would be defined by bent function output.

Defining conditions for templates and bent functions:

1) number of cells in each $i-$th template. including target cell, must be even;
2) obviously number of cells is equal to number of arguments of appropriate bent function in the same block;
3) each b_i block has abstract toroidal form without size distortions;
4) target cell could be chosen anywhere in neighborhood template.

Now, using bent function $f(x_0, x_1, x_2, x_3) = x_0x_1 + x_2x_3 + x_0 + x_1$ lets see how our target cell will change on the next iteration (other cells should be changed also, but we want to check the output of function).

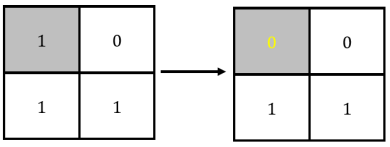


Fig. 2: How bent function works

C. Result

As we defined before the first block is our clasical CA with its own rule, but not all the rules produce statistically strong sequences. To improbe this scenario, we propose the second block in the grid, which would be producing by bent function. After needed number of evolutions block $b_1 XORing$ with block b_2 bit by bit, with appropriate size of sequence. So far, it can be concluded, that resulting sequence in block b_1 would be much more statistically strong. Thus, we can increase the set of rules, which produces strong, random and flexible sequences of bits.

D. Repeating

As our algorithm is deterministic, that leads to repeating evolution of CA with the same input and initial states. For this aim it can be generated abstract key to pass it through the channel for checking identity. So, the key K may be interpreted like the following sequence of bits. We use symbol | to mark concatenation.

$$K = p|q|l_{b_1}|w_{b_1}|...|l_{b_m}|w_{b_m}|t_{v_1^1}|t_{v_1^2}|...|t_{v_m^1}|t_{v_m^2}|V_1|...|V_m|T$$

where $t_{v_i^j}$ is the size of template parallelogram, including all possible cells in the template. And V_i are sent like the sequence of bits, where each bit, if it is 1 than cell is in the template, and when it is 0, than out of template.

Here are numbers of cells, instead of which should placed 1 or 0. T is bit sequence for determining all of bent functions, using in algorithm and other needed meta information. Thus, in spite of the fact that we must send such a long sized key, we can put some data in the boundary cells, placed around our main grid in order to save correct transition of state in case of software realization.

III. Test results

Our algorithm was tested with the help of NIST Test Suite, which developed for testing RNG and PRNG. The process of test involved: generation sequences of bits by our algorithm (1000000 bits), testing .txt file with sequence with NIST Test Suite. We tested various numbers of bent functions up to 16 arguments with different neighborhood templates appropriately, NESW method of filling the grid with initial states. Results of tests are averaged for all the experiments.

Head parameter of NIST Test Suite, showing the quality of sequence is $P-value$, which shows the difference between testing sequence and random sequence. The test statistic is used to calculate a $P-value$ that sumarizes the strength of the evidence against the null hypothesis. For these tests, each $P-value$ is the probability that a perfect random number generator would have produced a sequence less random than the sequence that was tested, given the kind of non-randomness assessed by the test. If a $P-value$ for a test is determined to be equal to 1, then the sequence appears to have perfect randomness. A $P-value$ of zero indicates that the sequence appears to be completely non-random. A significance level (α) can be chosen for the tests. If $P-value \geq \alpha$, then the null hypothesis is accepted; e.g., the sequence appears to be random. If $P-value < \alpha$, then the null hypothesis is rejected; e.g., the sequence appears to be non-random. The parameter α denotes the probability of the $TypeI$ error (if the data is, in truth, random, then a conclusion to reject the null hypothesis (e.g., conclude that the data is non-random) will occur a small percentage of the time) [10]. α is equal to 0.01.[10]

Here is the table (Fig. 3), which shows the difference on the NIST Test Suite between "clear" usage of $Rule30$ - $rule30$ column, with bent function of 4 arguments - $bent4$ column and bent function of 6 arguments - $bent6$ column. All bent functions were balanced: $f(x) + f(x + y)$ form, where x denotes vector of arguments, and y - random vector.

For $bent4$ variant we used the following function:

$$f(x_0, x_1, x_2, x_3) = x_0x_1 + x_1x_2 + x_2x_3$$

and as random vector was $y = 1011$

For $bent6$ variant we used the following function:

$$f(x_0, x_1, x_2, x_3, x_4, x_5) = x_0x_1 + x_2x_3 + x_4x_5$$

and as random vector was $y = 101110$

There will be represented results of average tests on about 100 rules without bent functions and with them of 4 and 6 arguments of Proportion criteria (Fig. 4) and P-value criteria (Fig. 5)

	bent4	***bent6***	***rule30***
Frequency	0.534146	0.979817	0.000002
BlockFrequency	0.911413	0.639187	0.000005
CumulativeSums	0.739918	0.977333	0.000003
Runs	0.534146	0.121244	0.000012
LongestRun	0.466882	0.667178	0.000000
Rank	0.304301	0.804337	0.350485
FFT	0.911413	0.542259	0.000000
NonOverlappingTemplate	0.739918	0.739163	0.004301
OverlappingTemplate	0.534146	0.237393	0.000000
Universal	0.276935	0.469075	0.000000
ApproximateEntropy	0.583423	0.955517	0.000000
RandomExcursions	0.655397	0.535641	0.000013
RandomExcursionsVariant	0.499590	0.424426	0.000023
Serial	0.534146	0.444016	0.000132
LinearComplexity	0.635174	0.736215	0.350485

Fig. 3: Statistical results of P-value for 3 variants of $Rule30$

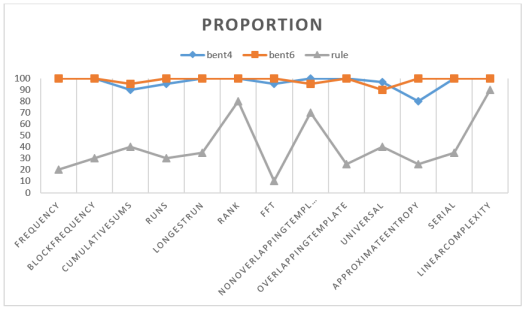


Fig. 4: Graph of difference between average test results (Proportion)

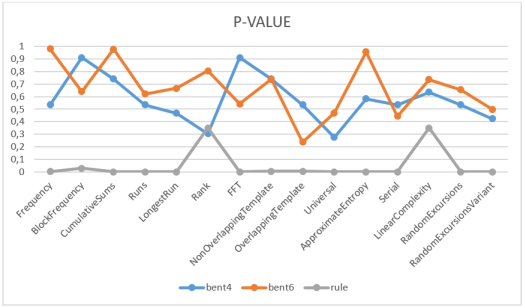


Fig. 5: Graph of difference between average test results (P-value)

IV. Conclusion

In spite of the fact, that not all the rules in classical CA can't be strongly recommended for generating pseudorandom sequences and numbers, we can find a way to increase statistical features up to hundred times with only using bent functions without loss of velocity and simplicity. Thus, the main advantage of this idea is that the set of using rules enlarges and can be used in a different way with results near needed distribution.

V. References

Acknowledgment

The research is supported by the grant of the Russian Science Foundation (Project 17-71 -20077)

REFERENCES

[1] Stephen Wolfram "Random sequence generation by cellular automata", *Advances in Applied Mathematics*, 1986

[2] Andrew Ilachinski "Cellular Automata: A Discrete Universe", *World Scientific*, 2001.

[3] M. Tomassini, M. Sipper, and M. Perrenoud "On the generation of high-quality random numbers by two-dimensional cellular automata" *IEEE Transactions on Computers*, 49(10):1146 - 1151, Oct 2000

[4] M. Tomassini, M. Perrenoud "Cryptography with cellular automata", *Appl. Soft Comput.*, 1(2):151 - 160, 2001

[5] Mads Haahr "True random number service" www.random.org, 1998

[6] M. Tomassini, M. Sipper, M. Zolla, M. Perrenoud, "Generating high-quality random numbers in parallel by cellular automata", *Future Gener.Comput. Syst., 16(2 - 3):291 - 305*, December 1999.

[7] V. B.Kudryavtsev, A.S.Podkolzin, "Cellular automata", *Intellectual systems* 1 - 4(10):657 - 692, 2006

[8] Ferguson, Niels; Schneier, Bruce; Kohno, Tadayoshi, "Chapter 9: Generating Randomness", *Cryptography Engineering: Design Principles and Practical Applications.*, Wiley Publishing, Inc. 2010

[9] Richard Brent "Uniform random number generators for supercomputers", 1992

[10] A. L. Rukhin, "A statistical test suite for random and pseudorandom number generators for cryptographic applications", *U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology*, rev. edition, 2010

[11] M. Tomassini, "Spatially Structured Evolutionary Algorithms: Artificial Evolution in Space and Time", *Springer*, 2005

[12] Toru Sasaki, Hiroyuki Togo, Jun Tanidaa and Yoshiki Ichiokab "Stream cipher based on pseudo-random number generation using optical affine transformation", *Applied Optics*, 39(14):2340 - 6 Âů June 2000

[13] Moore, Edward F "Gedanken-experiments on Sequential Machines", *Automata Studies, Annals of Math. Studies. Princeton, N.J.: Princeton University Press*, (34): 129 - 153, 1956

[14] Weisstein, Eric W. "von Neumann Neighborhood", *MathWorld–A Wolfram Web Resource*, http://mathworld.wolfram.com/vonNeumannNeighborhood.html

[15] Weisstein, Eric W. "Moore Neighborhood" *MathWorld–A Wolfram Web Resource*, http://mathworld.wolfram.com/MooreNeighborhood.html

[16] Alberto Dennunzio, Enrico Formenti, Julien Provillard Non-uniform cellular automata:Classes, dynamics, and decidability *Journal of Information and Computation*, Elsevier, 2012

[17] Wolfram S. "Cellular Automat" *Los Alamos Science* 9: 2 - 21, 1983

[18] Dobbertin H., Leander G. , "A survey of some recent results on bent functions" *Sequences and their applications.* , SETA, 2004.

[19] N. N. Tokareva, "Bent functions and their generalizations", *Prikl. Diskr. Mat.*, 2009, supplement 2, 5 - 17

[20] Carlet C., "On the higher order nonlinearities of Boolean functions and S-boxes, and their generalizations " *The Fifth Int. Conf. on Sequences and Their Applications* , SETA, 2008 P. 345 - 367 (Lecture Notes in *Comput. Sci.* V. 5203).

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 387–390
DOI: 10.15439/2018F254
ISSN 2300-5963 ACSIS, Vol. 15

Parametric Hash Function Resistant to Attack by Quantum Computer

Sergey Krendelev
Novosibirsk State University,
JetBrains research, Novosibirsk,
Russia
Email: s.f.krendelev@gmail.com

Polina Sazonova
Novosibirsk State University,
JetBrains research, Novosibirsk,
Russia
Email: psazonova@gmail.com

Abstract—**This paper describes an algorithm for creating hash function, resistant for quantum computer. The given approach is based on the problem of solving a system of polynomial equations in integers, where the number of equations is less than the number of unknown parameters. The developed algorithm is parameterized so the result of the hash function depends on several parameters, therefore, it will take considerably longer to select the solution of the task. The avalanche effect is about 50%, collision is impossible because the task to find a solution of the described system of equations with a degree greater than 3 is algorithmically unsolvable. This hash function was developed for blockchain to ensure its integrity, but it can also be used in any application where a hash function is needed.**

I. Introduction

SINCE Peter Shor has been demonstrated the solvability of the problem of discrete logarithm factorization using quantum computer in 1995 [1], there was become actually a post-quantum cryptography. It was necessary to develop such algorithms that could not be solved with the help of quantum computers.

Blockchain technology become popular for different kinds of applications: in banking, gambling, registries and etc. It uses hash function – cryptographically primitive for supporting invariability and consistency of data.

Hashing in blockchain is the process of converting an array of input data of arbitrary length into an output bit string. Hash function uses for making a digest of blocks or some another data, stored not only in blockchain. Hash functions guarantee the "irreversibility" of data.

But inventing quantum computers will force to develop a hash function resistant to the quantum computers.

In developing the hash function algorithm for the blockchain technology, some requirements is important: hash function should be resistance to collisions of first and second kind and it should have a high avalanche effect.

A. Motivation

At present, post-quantum cryptography is based on four approaches that guarantee resistance to quantum computers These are Code-based cryptography, Hash-based Digital Signature Schemes, Multivariate Public Key Cryptography, Lattice-based Cryptography [2].

Our algorithm is based on problem where the number of equations is less than the number of unknown parameters.

B. Algorithm Idea

As already mentioned, post-quantum cryptography is based on algorithmically unsolvable problems. We describe two complexity problems (we call it A and B) that are suitable for us. Our approach is constructed on Problem B, Problem A is its particular case. The work of Aitai [3] is equivalent to Problem A. In this section, we will show the transition from problem A to problem B and justify using of these computational problems.

Problem A. It is needed to find the solution of a system of linear Diophantine equations in integers.

Strongly underdefinished system of equations or a system where the number of equations is substantially less than the number of unknowns is given:

$$\sum_{j=1}^{n} a_{ij}x_j = d_i \; a_{ij}, d_i \in \mathrm{Z},$$

$$i = 1,2,\dots,m, j = 1,2,\dots,n, n > m$$

If there are restrictions, such as $x_j \geq 0,\ j = 1,2,\dots,n$ or $x_j \in \{0,1\}, j = 1,2,\dots,n$ - this task becomes the task of integer programming. Particularly interesting for encryption is the case where $n > m$ (it is a strongly underdefinished system of linear equations). In particular, if $m = 1$ and $x_j \in \{0,1\}, j = 1,2,\dots,n$, then this task is the task of the knapsack problem or subset-sum problem.

The scheme of the hash function, described by M. Aitai in 1996, is a special case of problem A. In the original article it tells about the lattice theory, but we show that the problem on lattices is equivalent to the described problem A.

Let us describe the scheme of the hash function of M. Aitai.

A randomly selected matrix $\mathrm{A} \in \mathrm{Z}_p^{n\times m}$ of dimension $n \times m$ is chosen, where $n < m$. Vector $\mathrm{x} \in \mathrm{Z}_p^m (d < p)$ will be hashed.

For this the system $\mathrm{A}x = mod(p) \in \mathrm{Z}_p^n$ is calculated, where $\mathrm{A}x$ is the hash of the vector x.

Note, that the parameters are set: n, m, q, $d > 1$, $n < m$, $q > d$, $\mathrm{A} \in \mathrm{Z}_p^{n\times m}$.

We note that the solution of equation $\mathrm{A}x = mod(p) \in \mathrm{Z}_p^n$ is a problem A, which is guarantees a solution. Consequently, the solution of the system of linear equation where the number of equations is less than the number of unknowns is equivalent to the problem on lattices.

This work was not supported by any organization

Problem B. It is necessary to find a solution of a system of polynomial equations in integers.

$$f_i(x_1, x_2, \dots, x_n) = 0, i = 1,2, \dots, m$$

Problem B is algorithmically unsolvable. In addition, if the degrees of polynomials ≥ 3 and $n > m$, then the problem is algorithmically unsolvable in integers. This conclusion follows from solution of 10th Hilbert problem.

In this paper, we consider a variant of constructing a hash function based on the problem B. In this type of hash function, a set of parameters can be used to enhances the persistence of the hash function. If you build a set of hash functions that depends on a large number of parameters, you get an object of the Universal hash type [4].

II. Algorithm Description

As our algorithm is parametric, first we need to choose parameters. In based version the parameters is: module *p*, size of dimension $m \times n$, set of starting coefficients $\propto_1, \propto_2, \dots, \propto_n$, size of block *b*, rules of forming summands $h_1(x), h_2(x), \dots, h_m(x)$.

Let us consider algorithms parameters. We also have developed requirements for parameters for the better result.

All calculations will be performed on the module *p*. The module should be a sufficiently large prime number.

We will generate some set of vectors according to special rules derived from the parameters $\propto_1, \propto_2, \dots, \propto_n$, $\propto_i \in Z_p^n, i = 1,2, \dots, n$, where n - is an arbitrary integer. The dimension of these vectors is *n*.

Suppose that some hashed document is described by a set of numbers $x = (x_1, x_2, \dots)$. Each number is a certain number of bits, assembled into a conditional block. Our block can be 8, 10, 12, etc. bit. The size of the block in bits *b* is another parameter of our algorithm.

A rule of generation of functions $h_1(x), h_2(x), \dots, h_m(x)$ should be defined as a parameter. It will determine the order of formation of the terms of our hash function.

The computation procedures of the proposed algorithm are illustrated as following.

Step 1: data preparation.

On this step, we prepare a string of decimal integers $x = (x_1, x_2, \dots, x_m)$ according to the input file.

Next, we prepare a matrix A = $(a_1, a_2, \dots, a_m)$, forming on the set of starting coefficients $\propto_1, \propto_2, \dots, \propto_n$. Vector a_i construct as a recurrent sequence according to the formula $a_i = \propto_1 a_1 + \propto_2 a_2 + \dots + \propto_n a_n$

Step 2: constructing a hash function.

Then the following vector will be a hash:

$$H(x) = [a_1 h_1(x) + a_2 h_2(x) + \dots + a_m h_m(x)] mod(p)$$

Functions $h_1(x), h_2(x), \dots, h_m(x)$ in hash function can be implemented as follows, but we can choose any rule for forming $h_i(x)$:

$$H(x) = [a_1 x_1 x_2 + a_2 x_2 x_3 + \dots + a_m x_m x_1] mod(p)$$

The size of the output string of the hash function is $n \times m$.

Step 3: modifications.

On the large file we have a high probability when some terms will be a zero. The main cause of it is a rule of forming a recurrent sequence, when zero in some terms is cumulated. To avoid it in a base version of algorithm we use a cyclic shift. In another version, we can use replacing on zero-component to fixed number which can be a parameter too.

Thus, we have constructed a hash-scheme with parameters, where the parameters are: module *p*, vector dimension *n*, block size *b*, terms generation rules $h_1(x), h_2(x), \dots, h_m(x)$.

III. The Toy Example

On the first step parameters is chosen. It is a simple number p, which will be a module; for example here p = 4049, the dimension of the vector n = 4, m = 4, the size of the block is b = 6 bits, the rule of generating multipliers in the term is $x_i x_{i+1}$, window size is 2, coefficients $\propto_1, \propto_2, \propto_3, \propto_4$ = (3174,3507,860,1294).

On the first step, we preparing a data.

Data from the file represented as decimal integers is x_1, x_2, ..., where each *x* is 6 bits. We separate 32 bits file on block of 6 bit and convert to decimal integers and the result is (34, 16, 23, 63).

Next, we need to generate coefficients A from starting coefficients $\propto_1, \propto_2, \dots, \propto_m =$ (3507,860,1294,3174)

$$\begin{pmatrix} \propto_1 \\ \propto_2 \\ \propto_3 \\ \propto_4 \end{pmatrix}, \propto_i \in \mathrm{Z}_{4049}, i = 1,2,3,4$$

Each a_k is calculate using recurrent sequence. On first step it will be $a_k = 3507a_{k-1} + 860a_{k-2} + 1294a_{k-3} + 3174a_{k-4}$.

Let a_i is:

$$a_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}; \; a_{-1} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}; \; a_{-2} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}; \; a_{-3} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Consequently:

$$a_1 = \begin{pmatrix} 3507 \\ 860 \\ 1294 \\ 3174 \end{pmatrix}$$

When we read the elements of a file by 2 items and calculate the product, it can turned to 0, if any one term turns to 0. Therefore, it is necessary to provide a decision of this problem in this case. We make a cyclic shift of numbers in vector a_i on 1 positions.

Next step we need to construct and calculate hash function:

$H(x) = [a_1 x_1 x_2 + a_2 x_2 x_3 + a_3 x_3 x_4 + a_4 x_4 x_1] mod(p)$

Let us construct the first term for the hash $a_1 x_1 x_2$:

$$\begin{pmatrix} 3507 \\ 860 \\ 1294 \\ 3174 \end{pmatrix} \times 34 \times 16 = \begin{pmatrix} 729 \\ 2205 \\ 3459 \\ 1782 \end{pmatrix} mod(4049)$$

Now the next vector according to the recurrence sequence should be calculated:

$$a_2 = 3174\begin{pmatrix}729\\2205\\3459\\1782\end{pmatrix} + 3507\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + 860\begin{pmatrix}0\\1\\0\\0\end{pmatrix} + 1294\begin{pmatrix}0\\0\\1\\0\end{pmatrix} = \begin{pmatrix}1325\\2858\\3321\\3664\end{pmatrix} mod(4049)$$

We calculate the product from the document data to the generated vectors:

$$\begin{pmatrix}1325\\2858\\3321\\3664\end{pmatrix} \times 16 \times 23 = \begin{pmatrix}1720\\3053\\3379\\35\end{pmatrix} mod(4049)$$

Modify the first two terms of hash:

$$\begin{pmatrix}729\\2205\\3459\\1782\end{pmatrix} + \begin{pmatrix}1720\\3053\\3379\\35\end{pmatrix} = \begin{pmatrix}2449\\1210\\2769\\1817\end{pmatrix} mod(4049)$$

And so on by induction. The final result of hash function will be (1679, 1137, 1883, 213).

IV. Possible modifications

The described algorithm can be modified as follows.

Modification 1. The nonlinear case of the formation of terms for the hash function should be considered.

We need to define a rule where the data from the input file will be combined and be distributed according to the hash function.

For example,

$$Ax = [a_1x_1x_2x_3 + a_2x_2x_3x_4 + \cdots + a_mx_mx_1x_2]mod(p)$$

The operations of multiplication are made on modulo p. Thus, we can consider other types of polynomials for the nonlinear case.

Modification 2. The nonlinear case should be considered when multiplications are made according to some multiplication table.

This table can be generated according to the hashed data.

V. Security Proof

C. Theoretical Foundation

For a hash function f will be cryptographically stable, it must satisfy the follow three basic requirements which most hash functions are based in cryptography:

1. Irreversibility or resistance to restoration of the prototype: for a given value of a hash function *y*, a data block *x* for which $f(x) = y$ must not be computed.
2. Resistance to collisions of the first kind or restoration of the second inverse images: for a given message *x* it must be computationally impossible to find another message *z* for which $f(x) = f(z)$.
3. Resistance to collisions of the second kind: it must be computationally impossible to select a pair of messages *x, z* having the same hash.

These requirements are not independent:

1. An invertible function is unstable to collisions of the first and second kind.
2. A function that is unstable to collisions of the first kind is not resistant to collisions of the second kind; the converse is not true.

Let us consider how the collision for our variant of the hash function will look. Let the same hash function be given for two different *x* and *z* documents:

$$H(x) = [a_1h_1(x) + a_2h_2(x) + \cdots + a_mh_m(x)]mod(p)$$

$$H(z) = [a_1h_1(z) + a_2h_2(z) + \cdots + a_mh_m(z)]mod(p)$$

Collision means that if $x \neq z$, but $H(x) = H(z)$.

Suppose for our algorithm the source document is known. Then, taking into account that the vectors $a_1, a_2, \dots, a_m$ are formed according to the parameters and the special rules to the function $h_i(x), i = 1,2, \dots, m$ calculations, the attacker is aware of the following information: vectors $a_1, a_2, \dots, a_m$, $\alpha_i = h_i(x), i = 1,2, \dots, m$ and $v = \sum_{j=1}^{m} \alpha_i a_j \, mod(p)$. The vector *v* is a hash of the document.

We need to solve equation $H(x) = d$ to find collisions, but this problem is equivalent to problem B, described in the section I.B of this article.

Thus, we demonstrated that a collision is theoretically possible. However, we affirm that there is no sense to find a collision for our algorithms, since the problem is algorithmically unsolvable if there are polynomials in the system of equations with a degree greater than 3.

For cryptographic hash functions it is also important that with the slightest change in the argument, the value of the function changes greatly (avalanche effect). In particular, the value of a hash should not give a leak of information, even about individual bits of the argument. This requirement is the key to the crypto-stability of algorithms for hashing user passwords to obtain keys.

D. Implementation Details

Describing algorithm was implemented in Python 3.3 for testing; measurements were made on a computer with an Intel Core i5-4210U of 2 cores, operating at 2.4Ghz. The PC contains 8 Gb RAM.

For testing avalanche effect, we calculated the hash function from the source file, changed an arbitrary bit in the source file and calculated the hash function from the modified file. Then a bitwise comparison was made. In the case of any documents of any size, when changing 1 bit in the source file, the hashes of the primary and modified files coincide only by 47-50% with a bitwise comparison.

Moreover, the best parameters at which the maximum number of discrepancies is reached is a sufficiently large prime number and the large dimension of the vector K is about 100.

The speed of the algorithm is about 0.007 sec for a 1 kb file, an average of 70 seconds for a 500 kb file, an average

500 seconds for a 1 mb file. The algorithm works both with text data, and with photo, video and audio content. Obviously, realization of this algorithm should be optimized to reduce the processing speed of the file.

VI. Conclusion

In this article is proposed an algorithm of hash function resistant to quantum computer. This algorithm uses algorithmically unsolvable problem of finding a solution to a system of polynomial equations in integers. Our algorithm is parametrized, which increases the decision-making time. It is resistant to collisions, because the problem on which the algorithm is built is algorithmically unsolvable (in the case where the degree of the polynomial is greater than 3). The avalanche effect is about 47-50% with a bitwise comparison. The algorithm can work both with text data, with photo, video and audio contents.

This algorithm was developed for blockchain technology to increase its resistance to attacks by quantum computer. It can also be used in any application where a hash function is needed.

VII. References

[1] Shor P.W. “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”, SIAM J. Com., 26:5, 1997, pp. 1484-1509.

[2] Bernstein D. J., Buchmann J., Dahmen E. “Post-Quantum Cryptography”. Springer-Verlag Berlin Heidelberg, 2009.

[3] M. Ajtai. “Generating Hard Instances of Lattice Problems”. In: 28th ACM Symposium on Theory of Computing, ACM, Philadelphia, USA, 1996, pp. 99–108.

[4] L. Carter and M. Wegman. “Universal Classes of Hash Functions”. In: J. Computer and System Sciences, Vol. 18(2), 1979, pp. 143–154.

[5] O. Goldreich, H. Krawczyk and M. Luby. “On the existence of pseudorandom generators”. In: SIAM J. on Computing, Vol. 22-6, 1993, pp. 1163–1175.

[6] J. Hastad, R. Impagliazzo, L.A. Levin and M. Luby. “A Pseudorandom Generator from any One-way Function”. In: SIAM J. on Computing, Vol. 28 (4), 1999, pp. 1364–1396.

[7] A.K. Lenstra, H.W. Lenstra, L. Lov´asz. “Factoring Polynomials with Rational Coefficients”. In: Mathematische Annalen, vol. 261(4), 1982, pages 515–534.

[8] C.P. Schnorr. “A more efficient algorithm for a lattice basis reduction”. In: Journal of Algorithms, Vol. 9, 1988, pages 47–62.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 391–395
DOI: 10.15439/2018F208
ISSN 2300-5963 ACSIS, Vol. 15

A new WAF-based architecture for protecting web applications against CSRF attacks in malicious environment

Michal Srokosz*, Damian Rusinek†, Bogdan Ksiezopolski†
*Polish-Japanese Academy of Information Technology
ul. Koszykowa 86, 02-008 Warszawa, Poland
msrokosz@pjwstk.edu.pl
†Maria Curie-Sklodowska University
pl. Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland
{damian.rusinek, bogdan.ksiezopolski}@umcs.lublin.pl

***Abstract*—Web application firewall is an application firewall for HTTP applications. Typical WAF uses static analysis of HTTP request, defined as a set of rules, to find potentially dangerous payloads in the requests. Generally, these rules cover common attacks such as cross-site scripting (XSS) and SQL injection which are server-related attacks. Cross-site scripting is client-side attack however the server is attacked and forced to return malicious response. Rule-based approach becomes useless when the attack is client-related, for example employing malware on the banking site. Malware allows to change the transfer data. This scenario is hard to detect because the browser displays valid transfer data and data is changed to the thieves' accounts number at the communication stage.**

In this paper we introduce a new web-based architecture for protecting web applications against CSRF attacks in malicious environemnt. In our approach we extend a classic, static WAF approach with historical and behavioral analysis, based on actions performed by the user in the past.

I. Introduction

ONE OF the ideas to increase Web applications security was Web Application Firewall, a proxy server used to defend web apps against attacks usually employed in the application layer in contrary to classic firewalls. WAFs are located between classic firewall and the application server. Such architecture allows the firewall to mitigate attacks on lower layers and WAF to detect and mitigate attacks on application layer. Both groups of mentioned attacks can have similiar consequences such as Denial of Service. DoS attacks, well known from lower layers[7], can also be performed on the application layer[5]. In [9] authors analyze 63 other articles about HTTP-GET flood attacks.

However, the application layer introduces a wide range of new threats to be detected and removed by WAF. OWASP Top Ten [10] is a powerful awareness document for web application security which represents a broad consensus about the most critical web application security flaws. The OWASP Top Ten list includes flaws, such as injections, cross-site scripting, cross-site request forgery, insecure session management, insecure direct object references, security misconfiguration, using components with known vulnerabilities and others. Most of web application firewalls focus on the technical attacks such as injections, cross-site scripting or cross-site request forgery while it is hard to detect other types such as insecure direct object references or business logic flaws because they are strictly application-dependent.

We are going to use request forgery attacks as an example of successful business attacks to present our new approach to detect and mitigate malicious requests. The most popular type of request forgery attacks are cross-site request forgery attack (CSRF) which makes a logged-on victim's browser send a forged HTTP request, together with the victim's session cookie and any other automatically attached authentication information to a vulnerable web application. In other words, the attacker forces the victim's browser to generate requests, which from the vulnerable application's perspective are legitimate. For this reason, on the server side we are not able to detect this only based on the technical attributes of the query. Although this is not a sophisticated attack, it indicates that the key players (Facebook, LinkedIn, etc.) had suffered from it.

Another type of request forgery attack is server-side request forgery (SSRF) and request forgery generated by malicious software. The SSRF differs from CSRF that the attacker forces a vulnerable application server to send a request. In the second type the attacker installs malicious software of victim's device which later sniffs the authentication data (eg. SMSes on the smartphone) and sends authenticated requests. According to reports by Symantec [11] and Kaspersky Lab [4] the malicious software is a significant problem with more than 30% of user computers subjected to at least one Malware-class attack and more than 170 mobile applications for credentials stealing in 2016. The financial Trojan threat landscape is dominated by three malware families: Ramnit, Bebloh (Trojan.Bebloh), and Zeus (Trojan.Zbot), responsible for 86 percent of all financial Trojan attack activity in 2016. Most anti-malware solutions is based on the detection of their presence.

Many commercial WAFs use signature-based techniques, attempting to find the malicious inputs appearing in the signature database. One can enumerate known WAFs, such

as F5, Juniper, Modsecurity and many others. In the scope of request forgery attacks the defense technique is to tokenize the requests. Such solutions are not only used by WAFs but many application frameworks provide such middleware as well.

Unfortunatelly, the use of tokens as the factor, which authenticate the requests is not sufficient in the malicious environemnt. In this paper, we are focused on a popular case of request forgery attack performed by malicious software installed on clients device (eg. mobile phone) and propose a mechanism to detect such attacks. The current Web Application Firewalls assume that the clients' devices is free from malicious software. This assumption in times of common malware can not take place.

The major contributions of the presented results can be summarized as follows:

- We present a successful request forgery attack on the application defended by classic WAF when client has malware installed.
- We propose new architecture for protecting web applications against request forgery attacks performed by malicious software.
- We extend our WAF proposal to include Two-Factor Authorization mechanism and user's history analysis.

The content of this paper is structured as follows. We discuss the related work in Section II. In section III we introduce the notation and describe the successful request forgery attacks leading to authorization bypass. Section IV describes our approach to detect and mitigate malicious business actions such as requests performed by malware. It includes the description of architecture, the detection algorithm. Finally, section V concludes this paper and describes the futher work.

II. Related Work

In the literature method of protecting web applications are not widely discussed. Researchers focus on non-standard attacks that can not be detected on classic firewalls and design new mechanisms for detecting these specific attacks.

In [3] authors propose an automatic method of HTTP attacks signature generation. Their approach relies on the use of a service-specific, semantic-aware anomaly detection scheme that combines stochastic learning with a model structure based on the HTTP protocol specification. The proposed soluction assume that the clienct is free from malicious software.

The article [8] proposes an approach that uses ontology models to detect web application attacks in HTTP protocol. Authors created three models of correct request, correct response and an attack. The HTTP requests are analyzed for compliance with the model and marked as a potential attack when forbidden values are found. This approach is similar to the whitelist approach, which is time-consuming and leads to many false positive alarms.

The authors in article [6] concentrate on SQL injection attack and propose the detection mechanism employing graphs and Support Vector Machine. The algorithm converts SQL query to the graph and uses previously trained SVM to detect SQL injection. The drawback of this algorithm is that it focuses only on the detection of tautology which is the first phase of the attack. When the mechanism blocks such query it can be considered as the presence of vulnerability.

In [1] authors conducted a review of the literature on popular web application attacks from OWASP Top 10 list, such as injections, access control or session management. Among the analyzed mechanisms were source code static analysis, dynamic detection of forbidden values and more complex such as comparison of responses which dropped responses outlying from the norm. The most popular solutions were based on the detection of forbidden values and authors stated that there does not exist a solution that is capable of detecting all injections, even in only one category such as cross-site scripting, because of many special cases of such flaw. All the discussed protection mechanisms assume that the client is free from malicious software.

III. Authorization bypasss with request forgery attacks

The aim of authorization bypass attacks is to perform an unauthorized action on behalf of authroized user. There exist many attack vectors and scenarios. In this section we describe four examples of such attacks, beginning with the simplest one employing social engineering techniques, to more complicated which uses malicious software.

A. Notation

We are going to use the following notation to describe the attack flows. The steps described correspond to the numbers in square brackets on matching figures.

- Actors:
  - $Client$ - the mobile or web client of the system,
  - WAF - web application firewall,
  - $Server$ - system endpoint server (reverse proxy),
  - $Attacker$ - an attacker (eg. malware).
- Messages:
  - $CREDENTIALS_{Client}$ - Client's credentials,
  - $SESS_{Client}$ - Client's session created by Server,
  - $EMAIL_{MAL}$ - malicious e-mail message with malicious link,
  - REQ_{BA} - a request to obtain form for business action BA,
  - $RESP_{FORM:BA}$ - a response that contains the form of business action BA,
  - $DATA_{FORM:BA}$ - form data to performa a business action BA,
  - $MODDATA_{FORM:BA}$ - modified (by malware) form data to performa a business action BA with malicious result,
  - $RESP_{BA}$ - a response that confirms the execution of business action BA,
  - $RESP_{MODBA}$ - a response that confirms the execution of modified business action BA,
  - $CSRF_{TAG}$ - a anti-CSRF tag,
  - $2FA_{OTP}$ - one time password from 2FA device, mandratory to authorize business action,

- Actions:
  - $ClickLink(EMAIL_{MAL})$ - user clicks the link from malicious e-mail,
  - $Verify(CREDENTIALS_{Client})$ - Server verifies Client's credentials and creates new session for him,
  - $WafVerify(Data)$ - WAF verifies the correctness of Data,
  - $CreateUser(DATA_{FORM:USER})$ - Server creates user record with form data,
  - $FormTag(RESP_{FORM:BA})$ - WAF creates a form tag to prevent CSRF attack,
  - $DoBA(DATA_{FORM:BA})$ - Server executes business action BA using form data,
  - $Intercept(Data)$ - Attacker intercepts data,
  - $ModifyBA(DATA_{FORM:BA})$ - Attacker modifies business action attributes (form data),
  - $AdditionalAuthRequired(Data)$ - WAF checks whether additional authorization is required for given action described by data,
  - $Send2FARequestForData(MODDATA_{FORM})$ - WAF sends challange to 2FA device for given business data for additional authorization,
  - $AbortForSecurityReason()$ - Client aborts operation (hacking attempt found).

B. The CSRF attack using a malware to bypass RSA Token and WAF

In this example we present an attack which bypasses the RSA Token and Web Application Firewall with the use of financial Trojan like ZEUS. RSA Token is a two-factor authentication device which generate a cryptographically-secure token to authorize the business action. ZEUS malware, on the other hand, allows to change the bank transfer data in online banking system. The attack is hard to detect by user because the browser displays valid transfer data and data is changed to the thieves' account number during the communication. Two-factor authorization, which does not user a device that displays the decription of operation to be authorized, is not effective for this type of attack.

The case background is the following. The Client has a bank account in the bank which requires that the transactions commissioned on the online service must be confirmed using one time password (OTP) generated by RSA Token. Client's device is infected with malware that is specialized in stealing money from the bank transaction system (eg. ZEUS).

The scenario of the attack is presented on figure 1.

(1) The Client logs in to the bank system.
(2) WAF performs static analysis if request is technically correct.
(3) Request was validated and pass to the banking system.
(4) The system validated the data entered.
(5) System created the session and returned it to the client.
(6) He wants to transfer money to his contractor.
(7) WAF verifies request.
(8) Pass it to the banking system.
(9) WAF receives form.
(10) WAF tags it with CSRF token.
(11) WAF sends it back to the user.
(12) The account number to which he wants to transfer the money is not added to any trusted transfer templates - the system will require authorization and one time password (OTP) code from the RSA Token.
(13) Malware detects an attempt to perform a transfer and, at the communication stage.
(14) Malware swaps the contractor account to the Attacker's account.
(15) A malicious request is sent to the system.
(16) WAF validates modified request.
(17) Pass it to the system.
(18) System performs transfer to Attacker's account.
(19) Malware, on the summary screen of the transfer, presents the Client with the account number of the contractor. The unaware Client gives the OTP code and authorizes the transfer to the Attacker's account.

IV. The new architecture for protecting web applications against request forgery attacks performed by malicious software

In this section we describe our approach to extend WAF security with behavioral analysis. The solution we want to propose increases the security and usability of the application that the WAF protects. It reduces the risk of a successful attack, even if your device is infected with malware.

We introduce behavioral analysis and user action history. The user request is analyzed by WAF before it reaches the target system. The WAF analyzes whether the user has performed similar actions in the past and whether they have been successfully commissioned. The similarity is calculated on the base of technical and business attributes describing actions. When the requested action is similar, the WAF does not require additional authorization. If not, the WAF asks for additional authorization to confirm the operation for the data entered. It would speed up the use of the target system and minimize the risk that the user confirms the operation with input altered by the Attacker.

A. History analyze and the similarity function

The added value of our solution to the classic Web Application Firewall is the Hisotry Analyzer module. With this module we are able to detect potencial abuse using cross-site request forgery. The key element of the proposed solution is the similarity function. We are going to use the following model to describe the proposed solution. In order to implement historical analysis, we need to introduce a concept of action. Actions, ie, business orders that a user has performed on a system that protects the WAF. The module checks to see whether similar operations have been performed by the user in the past. Similarity is calculated using an algorithm 1. If the action is similar to past operations, additional authorization is not required.

Definition 1. Technical attributes.

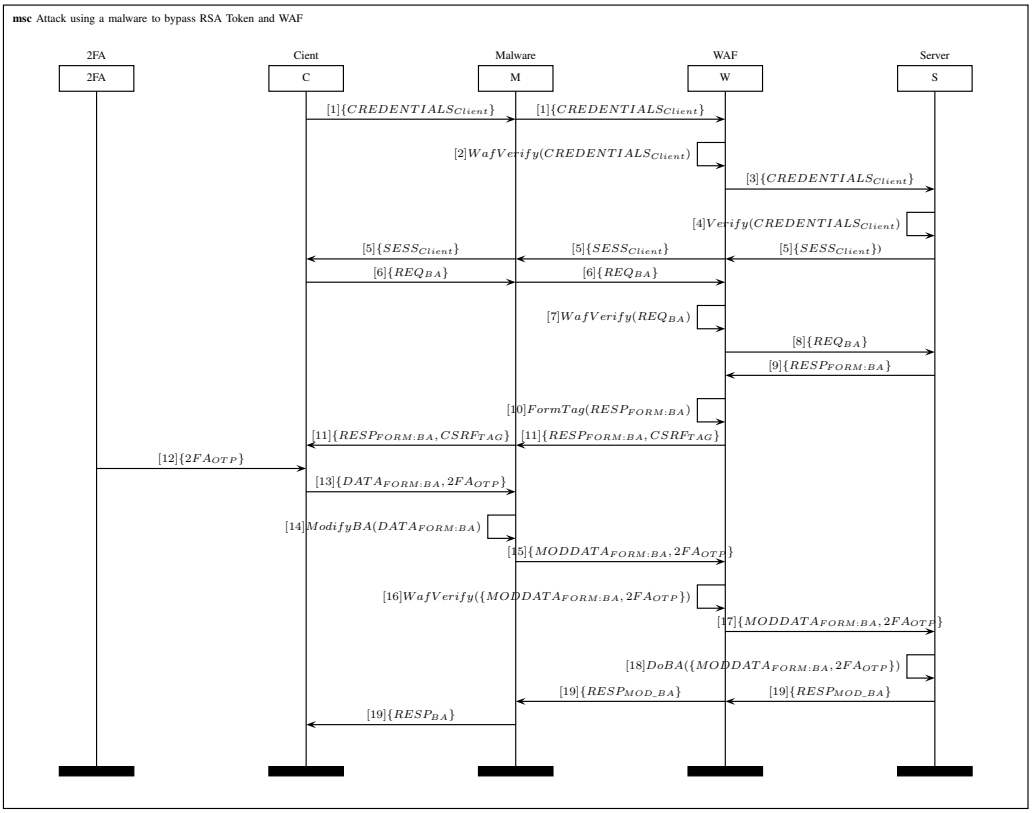


Fig. 1. Attack using a malware to bypass RSA Token and WAF.

$$T = \{t_1, t_2, ..., t_n\}$$
$$VT = \{vt_1, vt_2, ..., vt_n\}$$

The T set is a set of technical attributes and the VT set contains all of possible values of technical attributes. Technical attributes are not directly related to business data; they are rather a description of where and from what machine the action was initiated. This can be an ip address, browser fingerprint, country, time zone, and so on.

Definition 2. Business attributes.

$$B = \{b_1, b_2, ..., b_n\}$$
$$VB = \{vb_1, vb_2, ..., vb_n\}$$

The B set is a set of business attribute, the VB set represents all of possible values of business attributes. For example, the account number of the destination, the amount of the transfer for the service that executes the transfer.

Definition 3. Other variables.

$$Boolean = \{true, false\},$$
$$R,$$
$$Time$$

The $Boolean$ set is a set of boolean values, the R - set represents real numbers and $Time$ is a set of all possible timestamps

Definition 4. Actions and set of all possible actions.

$$A = \{a_1, a_2, ..., a_n\}$$
$$a_n = (T, B, VT, VB, Boolean, Boolean, Time)$$

The A set is a set of all possible actions and the a_n represents an action. This is a collection of all actions provided by a WAF-protected system with information about the actions taken by the user at a specific time along with the specific business effect. An action is defined by sets of technical and business attributes along with their values, two boolean values which states whether an action has been authorized with additional mechanisms and whether it has been allowed, and action's timestamp.

Definition 5. Function additionalAuthRequired

$$additionalAuthRequired(A \times 2^A) \rightarrow Boolean$$

Function $additionalAuthRequired$ returns whether additional authorization is required for given action.

Definition 6. Return elements.
The functions $attrs$, $values$, $time$ and $passed$ are return the elements of an action tuple a_i passed as an argument.

Definition 7. History.
The function $history$ is defined as follows: $history(a_i, A_m) \rightarrow A_n$, where $a_i \in A$, $A_m \subset A$ and

$A_n \subseteq A_m : a_j \in A_n \iff (passed(a_j) \wedge (time(a_j) <= time(a_i)))$

Definition 8. actionSimilarityThreshold function.

$$actionSimilarityThreshold(a_i) = max(similarityThreshold(vbt_i), \forall vbt_i \in 2^{VB \cup VT} : vbt_i \subseteq values(a_i))$$

Function $actionSimilarityThreshold$ returns similarity threshold for given action based on its attributes. It is calculated as the maximum similarityThreshold for all possible subsets of action's values of attributes

Definition 9. additionalAuthRequired function

$$additionalAuthRequired(a_i, A_m) = \begin{cases} true, & similarity(a_i, A_m) <= \\ & actionSimilarityThreshold(a_i) \\ false, & otherwise \end{cases}$$

The function $similarity$ depends on the method to be used to compare actions. The algorithm of calculating the similarity of action is presented in Alg. 1. In this state of study we are using simply algorithm based on weighted wage. The values of the similarity function are taken later in the verification ($additionalAuthRequired$) that the action is similar enough to those previously performed that no further verification is needed.

Algorithm 1: Function that returns the similarity between the action and the history

```
SIMILARITY (a_i, A_n)
    inputs: Action and Action set
    output: Similarity factor
    tmpSum ← 0
    tmpWage ← 0
    foreach a_j ∈ history(a_i, A_n) do
        foreach attr_i ∈ attrs(a_j) do
            if value(attr_i, a_i) ≈ value(attr_i, a_j) then
                tmpSum ← tmpSum + wage(attr_i)
            tmpWage ← tmpWage + wage(attr_i)
    if tmpWage = 0 then
        return 0;
    return tmpSum ÷ tmpWage;
```

V. Conclusions

In the article we presented the weaknesses of Web Application Firewalls which use signature-based and rule-based static analysis. We presented a successful request forgery attack on the application defended by classic WAF when client has malware installed.

To protect against such attacks we introduced an approach, based on historical and behavioral analysis of user requests, which reduces the need for additional forms of authorization. After sufficiently collecting and analyzing user's history, the additional authorization appears only in the situation that actually requires it. Such approach increases the responsiveness and general feel of the application.

In the future work, we plan to implement the proposed system and check the efficiency and accuracy of it.

References

[1] Deepa, G., Thilagam, P.S.: Securing web applications from injection and logic vulnerabilities: Approaches and challenges. Information and Software Technology 74, 160 – 180 (2016), http://www.sciencedirect.com/science/article/pii/S0950584916300234

[2] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Rfc 2616, hypertext transfer protocol – http/1.1 (1999), http://www.rfc.net/rfc2616.html

[3] Garcia-Teodoro, P., Diaz-Verdejo, J., Tapiador, J., Salazar-Hernandez, R.: Automatic generation of {HTTP} intrusion signatures by selective identification of anomalies. Computers & Security 55, 159 – 174 (2015), http://www.sciencedirect.com/science/article/pii/S0167404815001297

[4] Garnaeva, M., Sinitsyn, F., Namestnikov, Y., Makrushin, D., Liskin, A.: Overall statistics for 2016. Special report, Kaspersky Lab (December 2016), https://kasperskycontenthub.com/securelist/files/2016/12/Kaspersky_Security_Bulletin_2016_Statistics_ENG.pdf

[5] Jazi, H.H., Gonzalez, H., Stakhanova, N., A.Ghorbani, A.: Detecting http-based application layer dos attacks on web servers in the presence of sampling. Computer Networks 121, 25 – 36 (2017), http://www.sciencedirect.com/science/article/pii/S1389128617301172

[6] Kar, D., Panigrahi, S., Sundararajan, S.: Sqligot: Detecting {SQL} injection attacks using graph of tokens and {SVM}. Computers & Security 60, 206 – 225 (2016), http://www.sciencedirect.com/science/article/pii/S0167404816300451

[7] Mazur, K., Ksiezopolski, B., Nielek, R.: Multilevel modeling of distributed denial of service attacks in wireless sensor networks. Journal of Sensors 2016 (2016), https://www.hindawi.com/journals/js/2016/5017248/

[8] Razzaq, A., Anwar, Z., Ahmad, H.F., Latif, K., Munir, F.: Ontology for attack detection: An intelligent approach to web application security. Computers & Security 45, 124 – 146 (2014), http://www.sciencedirect.com/science/article/pii/S0167404814000868

[9] Singh, K., Singh, P., Kumar, K.: Application layer http-get flood {DDoS} attacks: Research landscape and challenges. Computers & Security 65, 344 – 372 (2017), http://www.sciencedirect.com/science/article/pii/S0167404816301365

[10] Wichers, D.: OWASP Top Ten Project. https://www.owasp.org/ (2013), [Online; accessed 12-March-2017]

[11] Wueest, C.: Istr financial threats review 2017. Special report, Symantec (May 2017), https://www.symantec.com/content/dam/symantec/docs/security-center/white-papers/istr-financial-threats-review-2017-en.pdf

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 397–405
DOI: 10.15439/2018F204
ISSN 2300-5963 ACSIS, Vol. 15

On the implementation of new symmetric ciphers based on non-bijective multivariate maps

Vasyl Ustymenko, Aneta Wróblewska
Maria Curie-Skłodowska University,
Institute of Mathematics,
pl. Marii Curie-Skłodowskiej 1,
20-031 Lublin, Poland
Email: vasylustimenko@yahoo.pl
awroblewska@hektor.umcs.lublin.pl

Urszula Romańczuk-Polubiec
Independent Researcher, Poland
Email: urszula_romanczuk@yahoo.pl

Monika Polak
Rochester Institute of Technology,
Department of Computer Science,
20 Lomb Memorial Dr,
Rochester, NY 14623, USA
Email: mkp@cs.rit.edu

Eustrat Zhupa
University of Rochester,
Department of Computer Science,
Rochester, NY 14627, USA
Email: ezhupa@cs.rochester.edu

***Abstract*—Certain families of graphs can be used to obtain multivariate polynomials for cryptographic algorithms. In particular, in this paper, we introduce stream ciphers based on non-bijective multivariate maps. The presented symmetric encryption algorithms are based on three families of bipartite graphs with partition sets isomorphic to $\mathbb{K}^n$, where $\mathbb{K}$ is selected as the finite commutative ring. The plainspace of the algorithm is $\Omega = \{x | \sum x_i \in \mathbb{K}^*, x \in \mathbb{K}^n\} \subset \mathbb{K}^n$, $\Omega \cong \mathbb{K}^* \times \mathbb{K}^{n-1}$. We describe the algorithm for the case $\mathbb{K} = \mathbb{Z}_{2^m}$, $m \geq 2$. In fact, we use the relation $d * d_{dec} \equiv 1 (\mathbf{mod}\ 2^{m-1})$, $d, d_{dec} \in \mathbb{Z}^*_{2^{m-1}}$ to obtain encryption polynomial map of degree greater than or equal to $d+2$ and decryption map of degree greater than or equal to $d_{dec}+2$. We assume d_{dec} grows with the growth of parameter m, because this makes cryptanalysis very difficult task. Symmetric encryption and decryption algorithms for users are numerical recurrent processes, not requiring generation of encryption and decryption maps in their symbolic forms. They use arithmetical operations of addition, subtraction, and multiplication. That's why the algorithms are robust (execution speed is $O(n)$). To break the algorithm an adversary must use linearization attacks for recovering non-bijective "decryption map" of degree greater than $d_{\mathrm{dec}}+2$ in its symbolic form. To achieve this, the adversary needs at least $O(n^{d_{\mathrm{dec}}}+2)$ pairs of plaintext and corresponding ciphertext to restore the non-bijective map of degree greater than or equal to $d_{dec}+2$. We present tables for evaluation of execution time for m = 8 with various length of passwords and sizes of files. Computer simulations demonstrate good mixing properties of the encryption functions.**

FEW graph based algorithms have been implemented since 1998 (see [1] - [25]). So there is some history of the usage of sparse algebraic graphs in symmetric cryptographical algorithms. The following known graphs defined over finite commutative ring $\mathbb{K}$ were used: $D(n, \mathbb{K})$ (see [1], for $\mathbb{K} = \mathbb{F}_q$ graphs were defined and investigated in [26], [27]), $W(n, \mathbb{K})$ (Wenger graphs defined in [28]), graphs $A(n, \mathbb{K})$ introduced in [45] and graphs $\widetilde{D(n, \mathbb{K})}$ of [25]. Popular choices of $\mathbb{K}$ are finite fields $\mathbb{F}_{127}$, $\mathbb{F}_{2^7}$, $\mathbb{F}_{2^8}$, $\mathbb{F}_{2^{16}}$ and $\mathbb{F}_2^{32}$ and rings modular arithmetics $\mathbb{Z}_{2^7}$, $\mathbb{Z}_{2^8}$, $\mathbb{Z}_{2^{16}}$. We present this research history in the next section.

In section 3 we introduce a class of bivariate graphs containg all the above mentioned graphs. Such concept is convenient for uniform description of encryption scheme and observation of common properties of graphs from this class (section 4). We compare graphs and related algorithms corresponding to different families ($W(n, \mathbb{K})$, $D(n, \mathbb{K})$, $A(n, \mathbb{K})$ and $\widetilde{D(n, \mathbb{K})}$) in section 5.

Here the reader can find remarks on multivariate cryptography and its connections with cryptographical applications of Algebraic Graph Theory.

RSA is one of the most popular cryptosystems. It is based on a number factorization problem and on Euler's Theorem. Peter Shor discovered that factorization problem can be effectively solved by using a theoretical quantum computer. It means that RSA could not be a security tool in the future postquantum era. One of the research directions leading to a postquantum secure public key is the Multivariate Cryptography which uses a polynomial maps of affine space $\mathbb{K}^n$ defined over a finite commutative ring $\mathbb{K}$ into itself as encryption tools (see [29]). This is a young promising research area because of the current lack of known cryptosystems with the proven resistance against attacks with the use of Turing machines. Other important direction of Postquantum Cryptography is the study of Hyperelliptic Curves Cryptosystems. We have to say that classical elliptic curves encryption will be not secure in the Postquantum era.

Applications of Algebraic Graph Theory to Multivariate Cryptography were shown in our talks at Erdös Centennial (2013, Budapest) and Central European Conference on Cryptology 2014 (Alfred Renyi Institute, Budapest) [30], [31]. Talks were devoted to algorithms based on bijective maps of affine spaces into itself. Applications of algebraic graphs to cryptography started with symmetric algorithms based on explicit constructions of extremal graph theory and their directed analogues (see survey [11], [32]). The main idea is to convert an algebraic graph in a finite automaton and to use the pseudorandom walks on the graph as encryption tools.

This approach can also be used for the key exchange protocols. Nowadays the idea of "symbolic walks" on algebraic graphs, when the walk on the graph depends on parameters given as special multivariate polynomials in variables depending from plainspace vector, appears in several public key cryptosystems. Another source of graphs suitable for cryptography is connected to finite geometries and their flag system (see [33] and further references).

Multivariate cryptography started from the study of potential for the special quadratic encryption multivariate bijective map of $\mathbb{K}^n$, where $\mathbb{K}$ is an extention of finite field $\mathbb{F}_q$ of characteristic 2. One of the first such cryptosystems was proposed by Imai and Matsumoto and cryptanalysis for that system was invented by J. Patarin. A survey on various modifications of this algorithm and corresponding cryptanalysis can be found in [29] or [34].

One of the first uses of non-bijective map of multivariate cryptography was in the *oil and vinegar* cryptosystem proposed in [35] and analyzed in [36]. Nowadays, this general idea is strongly supported by publication [37] devoted to security analysis of direct attacks on modified unbalanced oil and vinegar systems. It looks like such systems and rainbow signature schemes may lead to promising Public Key Schemes of Multivariate Encryption defined over finite fields. Non-bijective multivariate sparse encryption maps of degree 3 and ≥ 3 based on walks on algebraic graphs $D(n, \mathbb{K})$ defined over general commutative ring and their homomorphic images were proposed in [38]. Security of the corresponding cryptosystem rests on the idea of hidden discrete logarithm problem. U. Romańczuk-Polubiec and V. Ustimenko combine an idea of "oil and vinegar signature cryptosystem" with the idea of linguistic graph-based map with partially invertible decomposition to introduce a new cryptosystem [38]. This algorithm can be implemented with the use of families $D(n, \mathbb{K})$ and $A(n, \mathbb{K})$ and natural homomorphism between them. Finally, in [39] "hidden RSA multivariate encryption" based on graphs $D(n, \mathbb{K})$ were proposed.

In this paper we modify the encryption map (private key) of the above mentioned cryptosystem in terms of family of bivariate graphs defined over the commutative ring $\mathbb{K}$. These maps have multivariate nature despite the "numerical implementation" in symmetric ciphers mode with the plainspace isomorphic to $\mathbb{K}^* \times \mathbb{K}^{n-1}$.

I. On some implementation of algorithms based on bijective maps

We worked on a software package that allows us to investigate strongly symmetric cases of stream ciphers based on graphs $W(n, \mathbb{K})$, $D(n, \mathbb{K})$, $\widetilde{D(n, \mathbb{K})}$ and $A(n, \mathbb{K})$, where $\mathbb{K}$ is the arithmetic ring. Some cases are already implemented by our team at the level of prototype model.

In very special cases the algorithms were previously implemented. The first implementation of $D(n, \mathbb{K})$ encryption was done in 2000 at the University of South Pacific (USP, Fiji Islands). The research team was composed by Prof. V. Ustimenko, PhD Dharmendra Sharma (currently professor of University of Canberra), postgraduate students V. Gounder and R. Prasad (see [2], [3]). The work was supported by the University Research Committee of the University of South Pacific (USP) grant. The implementation of this case on asymmetric mode was discussed in [5]. The chosen case for $\mathbb{K}$ was $\mathbb{F}_{127}$, which is the closest prime number to the size of ASCII code alphabet. It means that one has to delete just the *delete* service symbol and can encrypt arbitrary files of type txt. The chosen string was $\alpha_i(x) = x + d_i$, where d_i are elements of chosen ring $\mathbb{K} = \mathbb{F}_{127}$ chosen in pseudorandom fashion. So that was a case of shifting encryption.

The affine transformations L_1 and L_2 were simply identities. Implemented cipher on ordinary PC was rather robust in performance, but with average mixing properties. It's been used at USP digital network working for campuses and USP centers located in 11 island countries of South Pacific region. The package was also used by ORACLE based system of the bursary office (see [8]). Recently group of students from Okanagan college (affiliated with the University of British Columbia) implemented that stream cipher on a cluster network of PC's. It was used for a large data encryption [10]. The implementation of that security algorithm for protection of Geo Information Systems was described in [6], [7].

Another case for $\mathbb{K} = \mathbb{Z}_{256}$ and graph $D(n, \mathbb{K})$ was implemented under the Research Committee of Sultan Qaboos University (SQU, Oman) grant. The research team was composed of professors Vasyl Ustimenko and Abderezak Tousane and students Rahma Al Habsi and Huda Al Naamani. The software uses one to one correspondence between element of $\mathbb{Z}_{256}$ and symbols of binary alphabet. It allows encryption of various file types (with extension doc, jpg, htm, avi, pdf, ...) in a way that encrypted file is presented in the same format with the plaintext. The symmetric algorithm was used at academical networks of SQU and Kiev Mohyla Academy [9], [10].

The cases of $D(n, \mathbb{K})$, where $\mathbb{K}$ is the finite field $\mathbb{F}_{2^7}$ $\mathbb{F}_{2^8}$, $\mathbb{F}_{2^{16}}$, the shifting encryption was implemented and investigated in [20].

The systematic study of shifting encryption for cases of shifting encryptions of $D(n, \mathbb{K})$ was conducted at UMCS (Lublin, Poland). J. Kotorowicz used arithmetical rings $\mathbb{Z}_2^7$, $\mathbb{Z}_2^8$, $\mathbb{Z}_2^{16}$ for the implementation with various affine transformation τ_L and τ_R (see [14], [16]). The encryption was essentially faster than in all previously known cases. The selected affine transformation leads to an encryption with very good mixing properties: the change of a single character of the plaintext or the change of a single character of the encryption string $d_1, d_2, \ldots, d_s$ causes the change of at least 98 percent of the ciphertext characters. In [23] these cases were implemented for graphs $A(n, \mathbb{K})$ with very similar results on the mixing properties. In the case of $\tau_R = {\tau_L}^{-1}$ it can be proved that the order of $A(n, \mathbb{K})$ and $D(n, \mathbb{K})$ based encryption map grows with the growth of parameter n. The comparison of orders was completed through the study of cycles structures of $A(n, \mathbb{K})$ and $D(n, \mathbb{K})$ encryptions. Results demonstrated similarity in both cases.

M. Klisowski implemented $D(n,\mathbb{K})$ and $A(n,\mathbb{K})$ shifting encryption on symbolic level in the cases of finite fields $\mathbb{F}_{2^7}$ $\mathbb{F}_{2^8}$, $\mathbb{F}_{2^{16}}$ $\mathbb{F}_{2^{32}}$ ([21], [22], [24]). In [40] A. Wróblewska proved that shifting $D(n,\mathbb{K})$ encryption is given by a cubical multivariate map. A similar result for $A(n,\mathbb{K})$ based encryption was stated in [41]. Simulation results of [22], [23] allow to estimate time of generation of these maps as functions of parameter n and densities of such multivariate cubic encryption and decryption maps. A comparison of cases $A(n,\mathbb{K})$ and $D(n,\mathbb{K})$ for the above fields an be found in [24]. Similar results for cases of Boolean rings of sizes 2^7, 2^8, 2^{16}, 2^{32} are obtained via computer simulations.

The PhD Thesis of M. Klisowski [42] contains the first results on $D(n,\mathbb{K})$ and $A(n,\mathbb{K})$ based multivariate maps which are not defined via shifting encryptions. He used symbolic strings of kind $\alpha_1(x) = x+c_1, \alpha_2(x) = x+c_2, \ldots, \alpha_{s-1}(x) = x + c_{s-1}$, $\alpha_s(x) = x^3 + c_s$ with constants c_i, $i = 1, 2, \ldots, s$ for special fields $\mathbb{F}_q$ in which $x^3 = b$ has unique solution. It was shown that such a choice makes direct linearization attacks impossible.

The first implementation for the case of Wenger graph based encryption was completed at the University of Sao Paolo (USP, Brasil) (see [12] and further references). Professors V. Futorny and V. Ustimenko chose field $\mathbb{F}_{253}$ which size is the closest from below prime to the size of binary alphabet. This research was partially supported by FAPESP foundation (grant for international cooperation with USP). Computer simulation demonstrated high speed of encryption. In [12] authors evaluated the diameter of graph $W(n,\mathbb{F}_q)$ and proved that the family of these graphs $W(n,q)$, $n \leq q$ is a family of small world graphs.

Professor Routo Terada (USP, Brasil) suggested to investigate the behaviour of these algorithms under linearization attacks. Computer simulation supports the conjecture on a good resistance of the encryption scheme to such attacks.

The idea of using graphs $A(n,\mathbb{K})$ in cryptography was proposed by U. Romańczuk-Polubiec and V. Ustimenko in [45]. Evaluation of the order of encryption map based on $A(n,q)$ was presented in [23]. A theoretical study of orders and cycles can be found in [44], [45].

Some stream ciphers defined via graphs $\widetilde{D(n,\mathbb{K})}$ were proposed by M. Polak and V. Ustymenko in [25]. Furthermore, M. Polak compared LDPC codes corresponding to $A(n,\mathbb{K})$, $D(n,\mathbb{K})$ and $\widetilde{D(n,\mathbb{K})}$ in [49].

The importance of such graphs was justified in [44]. The encryption algorithm was implemented and some properties (speed, mixing properties, order) were investigated in the paper.

II. On the class of bivariate graphs

Let $\mathbb{K}$ be a commutative ring. We define $T(n,\mathbb{K})$ as a bipartite graph with the set of vertices $V(T) = P \cup L$, $P \cap L = \emptyset$. We call $P = \mathbb{K}^n$ a set of points and $L = \mathbb{K}^n$ a set of lines (two copies of a Cartesian power of $\mathbb{K}$ are used). We will use two types of brackets to distinguish points $(p) \in P$ and lines $[l] \in L$:

$$(p) = (p_1, p_2, \ldots, p_n) \in P,$$

$$[l] = [l_1, l_2, \ldots, l_n] \in L.$$

p_i, l_i $(1 \leq i \leq n)$ are elements of $\mathbb{K}$. We say that vertex (p) (point (p)) is incident with the vertex $[l]$ (line $[l]$) and we write: $(p)I_T[l]$, if the following relations between their coordinates hold:

$$\begin{cases} p_2 - l_2 = e_2^1 p_1 l_1 \\ p_3 - l_3 = e_3^1 p_1 l_2 + e_3^2 l_1 p_2 \\ \vdots \\ p_s - l_s = e_s^1 p_1 l_{i_s} + e_s^2 l_1 p_{j_s} \\ \vdots \\ p_n - l_n = e_n^1 p_1 l_{i_n} + e_n^2 l_1 p_{j_n} \end{cases} \quad (1)$$

where $e_2^1, e_s^1, e_s^2 \in \{0, 1, -1\}$, $1 \leq i_s < s$, $1 \leq j_s < s$. So the incidence relations for graph $T = T(n,\mathbb{K})$ are given by condition $(p)I_T[l]$. The set of edges consists of all pairs $\{(p), [l]\}$ for which: $(p)I_T[l]$. Let us consider the case of finite commutative ring $\mathbb{K}$, $|\mathbb{K}| = k$. As it instantly follows from the definition, the order of our bipartite graph is $|V(T)| = 2k^n$ and the number of edges is $|E(T)| = k^n \cdot k = k^{n+1}$. Graphs $T = T(n,\mathbb{K})$ are k-regular. In fact, the neighbour of a given point (p) is given by above equations, where parameters $p_1, p_2, \ldots, p_n$ are fixed elements of the ring and symbols $l_1, l_2, \ldots, l_n$ are variables. It is easy to see that if we set l_1 then this choice uniformly establishes values $l_2, l_3 \ldots, l_n$. So each point has precisely k neighbours. In a similar way we observe that the neighbourhood of any line also contains k neighbours. Notice, that the order and degree of our graph defined via strings i_s, j_s, e_2^1, e_s^1, e_s^2, where $s = 2, 3, \ldots, n$, does not depend on the strings.

Let us consider some examples.

Wenger graphs $W(n,\mathbb{K})$

In 1991 Wenger defined the family of bipartite, p-regular graphs $H_n(p)$, where p prime number [28]. In [26] Lazebnik and Ustimenko introduced straight forward generalization $W(n,q)$ of these graphs via change of $\mathbb{F}_p$ to $\mathbb{F}_q$, where q is a prime power. They used special Lie algebra and proved that the family of bipartite, q-regular graphs $W(n,q)$, where q is prime power and $n \geq 2$. Graphs $W(n,q)$ are defined for all prime powers and $H_n(p) = W(n,p)$ are defined only for primes.

The set of vertices of infinite incidence structure (P, L, I) is $V = P \cup L$ and the set of edges E consists of all pairs $\{(p), [l]\}$ for which $(p)I[l]$. Bipartite graphs $W(n,q)$ have partition sets P_n (collection of points) and L_n (collection of lines) isomorphic to vector space $\mathbb{F}_q^n$, where $n \in \mathbb{N}_+$. Let us use the following notations for points and lines in graph $W(n,q)$:

$$(p) = (p_1, p_2, p_3, \ldots, p_n) \in P,$$

$$[l] = [l_{1,}, l_2, l_3, \ldots, l_n] \in L.$$

The point (p) is incident with the line $[l]$, and we write $(p)I_W[l]$, if the following relations between their coordinates hold:

$$\{\ l_i - p_i = p_1 l_{i-1}, \tag{2}$$

for $2 \leq i \leq n$. The graphs $W(n, \mathbb{F}_q)$ have cycles of length 8.

One can change finite field $\mathbb{K}$ for general commutative ring $\mathbb{K}$ and work with graph $W(n, \mathbb{K})$.

Graphs $A(n, \mathbb{K})$

Graphs $A(n, \mathbb{K})$, formally appearing as graphs $E(n, \mathbb{K})$ in [43], are used as tools for the study of $D(n, \mathbb{K})$ properties. Later on the graphs $E(n, \mathbb{K})$ were presented with another name as an independent family $A(n, q)$ for the first time in [45] for cryptographic applications.

Let us use the following notations for points and lines in the graph $A(n, \mathbb{K})$:

$$(p) = (p_1, p_2, p_3, \ldots, p_n) \in P,$$
$$[l] = [l_{1,}, l_2, l_3, \ldots, l_n] \in L.$$

The point (p) is incident with the line $[l]$, and we write $(p)I_A[l]$, if the following relations between their coordinates hold:

$$\begin{cases} l_2 - p_2 = l_1 p_1 \\ l_3 - p_3 = p_1 l_2 \\ l_4 - p_4 = l_1 p_3 \\ l_i - p_i = p_1 l_{i-1} \text{ for odd } i \\ l_i - p_i = l_1 p_{i-1} \text{ for even } i \end{cases} \tag{3}$$

for $3 \leq i \leq n$.

Graphs $D(n, \mathbb{K})$

The following interpretation of a family of graphs $D(n, \mathbb{K})$ in case $\mathbb{K} = \mathbb{F}_q$ can be found in [27]. By I_D we denote the incidence relation for this graph. Let us use the following notations for points and lines:

$$(p) = (p_1, p_2, p_3, \ldots, p_n) \in P,$$
$$[l] = [l_{1,}, l_2, l_3, \ldots, l_n] \in L.$$

Two types of brackets allow us to distinguish points from lines. Points and lines are elements of two copies of the vector space over $\mathbb{K}$. Point (p) is incident with the line $[l]$, and we write $(p)I_D[l]$, if the following relations between their coordinates hold:

$$\begin{cases} l_2 - p_2 = l_1 p_1 \\ l_3 - p_3 = p_1 l_2 \\ l_4 - p_4 = l_1 p_2 \\ l_i - p_i = p_1 l_{i-2} \text{ for } i \bmod 4 \equiv 2 \text{ or } i \bmod 4 \equiv 3 \\ l_i - p_i = l_1 p_{i-2} \text{ for } i \bmod 4 \equiv 0 \text{ or } i \bmod 4 \equiv 1 \end{cases} \tag{4}$$

where $3 \leq i \leq n$.

The set of vertices is $V = P \cup L$ and the set of edges E consists of all pairs $\{(p), [l]\}$ for which $(p)I_D[l]$. Bipartite graphs $D(n, \mathbb{K})$ have partition sets P (collection of points) and L (collection of lines) isomorphic to vector space $\mathbb{K}^n$, where $n \in \mathbb{N}_+$.

Graphs $\widetilde{D(n, \mathbb{K})}$

Formal definitions for the family of graphs $\widetilde{D(n, \mathbb{K})}$ were presented in [25].

Construction of projective limits graphs of $\widetilde{D(n, \mathbb{K})}$ appears in papers motivated by results on embeddings of Chevalley group geometries in the corresponding Lie algebras and construction of blow-up for an incidence system of Weyl groups in [46], [47]. Moreover, this structure is the base for construction of family of graphs $D(n, \mathbb{K})$ (see [25, 27]).

Let us use the analogical notations for points and lines in graph $\widetilde{D(K)}$:

$$(p) = (p_1, p_2, p_3, \ldots, p_n) \in P,$$
$$[l] = [l_{1,}, l_2, l_3, \ldots, l_n] \in L.$$

In the incidence structure $\widetilde{(P, L, I)}$ the point (p) is incident with the line $[l]$, and we write $(p)I_{\widetilde{D}}[l]$, if the following relations between their coordinates hold:

$$\begin{cases} l_2 - p_2 = l_1 p_1 \\ l_3 - p_3 = p_1 l_2 \\ l_4 - p_4 = l_1 p_2 \\ l_5 - p_5 = l_1 p_3 - p_1 l_4 \\ l_i - p_i = p_1 l_{i-1} \text{ for } i \bmod 3 \equiv 0 \\ l_i - p_i = l_1 p_{i-2} \text{ for } i \bmod 3 \equiv 1 \\ l_i - p_i = l_1 p_{i-2} - p_1 l_{i-1} \text{ for } i \bmod 3 \equiv 2 \end{cases} \tag{5}$$

for $3 \leq i \leq n$.

Graphs from families $D(n, \mathbb{K})$ and $\widetilde{D(n, \mathbb{K})}$ are bipartite, k-regular, where $|\mathbb{K}| = k$. The girth of graphs from the described families increases with the growth of n. In fact $D(n, q)$ is a family of graphs of large girth and there is a conjecture that $\widetilde{D(n, q)}$ is another family of graphs of a large girth.

All graphs from the considered families are k-regular, bipartite and the set of vertices is $V = P \cup L$, $P \cap L = \emptyset$. They are sparse graphs.

It is clear that there is a natural homomorphism of $T(n+1, \mathbb{K})$ onto $T(n, \mathbb{K})$ of "deleting the last coordinate" that sends $(x_1, x_2, \ldots, x_n, x_{n+1})$ to $(x_1, x_2, \ldots, x_n)$ and $[y_1, y_2, \ldots, y_n, y_{n+1}]$ to $[y_1, y_2, \ldots, y_n]$. It means that there is a well defined projective limit $T(K)$ of graphs $T(n, \mathbb{K})$, $n \to \infty$. Bivariate graphs form a special subclass of so called *linguistic graphs* for which natural projective limits are defined in a similar way.

Recall that the girth $g = g(\Gamma)$ of the graph Γ is the length of its minimal cycle.

Let us assume that the girth $g(n)$ of graphs $T(n, \mathbb{K})$ is unbounded. The obvious inequality $g(n+1) \geq g(n)$ holds. It means that projective limit $T(\mathbb{K})$ has to be a $|\mathbb{K}|$-regular forest. We have such situation in cases of graphs $A(n, \mathbb{F}_q)$ and $D(n, \mathbb{F}_q)$ If $q \geq 2$ then $A(\mathbb{F}_q)$ is a single tree presented by the above equations. Graph $D(F_q)$ is an infinite forest containing infinitely many trees.

Projective limit $W(\mathbb{F}_q)$ of Wenger graphs is an infinite connected graph containing cycles of length 8.

III. General encryption algorithm

We can convert graph $T(n,\mathbb{K})$ to finite automaton in the following way. Let $\mathrm{v}=(v_1,v_2,v_3,v_4,\dots,v_n)\in V(T(n,\mathbb{K}))$ (or $\mathrm{v}=[v_1,v_2,v_3,v_4,\dots,v_n]\in V(T(n,\mathbb{K}))$) and $N_\alpha(\mathrm{v})$ be the operator of taking neighbor of vertex v where the first coordinate is α:

$$N_\alpha(v_1,v_2,v_3,v_4,\dots,v_n)\to[\alpha,*,*,*,\dots,*],$$
$$N_\alpha[v_1,v_2,v_3,v_4,\dots,v_n]\to(\alpha,*,*,*,\dots,*),$$

where $\alpha\in\mathbb{K}$. The remaining coordinates can be determined uniquely using relations describing the chosen graph $T(n,\mathbb{K})$.

We convert $T(n,\mathbb{K})$ to finite automaton via joining v an $N_\alpha(\mathrm{v})$ by directed arrow with weight α. We assume that all vertices of the graph are accepting states.

A bit more interesting object is a symbolic bivariate automaton. Let $\mathrm{a}(x)=(\alpha_1(x),\alpha_2(x),\dots,\alpha_s(x))$ be a string of elements from $\mathbb{K}[x]$ (totality of polynomials in variable x with coefficients from K).

We introduce operator $N^s_{\mathrm{a(x)}}(\mathrm{v})$, where v is a point or a line with coordinates $v_1,v_2,\dots,v_n$, of taking the last vertex u of the path v, $\mathrm{v}_1=N_{\alpha_1(v_1)}(\mathrm{v})$, $\mathrm{v}_2=N_{\alpha_2(v_1)}(\mathrm{v}_1)$, …, $\mathrm{v}_s=N_{\alpha_s(v_1)}(\mathrm{v}_{s-1})=\mathrm{u}$.

We refer to $N^s_{\mathrm{a(x)}}$ as a computation of the symbolic automaton with the string

$$\mathrm{a(x)}=(\alpha_1(\mathrm{x}),\alpha_2(\mathrm{x}),\dots,\alpha_\mathrm{s}(\mathrm{x}))$$

$\alpha_i\in\mathbb{K}[x]$, $i=1,\dots,s$ and initial state $\mathrm{v}=(v_1,v_2,v_3,v_4,\dots,v_n)\in T(n,\mathbb{K})$ (or $\mathrm{v}=[v_1,v_2,v_3,v_4,\dots,v_n]\in T(n,\mathbb{K})$). We can consider $F_s(\mathrm{v})=N^s_{\mathrm{a(v_1)}}(\mathrm{v})$ as a map on $P\cup L$.

It is easy to see that the restriction of this map on P is a polynomial transformation of $P=\mathbb{K}^n$ into P (parameter s is even) or L (parameter s is odd) of kind

$$\begin{aligned}x_1&\to f_1(x_1,x_2,\dots,x_n),\\ x_2&\to f_2(x_1,x_2,\dots,x_n),\\ &\vdots\\ x_n&\to f_n(x_1,x_2,\dots,x_n).\end{aligned}$$

Notice that generally F_s is not a bijection. Let us consider an invertibility condition for F_s.

Proposition III.1. *Let the equations of kind $\alpha_s(x)=b$, $b\in\mathbb{K}$ have exactly one solution. Then map F_s is invertible.*

Proof: It is easy to check that if $F_s(\bar{x})=\bar{y}$ then $F_s^{-1}(\bar{y})=\bar{x}$. It is easy to see that $f_1(x_1,x_2,\dots,x_n)=\alpha_s(x_1)$. Let p be some point from P_n and $F_n(\mathrm{p})=(c_1,c_2,\dots,c_n)$ (point or line). Then the equation $\alpha_s(x_1)=c_1$ has a unique solution η. So we can compute $\eta_1=\alpha_1(\eta)$, $\eta_2=\alpha_2(\eta)$, …, $\eta_{s-1}=\alpha_{s-1}(\eta)$.

We can compute the chain $\mathrm{c}=(c_1,c_2,\dots,c_n)$, $N_{\eta_{s-1}}(\mathrm{c})=\mathrm{c}_1$, $N_{\eta_{s-2}}(\mathrm{c}_1)=\mathrm{c}_2$, …, $N_{\eta_1}(\mathrm{c}_{s-2})=\mathrm{c}_{s-1}$, $N_\eta((\mathrm{c}_{s-1}))=\mathrm{c}_s=(p_1,p_2,\dots,p_n)$ with $\eta=p_1$. So F_n is a bijection. ■

Notice that $N^s_{\mathrm{a}(x)}$ for $\mathrm{a}(x)$ of kind $\alpha_1(x)=\beta_1(x)$, $\alpha_2(x)=\beta_2(\alpha_1(x))$, $\alpha_3=\beta_3(\alpha_2(x))$, …, $\alpha_s(x)=\beta_s(\alpha_{s-1}(x))$ is a composition of ${N^1}_{\beta_1(x)}$, ${N^1}_{\beta_2(x)}$, …, ${N^1}_{\beta_s(x)}$. In this case invertibility of each $\beta_i(x)$, $i=1,2,\dots,s$ guarantees the bijectivity of $N^s_{\mathrm{a}(x)}$. We refer to such case as recurrently defined string.

Let L_1 and L_2 be sparse affine bijective transformation of the affine space (free module in other terminology) $\mathbb{K}^n$

$$\begin{aligned}L_1=T_{A,b}:\bar{x}&\longrightarrow\bar{x}A+b,\\ L_2=T_{C,d}:\bar{x}&\longrightarrow\bar{x}C+d,\end{aligned}$$

where $A=\left[a_{i,j}\right]$ and $C=\left[c_{i,j}\right]$ are $n\times n$ matrices with $a_{i,j},c_{i,j}\in\mathbb{K}$. It is clear that

$$\begin{aligned}L_1^{-1}&=T^{-1}_{A,b}=T_{A^{-1},-bA^{-1}},\\ L_2^{-1}&=T^{-1}_{C,d}=T_{C^{-1},-dC^{-1}}.\end{aligned}$$

Let F_n be a polynomial map of $\mathbb{K}^n$ to itself. We refer to $G_n=\tau_L F_n\tau_R$ as affine deformation of F_n.

Symmetric algorithm

We can use the data on the graph $T(n,\mathbb{K})$, the symbolic computation given by the string $\mathrm{a}=\mathrm{a}(x)$ of polynomials $\alpha_1(x)$, $\alpha_2(x)$, …, $\alpha_s(x)$, where $\alpha_s(x)$ is a bijective map of $\mathbb{K}$ to itself and affine transformations L_1 and L_2 in the following encryption scheme.

Correspondents Alice and Bob agree on a private encryption key

$$K_p=(L_1,L_2,\alpha=(\alpha_1,\alpha_2,\dots,\alpha_s)),$$

and keep the key in secret. Messages are written using characters belonging to the alphabet $\mathbb{K}$. So the plainspace is K^n and its elements must be treated as points (or lines) of the graph. To encrypt they use the composition

$$L_1\circ N^s_{\mathrm{a}}\circ L_2.$$

Notice that the computation has to be executed in numerical level:

1) Correspondent Alice writes plaintext $\mathrm{p}=(p_1,p_2,\dots,p_n)$ and treats it as point of the bivariate graph.
2) She computes parameters $\mu_i=\alpha_i(v_1)$ for $i=1,2,\dots,s$.
3) She computes p_0 as $L_1(\mathrm{p})$, p_1 as $N_{\mu_1}(\mathrm{p}_0)$, p_2 as $N_{\mu_2}(\mathrm{p}_1)$, …, p_s as $N_{\mu_s}(\mathrm{p}_{s-1})$.
4) She computes the ciphertext c as $L_2(\mathrm{p}_s)$.proo

Alice and Bob can use their knowledge about triple (L_1, L_2, a) for the decryption. Let us assume that Bob receives the ciphertext c from Alice. To decrypt the ciphertext Bob proceeds as follows:

1) He has to compute c_0 as ${L_2}^{-1}(\mathrm{c})$.
2) He treats the string of coordinates of this tuple as a vertex of the graph, which is a point in case of even s or the line in case of odd s with coordinates c_1^0, c_2^0, …, c_n^0.
3) Bob must find a solution η of $\alpha_s(x)=c_1^0$ and form a string $\eta_0=\eta$, $\eta_1=\alpha_1(\eta)$, $\eta_2=\alpha_2(\eta)$, …, $\eta_{s-1}=\alpha_{s-1}(\eta)$.
4) He computes c_1 as $N_{\eta_{s-1}}(\mathrm{c}_0)$, c_2 as $N_{\eta_{s-2}}(\mathrm{c}_1)$, …, c_s as $N_{\eta_0}(\mathrm{c}_{s-1})$.
5) He computes the plaintext p as ${L_1}^{-1}(\mathrm{c}_s)$.

Remark III.2. *In the case of identity maps L_1 and L_2 one can try Dijkstra's algorithm for finding the shortest path between plaintext and ciphertext. Notice that its complexity is $O(v \log v)$, but here v is exponential q^n. Therefore we get worse complexity even than brute force search via the key space.*

In the case of recurrently defined symbolic computation as above the encryption bijective map is $F_s = L_1 N^1{}_{\beta_1(x)} N^1{}_{\beta_1(x)} \dots N^1{}_{\beta_s(x)} L_2$. As we already see, this encryption transformation is equivalent to $L_1 N^s{}_{\mathrm{a}(x)} L_2$, where $\mathrm{a}(x) = (\beta_1(x), \beta_2(\beta_1(x)), \dots, \beta_s(\beta_{s-1}(\dots(\beta_1(x)))))$. Recurrently defined symbolic computation is an example of the polynomial map with an invertible decomposition in the sense of [31]. It has various applications in the development of multivariate key exchange protocols and asymmetric multivariate algorithm. The most popular case of implementation is related to graphs $D(n, \mathbb{K})$ (see [1, 21]) and $A(n, \mathbb{K})$ (see [22, 23]), where $\mathbb{K}$ is a finite field of arithmetical rings $\mathbb{Z}_m$ and strings of kind $\beta_1 = x + d_1$, $\beta_2 = x + d_2$, $\dots$, $\beta_s = x + d_s$, where $d_i + d_{i+1}$, $i = 1, 2, \dots, s-2$ are regular elements of the ring $\mathbb{K}$. We refer to such case as shifting encryption.

Let us consider the case of strong symmetric encryption, when the function is $\alpha_s(x) = ax+b$, with a regular (invertible) element of $\mathbb{K}$. In this case it is easy to show that degrees of encryption map F_n and decryption map ${F_n}^{-1}$ are the same. The advantage of this case is its universality. One can implement it in case of arbitrary chosen finite ring $\mathbb{K}$.

IV. On the properties of bivariate graph based bijective encryption maps

The girth G of simple graph G is the length of its shortest cycle. As it was established in [27] the girth of the graph $D(n, \mathbb{F}_q)$ is $\geq n+5$. So in the case of shifting encryption the map with the password $x + d_1$, $x + d_2$, $\dots$, $x + d_s$, $s < n+5$ the encryption map F_n has no fixed points. So ciphertext is always different from the plaintext. Let us consider deformated shifting encryption of kind $\tau_L F_n \tau_R$. We assume that affine maps τ_L and τ_R are fixed. Correspondents are able to change string $d_1, d_2, \dots, d_s$ for another one.

We assume that $d_i + d_{i+1} \neq 0$ for $i = 1, 2, \dots, s-2$. Such choice means that encryption map corresponds to the path of length s. The inequality $g(D(n, q)) \geq n + 5$ implies that different strings of length $s < (n+5)/2$ produce different ciphertexts. So even in the case when τ_L and τ_R are known to adversary the complexity of attacks without an access to unencrypted information is bounded from below by $q^{(n+5)/2}$.

In [44] these results were generalized for the case of general commutative ring $\mathbb{K}$. Let $\mathbb{M}$ be a multiplicative subset of $\mathbb{K}$, i. e. $\mathbb{M}$ is closed under the ring multiplication and it does not contain 0. We say that a string $d_1, d_2, \dots, d_s$ is $|\mathbb{M}|$-regular if $d_i + d_{i+1} \in \mathbb{M}$ for $i = 1, 2, \dots, s-2$. It was proven that different M-regular strings of length $s < (n+5)/2$ produce distinct ciphertexts from the same plaintext. So in the case of $|\mathbb{K}| = k$, $|\mathbb{M}| = m$ the resistance to attacks without access to unencrypted data is bounded from below by $mk^{(n+5)/2-1}$.

It was proven that graphs $A(n, \mathbb{F}_q)$ form a family of graphs of increasing girth $h(n)$ that tends to infinity as n grows. The speed of growth of $h(n)$ needs further evaluation. In [44] it was proven that different $|\mathbb{M}|$-regular strings of length $s < n$ produce different encryption maps.

Results on $|\mathbb{M}|$-regular strings of length restricted maps are obtained in terms of dynamical systems corresponding to graphs $D(n, \mathbb{K})$ and $A(n, \mathbb{K})$.

Let us assume that maps τ_L and τ_R are identities and consider the groups of transformations $GD(n, \mathbb{K})$ and $GA(n, \mathbb{K})$ generated by shifting encryption maps corresponding to strings of even length. In [40] was proven that all elements of $GD(n, \mathbb{K})$ are cubical transformations of affine spaces P_n and L_n. Similar result for $GA(n, \mathbb{K})$ is stated in [44]. As it follows instantly from this result transformation $F'_n = \tau_L F_n \tau_R$ and its inverse are cubical transformations.

The cryptanalytic corollary of this statement is justification of linearization attacks on stream ciphers corresponding to stream ciphers based on graphs $D(n, \mathbb{K})$ and $A(n, \mathbb{K})$.

Let correspondents use the transformation F'_n. The adversary has knowledge on the general scheme of open algorithm but not on the data for τ_L and τ_R and shifting string. So he knows about cubic nature of encryption. We assume that he has access to the unencrypted information and is able to intercept quite many pairs of kind (p, c), where p is plaintext and c corresponding ciphertext.

Then adversary writes G_n which is a formal cubical map in standard form with the unknown coefficients in front of monomial terms. He or she is able to solve system of $O(n^3)$ equations of kind $G_n(\mathrm{c}) = \mathrm{p}$ and restore the map G_n. So adversary could control the communication channel. The complexity of such direct linearization attack is $O(n^{10})$.

V. On the implementation of graph based stream cipher based on non bijective maps

Let us describe an implemented algorithm, which can run in the case of arbitrary commutative ring $\mathbb{K}$ and arbitrary bivariate graph $T(n, \mathbb{K})$. We slightly modify the above described symmetric algorithm based on bivariate graphs $T(n, \mathbb{K})$ which is not a case of shifting encryption. Firstly, we take a symbolic computation for string $\mathrm{a} = \mathrm{a}(x) = (\alpha_1(x), \alpha_2(x), \dots, \alpha_s(x))$, with $\alpha_i(x) = x^d + d_s$, $i = 1, 2, \dots, s$ where d is mutually prime with the order of $\mathbb{K}^*$. So equation $x^d + d_s = c$, $x \in \mathbb{K}^*$ has at most one solution. We take L_1 as an affine bijective transformation of kind $x_1 \to x_1 + x_2 + \dots + x_n$, $x_2 \to l_2(x_1, x_2, \dots, x_n)$, $x_3 \to l_3(x_1, x_2, \dots, x_n)$, $\dots$, $x_n \to l_n(x_1, x_2, \dots, x_n)$, where l_i are linear functions from $K[x_1, x_2, \dots, x_n]$. Correspondents will use the plainspace

$$\Omega = \{(x_1, x_2, \dots, x_n) | x_1 + x_2 + \dots + x_n \in \mathbb{K}^*, \\ x_i \in \mathbb{K}, i = 1, 2, \dots, n\}.$$

They will use $L_1 N^s_{\mathrm{a}(x)} L_2$ as encryption map. To execute computation in time $O(n)$ they take finite parameter s and use loaded tables for $\alpha_i(x)$, $i = 1, 2, \dots, s$ (one dimensional arrays $a_i(x)$, $x \in \mathbb{K}^*$). So they will compute $L_1(\mathrm{p}) = \mathrm{v} = (v_1, v_2, \dots, v_n)$, form sequence $\mu_i = \alpha_i(v_1)$, $i = 1, 2, \dots, s$

TABLE I
ENCODING AND DECODING TIME

		$A(n,\mathbb{K})$		$D(n,\mathbb{K})$		$\widetilde{D(n,\mathbb{K})}$	
Password	Filesize	Enc	Dec	Enc	Dec	Enc	Dec
3	1K	0.0021	0.0029	0.0030	0.0026	0.0039	0.0041
	10K	0.0217	0.0253	0.0234	0.0249	0.0322	0.0366
	50K	0.1030	0.1338	0.1034	0.1423	0.1572	0.1859
	100K	0.2158	0.2701	0.2115	0.2683	0.3309	0.3800
	500K	1.2202	1.3863	1.0432	1.3556	1.6161	1.9323
	1M	2.1955	2.8346	2.1452	2.7285	3.2809	3.9029
	10M	21.9597	27.4227	21.3803	26.6821	32.8819	38.3860
4	1K	0.0416	0.0033	0.0400	0.0032	0.0401	0.0047
	10K	0.0311	0.0320	0.0302	0.0360	0.0420	0.0466
	50K	0.1393	0.1639	0.1374	0.1580	0.2125	0.2366
	100K	0.2800	0.3314	0.2738	0.3280	0.4259	0.4816
	500K	1.4381	1.7109	1.3918	1.6541	2.1278	2.4159
	1M	2.9271	3.5035	2.8457	3.4055	4.3633	4.9664
	10M	29.5728	34.6022	28.6899	33.7773	43.7334	49.4341
5	1K	0.0402	0.0045	0.0336	0.0039	0.0437	0.0058
	10K	0.0355	0.0395	0.0382	0.0440	0.0533	0.0596
	50K	0.1764	0.2038	0.1718	0.1909	0.2589	0.2876
	100K	0.3510	0.4097	0.3391	0.3922	0.5243	0.5781
	500K	1.7778	2.0589	1.7237	2.0015	2.7088	3.0049
	1M	3.6421	4.2418	3.5507	4.1302	5.4671	6.0630
	10M	37.3170	42.0697	36.2427	40.9556	55.1103	60.4248
6	1K	0.0445	0.0053	0.0412	0.0046	0.0445	0.0069
	10K	0.0426	0.0481	0.0453	0.0448	0.0705	0.0667
	50K	0.2132	0.2371	0.1987	0.2325	0.3123	0.3462
	100K	0.4176	0.4830	0.4069	0.4678	0.6303	0.6890
	500K	2.1494	2.4572	2.0897	2.3724	3.2690	3.5826
	1M	4.3851	4.9386	4.2630	4.8109	6.7762	7.2091
	10M	47.8490	50.3557	42.6451	47.7372	65.8464	71.6511
7	1K	0.0434	0.0055	0.0435	0.0059	0.0487	0.0091
	10K	0.0477	0.0540	0.0475	0.0533	0.0754	0.0848
	50K	0.2437	0.2699	0.2324	0.2671	0.3651	0.3979
	100K	0.4903	0.5457	0.4751	0.5275	0.7315	0.7938
	500K	2.5089	2.8124	2.5655	2.7524	3.7086	4.0025
	1M	5.0959	5.7679	5.1230	5.6692	7.5859	8.2276
	10M	51.0014	56.3961	49.8712	54.9345	76.4318	87.4684

Algorithm 1 Encoding with graph $A(n,\mathbb{Z}_{256})$

Input: password $p = (p_0, p_1, \ldots, p_{k-1})$, message $m = (m_0, m_1, \ldots, m_{n-1})$
Output: encrypted message
$x = m$
for $i = 0, 1, ..., k-1$ **do**
  if $i \mod 2 = 0$ **then**
    $y_0 = (m_0^3 + p_i) \mod 256$
    for $j = 1, 2, \ldots, n$ **do**
      if $j \mod 2 \equiv 1$ **then**
        $y_j = (x_j - x_{j-1} \cdot y_0) \mod 256$
      else
        $y_j = (x_j - x_0 \cdot y_{j-1}) \mod 256$
  else
    $x_0 = (m_0^3 + p_i) \mod 256$
    $x_1 = (y_1 - y_0 \cdot x_0) \mod 256$
    for $j = 1, 2, \ldots, n$ **do**
      if $j \mod 2 \equiv 0$ **then**
        $x_j = (y_j + x_0 \cdot y_{j-1}) \mod 256$
      else
        $x_j = (y_j + x_{j-1} \cdot y_0) \mod 256$
if $k \mod 2 \equiv 1$ **then**
  return y
else
  return x

and compute recurrently $\mathrm{v}_i = N_{\mu_i}(\mathrm{v}_{i-1})$, $i = 1, 2, \ldots, s$. They form the ciphertext c as $L_2(\mathrm{v}_s)$.

To decrypt they will take $\mathrm{c}_0 = (c_1^0, c_2^0, \ldots, c_n^0)$ as ${L_2}^{-1}(\mathrm{c})$ and find a solution η for the equation $x^d + d_s = c_1^0$. Loaded table of values for ${\alpha_s}^{-1}$ will allow to find η fast. Next they form a string $\eta_0 = \eta$, $\eta_1 = \alpha_1(\eta)$, $\eta_2 = \alpha_2(\eta)$, $\ldots$, $\eta_{s-1} = \alpha_{s-1}(\eta)$. So users take string $\mathrm{c}_1 = N_{\eta_{s-1}}(\mathrm{c}_0)$, $\mathrm{c}_2 = N_{\eta_{s-2}}(\mathrm{c}_1)$, $\ldots$, $\mathrm{c}_s = N_{\eta_0}(\mathrm{c}_{s-1})$. Finally they get plaintext as $L^{-1}(\mathrm{c}_s)$.

The case of this symmetric algorithm appears as a private key for a cryptosystem introduced in [39] with the plainspace $\mathbb{Z}_m^n$.

We selected string of polynomials as $\alpha_i = x^d + d_i$, $d_i \in \mathbb{K}$, $i = 1, 2, \ldots, s$ and special linear transformations L_1 and L_2, given by the lists of linear forms.

We can theoretically evaluate degrees of encryption d_{enc} and decryption d_{dec}. In cases of graphs $D(n,\mathbb{K})$ and $A(n,\mathbb{K})$, these parameters are bounded below by some constants depending from parameters α_i, $i = 1, 2, \ldots, s$. We can select string of parameters and get d_{dec} large enough to make cryptoanalysis a difficult task. In case $\widetilde{D(n,\mathbb{K})}$ the degrees are even larger, they have size $O(n)$. Notice that direct linearization attacks are formally impossible because the encryption map is not a bijective one.

The implementation of the algorithms in the present work was done using the Python programming language, in particular version 2.7. The code doesn't use any out-of-the-box libraries for facilitating operations with matrices. The tests for measuring the processing time have been executed on a machine with Intel Core2 Duo CPU 9600 1.60GHz x 2, RAM memory 4.8 GB, operating with Ubuntu 16.04 LTS. The complexity of the algorithms is of order $O(sn)$, where s is the length of the password. In particular, we implement this stream cipher for case of $\mathbb{K} = \mathbb{Z}_{256}$ and $\alpha_i(x) = x^3 + d_i$ ($d = 3$ and $d_{\mathrm{dec}} = 43$), $i = 1, 2, \ldots, s$ without using loaded tables for functions. A description of the "nonlinear part" of encryption process, i. e. computation of N_{a}^s is presented below. We recommend a password for which d_2 and $d_i - d_{i+2}$, $i = 1, 2, \ldots, s-2$ are regular elements of the ring.

VI. CONCLUSION

The paper presents a class of stream ciphers defined in terms of graphs given by equations over the finite commutative ring $\mathbb{K}$. The algorithm has multivariate nature: plaintext is a tuple from the free module $\mathbb{K}^n$, key string is also an element of $\mathbb{K}^m$, the encryption map is polynomial transformation of $\mathbb{K}^n$ into itself. Users have options to vary parameters n and m and ring $\mathbb{K}$. If the parameter m is bounded by a constant, then the speed of numerical recurrent of encryption is $O(n)$. The key can be given as a sequence of polynomials in a single variable x. We observe results on simplest case of key strings $x + d_1$, $x + d_2$, $\ldots$, $x + d_s$ obtained by theoretical studies and via computer simulation in case of finite fields or arithmetical rings of kind $\mathbb{Z}_{2^m}$. In case of graphs $D(n,\mathbb{K})$ and $A(n,\mathbb{K})$ simple conditions on d_i ensure that different keys produce distinct ciphertexts and allow to estimate the complexity of adversary attacks without access to plaintext. In the above

Algorithm 2 Encoding with graph $D(n, \mathbb{Z}_{256})$

Input: password $p = (p_0, p_1, \ldots, p_{k-1})$, message $m = (m_0, m_1, \ldots, m_{n-1})$
Output: encrypted message
$x = m$
for $i = 0, 1, ..., k-1$ **do**
  if $i \mod 2 = 0$ **then**
    $y_0 = (m_0^3 + p_i) \mod 256$
    $y_1 = (x_1 + x_0 \cdot y_0) \mod 256$
    if $n \geq 2$ **then**
      $y_2 = (x_2 + x_0 \cdot y_1) \mod 256$
      if $n \geq 3$ **then**
        for $j = 3, 4, \ldots, n$ **do**
          if $j \mod 4 \equiv 3$ or $j \mod 4 \equiv 0$ **then**
            $y_j = (x_j + x_{j-2} \cdot y_0) \mod 256$
          else
            $y_j = (x_j + x_0 \cdot y_{j-2}) \mod 256$
  else
    $x_0 = (m_0^3 + p_i) \mod 256$
    $x_1 = (y_1 - y_0 \cdot x_0) \mod 256$
    if $n \geq 2$ **then**
      $x_2 = (y_2 - y_1 \cdot x_0) \mod 256$
      if $n \geq 3$ **then**
        for $j = 3, 4, \ldots, n$ **do**
          if $j \mod 4 \equiv 3$ or $j \mod 4 \equiv 0$ **then**
            $x_j = (y_j - y_0 \cdot x_{j-2}) \mod 256$
          else
            $x_j = (y_j - y_{j-2} \cdot x_0) \mod 256$
if $k \mod 2 \equiv 1$ **then**
  return y
else
  return x

Algorithm 3 Encoding with graph $\widetilde{D(n, \mathbb{Z}_{256})}$

Input: password $p = (p_0, p_1, \ldots, p_{k-1})$, message $m = (m_0, m_1, \ldots, m_{n-1})$
Output: encrypted message
$x = m$
for $i = 0, 1, ..., k-1$ **do**
  if $i \mod 2 = 0$ **then**
    $y_0 = (m_0^3 + p_i) \mod 256$
    $y_1 = (x_1 + x_0 \cdot y_0) \mod 256$
    if $n \geq 2$ **then**
      for $j = 2, 3, \ldots, n$ **do**
        if $j \mod 3 \equiv 2$ **then**
          $y_j = (x_j + (x_0 \cdot y_{j-1}) \mod 256$
        else if $j \mod 3 \equiv 0$ **then**
          $y_j = (x_j + x_{j-2} \cdot y_0) \mod 256$
        else
          $y_j = (x_j + x_{j-2} \cdot y_0 - x_0 \cdot y_{j-1}) \mod 256$
  else
    $x_0 = (m_0^3 + p_i) \mod 256$
    $x_1 = (y_1 - y_0 \cdot x_0) \mod 256$
    if $n \geq 2$ **then**
      for $j = 2, 3, \ldots, n$ **do**
        if $j \mod 3 \equiv 2$ **then**
          $x_j = (y_j - y_{j-1} \cdot x_0) \mod 256$
        else if $j \mod 3 \equiv 0$ **then**
          $x_j = (y_j - y_0 \cdot x_{j-2}) \mod 256$
        else
          $x_j = (y_j - y_0 \cdot x_{j-2} + y_{j-1} \cdot x_0) \mod 256$
if $k \mod 2 \equiv 1$ **then**
  return y
else
  return x

mentioned case encryption and decryption maps are cubical and adversary after the interception of $O(n^3)$ pairs of kind plaintext-ciphertext can conduct a linearization attack in time $O(n^{10})$. In case of $\widetilde{D(n, \mathbb{K})}$ the degree of both maps grows linearly with the growth of parameter n, which makes the search for the inverse map via linearization attacks a difficult task. Additionally, authors started investigation of bijective and non-bijective encryption maps with keys of kind $x^d + d_1$, $x^d + d_2$, ..., $x^d + d_s$, where $d > 1$.

In the non-bijective case the plainspace is large subset of $\mathbb{K}^n$ and the adversary has to restore the multivariate encryption transformation E and search for polynomial map E' such that EE' fixes each plaintext. Known methods do not allow to solve this task in polynomial time. Special case with high degree E' is implemented. Loaded tables for x^d allow a fast encryption of text even in case of large parameter d.

REFERENCES

[1] V. Ustimenko, *Coordinatisation of Trees and their Quotients*, in the Voronoj's Impact on Modern Science, Kiev, Institute of Mathematics, 1998, vol. 2, 125-152.

[2] D. Sharma, V. Ustimenko, *Special Graphs in Cryptography*, The Poster Papers Collection,Third International Workshop on Practice and Theory in Public Key Cryptography (PKC 2000), Melbourne Exhibition Centre, Australia, January 2000, p. 16- 19.

[3] V. Ustimenko, *CRYPTIM: Graphs as Tools for Symmetric Encryption*, Lecture Notes in Computer Science, Springer, LNCS 2227, Proceedings of AAECC-14 Symposium on Applied Algebra, Algebraic Algorithms and Error Correction Codes, November 2001, p. 278 - 286.

[4] V. Ustimenko, *Graphs with special arcs and cryptography* , Acta Applicandae Mathematicae (Kluwer) 2002, 74,117-153

[5] Yu, Khmelevsky, V. Ustimenko, *Walks on graphs as symmetric and asymmetric tools for encryption*, 2002, South Pacific Journal of Natural Studies, 2002, vol. 20, 23-41. www.usp.ac.fj/spjns

[6] Yu.Khmelevsky, M. Govorov, P. Sharma, V.Ustimenko, S. Dhanjal, *Security Solutions for Spatial Data in Storage (Implementation Case within Oracle 9iAS)*, Proceedings of 8th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2004) Orlando, USA, in July 18-21, 2004, pp 318-323.

[7] M. Govorov, Yu Khmelevsky, A. Khorev, V. Ustimenko, *Security Control for Spatial Warehouses*, Proceedings ofthe 21th International Cartographic Conference (ICC), Durban, South Africa, 2003, 1784-1794.

[8] Yu Khmelevsky, V Ustimenko, Practical aspects of the Informational Systems reengineering, The South Pacific Journal of Natural Science, volume 21, 2003, p.75-21 (www.usp.ac.fj/spjns/volume21).

[9] A. Tousene, V. Ustimenko, *CRYPTALL - a System to Encrypt All types of Data*, Notices of Kiev - Mohyla Academy , v. 23, 2004, pp 12-15.

[10] A. Tousene, V. Ustimenko,*Graph based private key crypto-system*, International Journal on Computer Research, Nova Science Publisher, vol.13, issue 4, 2005, 12p.

[11] V. Ustimenko, *On the extremal graph theory for directed graphs and its cryptographical applications*, Advances in Coding Theory and Cryptography, Series on Coding Theory and Cryptology, vol. 3, World Scientific, 181-200 (2007).

[12] V. Futorny, V. Ustimenko, *On small world semiplanes with generalised Schubert cells*, Acta Applicandae Mathematicae, 98, N1 (2007) 47–61 (with V. Futorny).

[13] V. Ustimenko, *On the graph based cryptography and symbolic computations*, Serdica Journal of Computing, Proceedings of International Conference on Applications of Computer Algebra 2006, Varna, N1 (2007).

[14] J. Kotorowicz, V. A. Ustimenko, *On the implementation of cryptoalgorithms based on algebraic graphs over some commutative rings*, Condenced Matters Physics, Special Issue: Proceedings of the international conferences “Infinite particle systems, Complex systems theory and its application”, Kazimerz Dolny, Poland, 2006, 11 (no. 2(54)) (2008) 347–360.

[15] V. Ustimenko, *On the hidden discrete logarithm for some polynomial stream ciphers*, International Multiconference on Computer Science and Informational Technology, 20-22 October 2008, Wisla, Poland, CANA Proceedings. 13 pp.

[16] S. Kotorowicz, V. Ustimenko, *On the properties of Stream Ciphers Based on Extremal Directed graphs*, In ”Cryptography Research Perspectives”, Nova Publishers, Ronald E. Chen (the editor), 2008.

[17] A. Touzene, V. Ustimenko, *Private and Public Key Systems Using Graphs of High Girth*,In ”Cryptography Research Perspectives”, Nova Publishers, Ronald E. Chen (the editor), 2008, pp.205-216.

[18] M. Klisowski, V. Ustimenko, *On the public keys based on the extremal graphs and digraphs*, International Multiconference on Computer Science and Informational Technology, October 2010, Wisla, Poland, CANA Proceedings, 12 pp.

[19] Y. Khmelevsky, Gaetan Hains, E. Ozan, Chris Kluka, V. Ustimenko and D. Syrotovsky) International Cooperation in SW Engineering Research Projects, Proceedings of Western Canadien Conference on Computing Education, University of Northen British Columbia,Prince George BC, May 6-7, 2011, 14pp.

[20] A. Touzene, V. Ustimenko, Marwa AlRaisi, Imene Boudelioua*Performance of Algebraic Graphs Based Stream-Ciphers Using Large Finite Fields*, Annalles UMCS Informatica AI X1, 2 (2011), 81-93.

[21] M. Klisowski, V. Ustimenko, *On the implementation of cubic public rules based on algebraic graphs over the finite commutative ring and their symmetries*, MACIS 2011: Fourth International Conference on Mathematical Aspects of Computer and Information Sciences, Beijing, 2011, 13 pp.

[22] M. Klisowski, U. Romanczuk, V. Ustimenko,*On public keys based on a new family of algebraic graphs* , Annalles UMCS Informatica AI X1, 2 (2011), 127 -141.

[23] J. Kotorowicz, U. Romanczuk, V. Ustimenko, *Implementation of stream ciphers based on a new family of algebraic graphs*, Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, 13 pp.

[24] M. Klisowski, V. Ustimenko, *On the Comparison of Cryptographical Properties of Two Different Families of Graphs with Large Cycle Indicator*, Mathematics in Computer Science, 2012, Volume 6, Number 2, Pages 181-198.

[25] M. Polak, V. Ustimenko, *On stream cipher based on a family of graphs D(n, q) of increasing girth*, Albanian J. Math. 8 (2014), no. 2, 37-44.

[26] F. Lazebnik and V. Ustimenko, *Some Algebraic Constructions of Dense Graphs of Large Girth and of Large Size*, DIMACS series in Discrete Math- ematics and Theoretical Computer Science, V. 10 (1993), 75-93.

[27] F. Lazebnik, V. Ustimenko, *Explicit construction of graphs with an arbitrary large girth and of large size*, Discrete Appl. Math. , 60, (1995), 275 - 284.

[28] R. Wenger, *Extremal graphs with no C4, C6 and C10s*, 1991, J. Comb. Theory, Ser. B, 52(1),113-116.

[29] Ding J., Gower J. E., Schmidt D. S., *Multivariate Public Key Cryptosystems*, Springer, Advances in Information Security, V. 25, 2006.

[30] Polak M., Romańczuk U., Ustimenko V. and Wróblewska A., *On the applications of Extremal Graph Theory to Coding Theory and Cryptography*, Erdős Centennial, Proceedings of Erdős Centennial (EP 100), Electronic Notes in Discrete Mathematics,V43, P. 329–342 2013.

[31] Ustimenko V. A., *Explicit constructions of extremal graphs and new multivariate cryptosystems* Studia Scientiarum Mathematicarum Hungarica, Special issue ”Proceedings of The Central European Conference, 2014, Budapest”.

[32] V. A. Ustimenko, *On the cryptographical properties of extreme algebraic graphs*, in Algebraic Aspects of Digital Communications, IOS Press (Lectures of Advanced NATO Institute, NATO Science for Peace and Security Series - D: Information and Communication Security, Volume 24, July 2009, 296 pp.

[33] Ustimenko V. A., *On the flag geometry of simple group of Lie type and Multivariate Cryptography*, Algebra and Discrete Mathematics. V. 19. No 1. 2015. P. 130-144.

[34] Louis Goubin, Jacques Patarin, Bo-Yin Yang,*Multivariate Cryptography. Encyclopedia of Cryptography and Security*, (2nd Ed.) 2011, 824-828.

[35] Patarin J., *The Oil i Vinegar digital signatures*, Dagstuhl Workshop on Cryptography. 1997.

[36] Kipnis A., Shamir A., *Cryptanalisis of the Oil and Vinegar Signature Scheme* Advances in Cryptology - Crypto 96, Lecture Notes in Computer Science, V. 1462, 1996, P. 257–266.

[37] Bulygin S., A. Petzoldt A., and Buchmann J, *Towards provable security of the unbalanced oil and vinegar signature scheme under direct attacks*, In Guang Gong and KishanChand Gupta, editors, ”Progress in Cryptology - INDOCRYPT”, Guang Gong and Kishan Chand Gupta, editors, Lecture notes in Computer Science, V. 6498, 2010. P. 17–32.

[38] Romańczuk-Polubiec U., Ustimenko V, *On two windows multivariate cryptosystem depending on random parameters* Algebra and Discrete Mathematics. 2015. V. 19. No. 1. P. 101–129.

[39] V. Ustimenko, *On algebraic graph theory and non-bijective maps in cryptography*, Algebra and Discrete Mathematics, Volume 20 (2015). Number 1, pp. 152-170.

[40] A.Wróblewska, *On some properties of graph based public keys*, Albanian Journal of Mathematics, Volume 2, Number 3, 2008, 229-234, NATO Advanced Studies Institute: ”New challenges in digital communications”.

[41] U. Romańczuk, V. Ustimenko, *On regular forests given in terms of algebraic geometry, new families of expanding graphs with large girth and new multivariate cryptographical algorithms*, Proceedings of International conference ”Applications of Computer Algebra”, Malaga, 2013, p. 135-139.

[42] M. Klisowski, *Zwiekszenie bezpieczeństwa kryptograficznych algorytmów wielu zmiennych bazujacych na algebraicznej teorii grafów*, PhD thesis, Czestochowa, 2014.

[43] V. Ustimenko,*Linguistic Dynamical Systems, Graphs of Large Girth and Cryptography*, Journal of Mathematical Sciences, Springer, vol.140, N3 (2007) pp. 412-434.

[44] V. Ustimenko, U. Romańczuk, *On Dynamical Systems of Large Girth or Cycle Indicator and their applications to Multivariate Cryptography*, in ”Artificial Intelligence, Evolutionary Computing and Metaheuristics ”, In the footsteps of Alan Turing Series: Studies in Computational Intelligence, Volume 427/January 2013, pp. 231-256.

[45] V. Ustimenko, U. Romańczuk, *On Extremal Graph Theory, Explicit Algebraic Constructions of Extremal Graphs and Corresponding Turing Encryption Machines*, in ”Artificial Intelligence, Evolutionary Computing and Metaheuristics ”, In the footsteps of Alan Turing Series: Studies in Computational Intelligence, Volume 427/January 2013, pp. 257-285.

[46] U. Romańczuk, V. Ustimenko, *On the key exchange with matrices of large order and graph based nonlinear maps*, Albanian Journal of Mathematics, Volume 4, Number 4, pp. 203–211 (2010).

[47] V. Ustimenko, *Division algebras and Tits geometries*, DNAUSSR 296, No. 5 (1987), 1061-1065 (Russian)

[48] V. Ustimenko, *A linear interpretation of the flag geometries of Chevalley groups*, Kiev University, Ukrainskii Matematicheskii Zhurnal 42, No. 3 (March, 1990), 383-387; English transl.

[49] M. Polak, *On the applications of algebraic graph theory to coding*, PhD thesis, Maria Curie-Skłodowska, 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 407–416
DOI: 10.15439/2018F43
ISSN 2300-5963 ACSIS, Vol. 15

Group Anonymity in Security Protocols

Ferucio Laurenţiu Ţiplea and Cosmin Vârlan
Department of Computer Science, "Al.I.Cuza" University of Iaşi
Iaşi, Romania, e-mail: {ferucio.tiplea@uaic.ro, vcosmin@info.uaic.ro}

***Abstract*—Group anonymity, as an instance of information hiding, means that an agent is not identifiable within a group of agents with respect to an observer. In this paper we define group anonymity in security protocols by taking into account two types of observers: honest agents, as local observers of the protocol execution, and intruders (active or passive), as global observers of the protocol execution. It is shown that an action may be group anonymous in a protocol under a passive intruder but not in the same protocol under an active intruder, and vice versa. In case of basic-term actions, group anonymity in a protocol under an active intruder implies group anonymity in the same protocol under a passive intruder. A broad spectrum of relationships between group anonymity for various types of actions is developed, as well as relationships between group anonymity, minimal anonymity, and role interchangeability. Finally, the decidability and complexity status of the decision problems induced by these concepts is completely discussed. Thus, it is shown that group anonymity and role interchangeability are undecidable in unrestricted protocols. Group anonymity is complete for NEXPTIME when it is restricted to basic-term actions and bounded security protocols, and it is complete for NP when it is restricted to basic-term actions and 1-session bounded security protocols.**

***Index Terms*—Security protocol, anonymity, decision problem, epistemic logic**

I. Introduction

OVER the last two decades there has been a growing interest in methods for anonymous communication and in developing techniques for reasoning about information hiding properties in security protocols [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26]. This is mainly due to the applications of the anonymous communication to various fields such as e-voting, e-commerce, e-mail, e-cash and so on.

Information hiding embraces many forms such as *anonymity*, *unlinkability*, *indistinguishability*, *role interchangeability*, *undetectability*, *unobservability*, and *identity management*. In an effort to standardize the terminology on information hiding, Pfitzmann and Hansen have written and maintained since 2000 a consolidated proposal of terminology on information hiding properties [23].

Anonymity is a prominent information hiding property to which a lot of research has been dedicated. Using a multi-agent system and epistemic logic based framework, Halpen and O'Neill [11] classified the anonymity into:

- minimal anonymity – an action performed by an agent is not always seen by an observer;
- group anonymity – an agent who performed some action is not "identifiable", by an observer, within a set of agents. Sender and recipient anonymity in [23] are instances of group anonymity.

Contribution: In two earlier papers [24], [26], we have investigated the minimal anonymity in security protocols. In this paper we do the same for group anonymity. Using the framework developed in [24], [26], we define group anonymity for security protocols. As we are using six types of send actions and six types of receive actions, our approach covers a large spectrum of group anonymity concepts met in the literature on information hiding properties. Several basic relationships between all these concepts of group anonymity are established in the paper.

Anonymity is a property that depends on the observer of the protocol execution. We consider two types of observers: honest agents and the intruder. An honest agent is a local observer of the protocol execution; he can only record information about the actions which involve him. The intruder is a global observer of the protocol execution who can record all actions in the protocol execution. Honest agents and the intruder as observers may have incomparable deductive powers due to the fact that honest agents may know secret information unknown to the intruder, while the intruder may have more information about the actions performed in the protocol. This is clearly reflected in the results obtained in the paper.

As we prove in our paper, group anonymity highly depends on the intruder type: passive or active. Thus, we show that there are group anonymous actions in protocols under passive intruders which are not group anonymous if the intruder is active. That is, an active intruder may destroy the group anonymity property of an action. More interestingly is that the converse holds true as well: there are group anonymous actions in protocols under active intruders which are not group anonymous if the intruder is passive. That is, an active intruder may induce some degree of anonymity in security protocols.

The relationships between minimal anonymity, group anonymity, and role interchangeability are also discussed. Thus, we show that any group anonymous exclusive action is minimally anonymous (in the same security protocol), and any role interchangeable observable action is group anonymous (in the same security protocol).

As decision problem, group anonymity is proven undecidable in unrestricted security protocols. If we restrict group anonymity to basic-term actions and bounded security protocols, then it is complete for NEXPTIME. If one more restriction is added by requiring just 1-session protocol executions, then group anonymity becomes complete for NP.

Related work: The seminal work that marked the development of a formal study of anonymity-related properties is that of David Chaum [1], [2], [3] who proposed a method by which an agent can send a message to some other agent without revealing his identity. Since then, several formalisms for anonymity have been proposed. Thus, [4] proposes a formalization of anonymity in the CSP framework. [6] and [11] focus on anonymity in security protocols and multi-agent systems, respectively, by using an epistemic logic framework. These two papers have greatly influenced the research on anonymity based on an epistemic logic formalism [12], [15], [20], [24], [25], [26]. The roots of our paper can be traced back to these two papers too. While [6], [11] have just offered the basis for an epistemic logic based approach to anonymity, [24], [26] have proposed a rich inference system to reason about anonymity in security protocols. Moreover, based on this, many results on minimal anonymity, such as decidability and complexity results, have been developed in [24], [26]. Our paper continues along the same line, by offering a large spectrum of results on group anonymity in security protocols.

A rather different but very interesting approach to anonymity was proposed by Hughes and Shmatikov [8] by using *function views*. The *cryptographic protocol logic* (CPL) proposed in [27] came as an ambitious general framework for formalizing a very large class of security properties. While CPL seems very expressive, the model checking problem for it is undecidable and not too much about decidable fragments and proof systems for the core CPL is known.

Structure of the paper: The paper is organized into seven sections. The formal model we use in this paper for security protocols is recalled in Section 2. The group anonymity concepts studied in this paper are introduced in Section 3, while Section 4 presents basic properties of these concepts. Section 5 is dedicated to the study of the decidability status of the decision problems induced by our group anonymity concepts, while Section 6 discusses complexity issues. We conclude in Section 7.

Due to the space limitation, the proofs of the results were not included in the final version of the paper.

II. Modeling Security Protocols

The formalism used in this paper to model security protocols is precisely the one in [28], [29], [30], [24], [26]. Therefore, we will only recall its basic notations and terminology (for more details the reader is referred to the papers cited above).

A *security protocol signature* is a 3-tuple $\mathcal{S} = (\mathcal{A}, \mathcal{K}, \mathcal{N})$ consisting of a finite set $\mathcal{A}$ of *agent names* (or shortly, *agents*) and two at most countable sets $\mathcal{K}$ and $\mathcal{N}$ of *keys* and *nonces* (numbers once used), respectively. It is assumed that:

- $\mathcal{A}$ contains a special element denoted by I and called the *intruder*. All the other elements are called *honest agents* and Ho denotes their set;
- $\mathcal{K} = \mathcal{K}_0 \cup \mathcal{K}_1$, where $\mathcal{K}_0$ is the set of *short-term keys* and $\mathcal{K}_1$ is a finite set of *long-term keys*. The elements of $\mathcal{K}_1$ are of the form K_A^e (A's public key), or K_A^d (A's private key), or K_{AB} (shared key by A and B), where A and B are distinct agents;
- some honest agents A may be provided from the beginning with some *secret information* $Secret_A \subseteq \mathcal{K}_0 \cup \mathcal{N}$, not known to the intruder. $Secret_A$ does not contain long-term keys because they will never be communicated by agents during the runs;
- the intruder is provided from the beginning with a set of nonces $\mathcal{N}_I \subseteq \mathcal{N}$ and a set of short-term keys $\mathcal{K}_{0,I} \subseteq \mathcal{K}_0$. It is assumed that no elements in $\mathcal{N}_I \cup \mathcal{K}_{0,I}$ can be generated by honest agents.

The set of *basic terms* is $\mathcal{T}_0 = \mathcal{A} \cup \mathcal{K} \cup \mathcal{N}$. The set $\mathcal{T}$ of *terms* is defined inductively: every basic term is a term; if t_1 and t_2 are terms, then (t_1, t_2) is a term (meaning concatenation of messages); if t is a term and K is a key, then $\{t\}_K$ is a term (meaning t encrypted by K). We extend the construct (t_1, t_2) to $(t_1, \ldots, t_n)$ as usual by letting $(t_1, \ldots, t_n) = ((t_1, \ldots, t_{n-1}), t_n)$, for all $n \geq 3$. Sometimes, parenthesis will be omitted. Given a term t, $Sub(t)$ is the set of all *sub-terms* of t (defined as usual). This notation is extended to sets of terms by union.

The length of a term is defined as usual, by taking into consideration that pairing and encryption are operations. Thus, $|t| = 1$ for any $t \in \mathcal{T}_0$, $|(t_1, t_2)| = |t_1| + |t_2| + 1$, for any terms t_1 and t_2, and $|\{t\}_K| = |t| + 2$, for any term t and key K.

The *perfect encryption assumption* we adopt [31] states that a message encrypted with a key K can be decrypted only by an agent who knows the corresponding inverse K^{-1} of K, and the only way to compute $\{t\}_K$ is by encrypting t with K.

There are two types of actions, send and receive. A *send action* is of the form $A!B : (M)t$, and a *receive action* is of the form $A?B : t$. In both cases, A is assumed an honest agent who *performs the action*, $A \neq B$, $t \in \mathcal{T}$ is the *term of the action*, and $M \subseteq Sub(t) \cap (\mathcal{N} \cup \mathcal{K}_0)$ is the *set of new terms of the action*. $M(a)$ denotes M, if $a = A!B : (M)t$, and the empty set, if $a = A?B : t$; $t(a)$ stands for the term of a. When $M = \emptyset$ we will simply write $A!B : t$. For a sequence of actions $w = a_1 \cdots a_l$ and an agent A, define the *restriction of w to A*, denoted $w|_A$, as being the sequence obtained from w by removing all actions not performed by A. The notations $M(a)$ and $t(a)$ are extended to sequences of actions by union.

A *security protocol* (or simply, *protocol*) is a triple $\mathcal{P} = (\mathcal{S}, \mathcal{C}, w)$, where $\mathcal{S}$ is a security protocol signature, $\mathcal{C}$ is a subset of $\mathcal{T}_0$, called the set of *constants* of $\mathcal{P}$, and w is a non-empty sequence of actions, called the *body* of the protocol, such that no action in w contains the intruder. Constants are publicly known elements in the protocol that cannot be re-instantiated (as it will be explained below). As usual, $\mathcal{C}$ does not include private keys, elements in $Secret_A$ for any honest agent A, or elements in $\mathcal{N}_I$, $\mathcal{K}_{0,I}$ and $M(w)$. Any non-empty sequence $w|_A$, where A is an agent, is called a *role* of the protocol. A role specifies the actions a participant should perform in a protocol, and the order of these actions.

An example of a security protocol is given in Figure 1. In this example, the server S wants to get an opinion from its clients regarding the network services provided by it (the

clients are A and B in our example). Therefore, S generates a fresh short term key K and sends it to A and B. These agents compose some messages (their opinions) t and t' and send them, encrypted by K, to H. When H has collected all the messages (opinions), it forwards them to S.

$$\begin{array}{lcl} S\,!\,A & : & (\{K\})\{K,H\}_{K_{SA}} \\ A\,?\,S & : & \{K,H\}_{K_{SA}} \\ S\,!\,B & : & \{K,H\}_{K_{SB}} \\ B\,?\,S & : & \{K,H\}_{K_{SB}} \\ A\,!\,H & : & \{\{t\}_K,S\}_{K_{AH}} \\ H\,?\,A & : & \{\{t\}_K,S\}_{K_{AH}} \\ B\,!\,H & : & \{\{t'\}_K,S\}_{K_{BH}} \\ H\,?\,B & : & \{\{t'\}_K,S\}_{K_{BH}} \\ H\,!\,S & : & \{\{t\}_K,\{t'\}_K\}_{K_{SH}} \\ S\,?\,H & : & \{\{t\}_K,\{t'\}_K\}_{K_{SH}} \end{array}$$

Fig. 1. A running example

Instantiations of a protocol are given by *substitutions*, which are functions σ that map agents to agents, nonces to arbitrary terms, short-term keys to short-term keys, and long-term keys to long-term keys. Moreover, for long-term keys, σ should satisfy $\sigma(K_A^e) = K_{\sigma(A)}^e$, $\sigma(K_A^d) = K_{\sigma(A)}^d$, and $\sigma(K_{AB}) = K_{\sigma(A)\sigma(B)}$, for any distinct agents A and B. Substitutions are homomorphically extended to terms, actions, and sequences of actions. A substitution σ is called *suitable for an action* $a = AxB : y$ if $\sigma(A)$ is an honest agent, $\sigma(A) \neq \sigma(B)$, and σ maps distinct nonces from $M(a)$ into distinct nonces, distinct keys into distinct keys, and it has disjoint ranges for $M(a)$ and $Sub(t(a)) - M(a)$. σ is *suitable for a sequence of actions* if it is suitable for each action in the sequence, and σ is *suitable for a subset* $C \subseteq \mathcal{T}_0$ if it is the identity on C.

An *event* of a protocol $\mathcal{P} = (\mathcal{S}, \mathcal{C}, w)$ is any triple $e_i = (u, \sigma, i)$, where $u = a_1 \cdots a_l$ is a role of $\mathcal{P}$, σ is a substitution suitable for u and $\mathcal{C}$, and $1 \leq i \leq l$. $\sigma(a_i)$ is the *action of the event* e_i. As usual, $act(e_i)$ ($t(e_i)$, $M(e_i)$) stands for the the action of e_i (term of e_i, set of new terms of e_i). The *local precedence relation* on events is defined by $(u, \sigma, i) \rightarrow (u', \sigma', i')$ if and only if $u' = u$, $\sigma' = \sigma$, and $i' = i + 1$, provided that $i < |u|$. $\stackrel{+}{\rightarrow}$ is the transitive closure of $\rightarrow$. Given an event e, ${}^\bullet e$ stands for the *set of all local predecessors of* e, i.e., ${}^\bullet e = \{e' | e' \stackrel{+}{\rightarrow} e\}$.

Given X a set of terms, $analz(X)$ stands for the least set which includes X, contains t_1 and t_2 whenever it contains (t_1, t_2), and contains t whenever it contains $\{\{t\}_K\}_{K^{-1}}$ or $\{t\}_K$ and K^{-1}. By $synth(X)$ we denote the least set which includes X, contains (t_1, t_2), for any terms $t_1, t_2 \in synth(X)$, and contains $\{t\}_K$, for any term t and key K in $synth(X)$. Moreover, $\overline{X}$ stands for $synth(analz(X))$.

A *state* of a protocol $\mathcal{P}$ is an indexed set $s = (s_A | A \in \mathcal{A})$, where s_A is A's (local) state, for any agent A. The traditional approach to security protocols defines agent states as sets of messages (all messages sent and received by the agent during some computation). This approach is quite sufficient if one wants to reason about confidentiality [28], [29], [30]. However, this is not enough to reason about anonymity properties, where more information about the actions performed by agents in the protocol are needed. One way to solve this is to add *facts* to agent states [6], [24], [26]. A fact is a sentence of the form $P(t_1, \ldots, t_i)$, where P is a predicate symbol and $t_1, \ldots, t_i$ are message terms (facts beginning by the same predicate symbol P will also be called *P-facts*). Using the approach in [24], [26], we will use six classes of facts which are illustrated on the protocol in Figure 1:

1) $sent$-facts. Each agent X who sends a message t to some agent Y records a fact $sent(X, t, Y)$. For instance, when the first action of the protocol in Figure 1 will be performed, S records $sent(S, \{K, H\}_{K_{SA}}, A)$;
2) rec-facts. Two cases are to be considered here:
   a) *passive intruder*. If an action $X\,?\,Y : t$ was performed by X, then X may safely record a fact $rec(X, t, Y)$ because he knows that the message he received is from Y;
   b) *active intruder*. If an action $X\,?\,Y : t$ was performed by X, then X might not be sure whether t comes from Y or from the intruder. In such a case X records a fact $rec(X, t, (Y, I))$ which tells him that t may be from Y or from I.

   If the second action in the protocol in Figure 1 has been performed in some computation, and the intruder was active, then A records the fact $rec(A, \{K, H\}_{K_{SA}}, (S, I))$;
3) $shared\_key$-facts. In the first action of the protocol, the agent S generates a short-term key K and sends it to A. Therefore, K acts as a short term *shared key* between S and A. We use $shared\_key(Z, X, Y, K)$ to mean that Z randomly generated a short term key K to be used by X and Y as a shared-key. In our protocol in Figure 1, $shared\_key(S, S, A, K)$ is the fact to be recorded by A when the first action of the protocol is performed;
4) gen-facts. The message in the first action of the protocol in our running example is *generated by S for A* because it is encrypted by the long term shared key K_{SA}; denote this by $gen(S, \{K, H\}_{K_{SA}}, A)$ and record it in S's state. Similarly, A will record the fact $gen(A, \{t\}_K, S)$ in his state when fifth action of the protocol is performed;
5) $auth$-facts. A message t encrypted by X's private key K_X^d is authenticated by X. The fact $auth(X, (t, \{t\}_{K_X^d}))$ denotes this;
6) hop-facts. These are facts of the form $hop(A, C, B, t)$ whose meaning is that B can only received t from A via C (examples of hope facts can be found in [26]).

According to our discussion above, an agent A state is a pair $s_A = (s_{A,m}, s_{A,f})$, where $s_{A,m}$ is a set of messages and $s_{A,f}$ is a set of facts. Intuitively, $s_{A,m}$ represents the set of all messages the agent A sent or received in some computation from the initial state to the state s_A, and $s_{A,f}$ represents the set of facts which give information about the actions the agent A performed in that computation.

The protocol computation rule in [28], [29], [30] has to be changed accordingly [24], [26]. Given two states s and s' and

an action a, we write $s[a\rangle s'$ if and only if:

1) if a is of the form $A!B : (M)t$, then:
   a) $t \in \overline{s_{A,m} \cup M}$ and $M \cap Sub(s) = \emptyset$;
   b) $s'_{A,m} = s_{A,m} \cup M \cup \{t\}$, $s'_{I,m} = s_{I,m} \cup \{t\}$, and $s'_{C,m} = s_{C,m}$ for any $C \in \mathcal{A} - \{A, I\}$;
   c) the facts in s' are obtained as follows:
      i) add $sent(A, t, B)$ to $s_{A,f}$ and $s_{I,f}$;
      ii) if some term $t_1 = \{t'\}_{K_{AC}}$ or $t_1 = \{t'\}_{K^e_C}$ or $t_1 = \{t'\}_K$ has been built by A in order to build t, where K is a short-term shared key by A and some agent C and A owns this key, then add $gen(A, t_1, C)$ to $s_{A,f}$;
      iii) if some term $t_1 = (t', \{t'\}_{K^d_A})$ has been built by A in order to build t, then add $auth(A, t_1)$ to $s_{A,f}$;
      iv) if some short-term key K has been generated by A to be used as a shared key by two agents C and D, and K is a part of t, then add $shared\_key(A, C, D, K)$ to $s_{A,f}$;
      v) $s'_{C,f} = s_{C,f}$, for any $C \in \mathcal{A} - \{A, I\}$;
2) if a is of the form $A?B : t$, then:
   a) $t \in \overline{s_{I,m}}$;
   b) $s'_{A,m} = s_{A,m} \cup \{t\}$ and $s'_{C,m} = s_{C,m}$, for all $C \in \mathcal{A} - \{A\}$;
   c) the facts in s' are obtained as follows:
      i) add $rec(A, t, (B, I))$ to $s_{A,f}$ and $s_{I,f}$;
      ii) if A received a key K as part of t and he knows that K was generated by some agent C to be shared by A with another agent D, then add $shared\_key(C, A, D, K)$ to $s_{A,f}$;
      iii) if A received a message t' as part of t and he knows that this message comes from some agent C via another agent D, then add $hop(C, D, A, t')$ to $s_{A,f}$;
      iv) $s'_{C,f} = s_{C,f}$, for any $C \in \mathcal{A} - \{A, I\}$.

In the case of a passive intruder (2a) should be "$t \in \overline{s_{B,m}}$" and (2ci) above should be "add $rec(A, t, B)$ to $s_{A,f}$ and $s_{I,f}$". If we remove (1c) and (2c) from the computation rule above, we obtain the standard computation rule in [28], [29], [30].

At each point in the evolution of a protocol, each agent may derive new facts from the facts he owns at that point. The derivation process is guided by deduction rules. In order to present these rules we need first two basic concepts. A message t is called *decomposable* [24], [26] over an agent state $s = (s_m, s_f)$ if $t \in \mathcal{T}_0$, or $t = (t_1, t_2)$ for some messages t_1 and t_2, or $t = \{t'\}_K$ for some message t' and key K with $K^{-1} \in analz(s_m)$, or $gen(A, t, B) \in s_f$ for some honest agents A and B ("$gen(A, t, B)$" covers the case when A generates t for B by encrypting some message by B's public key. A does not know B's corresponding private key but knows how he built t and, from this point of view, we may say that t is decomposable). The function $trace(t, s)$ [24], [26], where t is a message and $s = (s_m, s_f)$ is an agent state, is given by:

- $trace(t, s) = \{t\}$, if $t \in \mathcal{T}_0$;
- $trace(t, s) = \{t\} \cup trace(t_1, s) \cup trace(t_2, s)$, if $t = (t_1, t_2)$ for some terms t_1 and t_2;
- $trace(t, s) = \{t\}$, if t is not decomposable over s;
- $trace(t, s) = \{t\} \cup trace(t', s)$, if $t = \{t'\}_K$ is an encrypted but decomposable message over s.

The deduction rules (Table I) we use are those from [24] with slight modifications [26] (a rule with one or two indexes specifies the current state where the rule should be applied; for instance, $(RShR)_{A,C}$ means that the rule $(RShR)$ should be applied in A's or C's current state. A rule with no indexes means that the rule can be applied in any state). We discuss just one of the rules in Table I, namely $(RShR)$ (for a complete discussion about them, the reader is referred to [26]). According to this rule, if A received a message $\{t\}_K$ encrypted by a short-term key distributed by C to him and to B, then surely B received the key from C.

Given a set M of messages and a set F of facts, denote by $Analz(M, F)$ the set of all facts that can be inferred from F and M. If $s = (s_m, s_f)$ is an agent state, then $Analz(s)$ stands for $Analz(s_m, s_f)$.

III. Defining Anonymity

Anonymity in a security protocol is a property that has to be defined w.r.t. an observer of the protocol execution. In our approach, the observer is either an honest agent (as in [11], [24], [26]) or the intruder (passive [6], [10], or active [24], [26]). Honest agents as observers are limited to observing some of the actions performed by the agents who interact with him, while passive intruders as observers are capable to observe the entire protocol execution. On the other side, honest agents may have more deductive power than passive intruders because they may know secret keys unknown to intruders. Therefore, from the anonymity point of view, honest agents and passive intruder as observers have incomparable powers.

While [6], [10] have considered only passive intruders as observers, in [24], [26] active intruders were taken into consideration too. This is because an action may be anonymous w.r.t. a passive intruder but not w.r.t. an active one, and vice-versa (see [26] and Theorem 5 in this paper).

Observers draw conclusions about protocol executions by analyzing their current states. If two current states are "equivalent", then the conclusions should be equivalent. We formalize this as follows. Given a pair of agent states (s, s') define the binary relation $\sim_{s,s'}$ on message terms by [24], [26]:

- $t \sim_{s,s'} t$, for any $t \in \mathcal{T}_0$;
- $t \sim_{s,s'} t'$, for any term t undecomposable over s and any term t' undecomposable over s';
- $(t_1, t_2) \sim_{s,s'} (t'_1, t'_2)$, for any terms t_1, t_2, t'_1, and t'_2 with $t_1 \sim_{s,s'} t'_1$ and $t_2 \sim_{s,s'} t'_2$;
- $\{t\}_K \sim_{s,s'} \{t'\}_K$, for any terms t and t' and any key K with $t \sim_{s,s'} t'$ and $K^{-1} \in analz(s_m) \cap analz(s'_m)$.

Extend $\sim_{s,s'}$ to facts by $P(t_1, \ldots, t_i) \sim_{s,s'} P(t'_1, \ldots, t'_i)$ if $t_j \sim_{s,s'} t'_j$ for any $1 \leq j \leq i$.

Two agent states $s = (s_m, s_f)$ and $s' = (s'_m, s'_f)$ are called *observationally equivalent* [24], [26], denoted $s \sim s'$, if:

TABLE I
DEDUCTION RULES

$$(S1)\ \frac{sent(A,t,B)}{sent(A,t), sent(A,B), sent(t,B)} \qquad (R1)\ \frac{rec(A,t,x)}{rec(A,t), rec(A,x), rec(t,x)} \qquad (RS)\ \frac{rec(A,t,B)}{sent(B,t,A)}$$

$$(S2)\ \frac{sent(A,B)}{sent(A)} \qquad (R2)\ \frac{rec(A,x)}{rec(A)} \qquad (RGS)_A\ \frac{rec(A,t),\ gen(B,t,A)}{sent(B,t,A)}$$

$$(S3)\ \frac{sent(A,t)}{sent(A), sent(t)} \qquad (R3)\ \frac{rec(A,t)}{rec(A), rec(t)} \qquad (RAS)\ \frac{rec(t),\ auth(A,t)}{sent(A,t)}$$

$$(S4)\ \frac{sent(t,B)}{sent(t)} \qquad (R4)\ \frac{rec(t,x)}{rec(t)} \qquad (SGS)_{A,B}\ \frac{sent(A,t),\ gen(A,t,B)}{sent(A,t,B)}$$

$$(S5)\ \frac{sent(A,t,B),\ t' \in trace(t,s)}{sent(A,t',B)} \qquad (R5)\ \frac{rec(A,t,x),\ t' \in trace(t,s)}{rec(A,t',x)} \qquad (RA)\ \frac{rec(t,\{t\}_{K_A^d})}{auth(A,(t,\{t\}_{K_A^d}))}$$

$$(RG)_A\ \frac{rec(A,\{t\}_{K_{AB}}),\ \neg gen(A,\{t\}_{K_{AB}},B)}{gen(B,\{t\}_{K_{AB}},A)} \qquad (RG')_A\ \frac{rec(A,\{t\}_K),\ shared\_key(C,A,B,K),\ \neg gen(A,\{t\}_K,B)}{gen(B,\{t\}_K,A)}$$

$$(RGR)_{A,B}\ \frac{rec(A,t,(B,I)),\ gen(B,t,A)}{rec(A,t,B)} \qquad (SGR)_{B,C}\ \frac{sent(A,t,B),\ gen(C,t,B), hop(C,A,B,t)}{rec(A,t,C)}$$

$$(RShR)_{A,C}\ \frac{rec(A,\{t\}_K),\ shared\_key(C,A,B,K),\ \neg gen(A,\{t\}_K,B)}{rec(B,K,C)}$$

- $analz(s_m) \cap \mathcal{T}_0 = analz(s'_m) \cap \mathcal{T}_0$;
- $(\forall \varphi \in Analz(s))(\exists \varphi' \in Analz(s'))(\varphi \sim_{s,s'} \varphi')$;
- $(\forall \varphi' \in Analz(s'))(\exists \varphi \in Analz(s))(\varphi' \sim_{s',s} \varphi)$.

That is, s and s' are observationally equivalent if the agent can derive the same meaningful information from any of these two states. In other words, these two states are *indistinguishable*.

Two protocol states s and s' are *observationally equivalent w.r.t. an agent* A, denoted $s \sim^A s'$, if $s_A \sim s'_A$.

It was shown in [24], [26] that the observational equivalence on agent states is an equivalence relation decidable in $\mathcal{O}(f^4 l^4)$ time complexity, where f is the maximum number of facts and l is the maximum length of the messages in the states.

We use a fragment of the epistemic logic in [32], [11] to reason about anonymity. Its syntax is

$$\varphi ::= p \,|\, \varphi \wedge \varphi \,|\, \neg\varphi \,|\, \mathtt{K}_A\varphi$$

where p ranges over a countable set Φ of atomic propositions, A ranges over a non-empty finite set $\mathcal{A}$ of agent names, and φ in $\mathtt{K}_A\varphi$ does not contain any $\mathtt{K}$ operator. Denote by $\mathcal{L}(\Phi,\mathcal{A})$ the set of all formulas defined as above. As usual we use $\mathtt{P}_A\varphi$ as an abbreviation for $\neg\mathtt{K}_A\neg\varphi$.

Let $\mathcal{P}$ be a security protocol. The *truth value of a formula* $\varphi \in \mathcal{L}(\Phi,\mathcal{A})$ *in* $\mathcal{P}$ is defined as follows:

- $\mathcal{P} \models \varphi$ iff $(\mathcal{P}, s) \models \varphi$, for any reachable state s in $\mathcal{P}$;
- $(\mathcal{P}, s) \models p$ iff $(\mathcal{P}, s_A) \models p$, for some agent $A \in \mathcal{A} - \{I\}$;
- $(\mathcal{P}, s) \models \neg\varphi$ iff $(\mathcal{P}, s) \not\models \varphi$;
- $(\mathcal{P}, s) \models \varphi \wedge \psi$ iff $(\mathcal{P}, s) \models \varphi$ and $(\mathcal{P}, s) \models \psi$;
- $(\mathcal{P}, s) \models \mathtt{K}_A\varphi$ iff $(\mathcal{P}, s'_A) \models \varphi$, for any reachable state s' with $s' \sim^A s$;
- for any formula φ without $\mathtt{K}$ operators and any $A \in \mathcal{A}$, $(\mathcal{P}, s_A) \models \varphi$ is defined as follows:
  - if $\varphi = p$ then $(\mathcal{P}, s_A) \models \varphi$ iff $p \in Analz(s_A)$;
  - if $\varphi = \varphi_1 \wedge \varphi_2$ then $(\mathcal{P}, s_A) \models \varphi$ iff $(\mathcal{P}, s_A) \models \varphi_1$ and $(\mathcal{P}, s_A) \models \varphi_2$;
  - if $\varphi = \neg\varphi_1$ then $(\mathcal{P}, s_A) \models \varphi$ iff $(\mathcal{P}, s_A) \not\models \varphi_1$.

We shall simply write $s \models \varphi$ ($s_A \models \varphi$) instead of $(\mathcal{P}, s) \models \varphi$ ($(\mathcal{P}, s_A) \models \varphi$), whenever $\mathcal{P}$ is understood from the context.

The formula $\mathtt{K}_A\varphi$ means "agent A knows φ". It holds in a reachable state s if it holds in any reachable state that is observationally equivalent to s w.r.t. A. $\mathtt{P}_A\varphi$ means "agent A thinks that φ is possible". It holds in a state s if it holds in some reachable state observationally equivalent to s w.r.t. A.

Anonymity in security protocols will be defined for *actions*

performed by agents, w.r.t. some observer. By an *action* we will understand a $sent$-fact (also called $sent$-*action*), or a rec-fact that does not contain terms of the form (B, I) (also called rec-*action*). Therefore, the $sent$-actions are of the form $sent(A, t, B)$, $sent(A, t)$, $sent(A, B)$, $sent(A)$, $sent(t)$, or $sent(t, B)$, while the rec-actions are of the form $rec(A, t, B)$, $rec(A, t)$, $rec(A, B)$, $rec(A)$, $rec(t)$, or $rec(t, B)$. By act we will denote a generic action of the one of the forms above.

Each action, except for $sent(t)$, $sent(t, B)$, $rec(t)$, and $rec(t, B)$, is performed by exactly one agent, namely, the first argument of the corresponding $sent$- or rec-fact. These actions are also called *mono-agent actions*. The actions $sent(t)$, $sent(t, B)$, $rec(t)$, and $rec(t, B)$ may be performed by more than one agent; they will be called *multi-agent actions*. If act is a mono-agent action performed by some agent A, then we also write $act(A)$ just to specify the agent who performs the action. If act is a multi-agent action, such as $sent(t)$, $sent(t, B)$, $rec(t)$, or $rec(t, B)$, we also write $act(t)$ just to specify the message term involved in the action.

Definition 1: Let $\mathcal{P}$ be a security protocol, G a nonempty set of agents, T a finite set of message terms, and X an observer (agent) not in G.

1) A mono-agent action $act(A)$ of $\mathcal{P}$ is *anonymous within G w.r.t. X* if $\mathcal{P} \models \psi(act(A), G, X)$, where $\psi(act(A), G, X) = (\mathtt{P}_X act(A) \Rightarrow \bigwedge_{C \in G} \mathtt{P}_X act(C))$.
2) A multi-agent action $act(t)$ of $\mathcal{P}$ is *anonymous within T w.r.t. X* if $\mathcal{P} \models \psi(act(t), T, X)$, where $\psi(act(t), T, X) = (\mathtt{P}_X act(t) \Rightarrow \bigwedge_{t' \in T} \mathtt{P}_X act(t'))$.
3) An action $sent(A, t)$ is *role interchangeable* within $G \times T$ w.r.t. X if the following property holds:

$$\mathcal{P} \models \mathtt{P}_X sent(A, t) \Rightarrow \bigwedge_{C \in G,\, t' \in T} (\mathtt{P}_X sent(C, t') \Rightarrow \mathtt{P}_X (sent(A, t') \wedge sent(C, t))).$$

A few explanations about these concepts are in order:

- Anonymity of $act(A)$ within G w.r.t. X in $\mathcal{P}$ means that, whenever X thinks that $act(A)$ is possible at some state s then, for any $C \in G$, X thinks that $act(C)$ is possible at some state s' observationally equivalent to s.
- Role interchangeability simply means that, from the observer's point of view, two actions may be interchanged between two distinct agents.

Sender (receiver) anonymity within a set of senders (receivers), as defined in [23], is a special case of anonymity of a mono-agent $sent$-action (rec-action) within a set of agents.

The anonymity concepts introduced in Definition 1(1)(2) are also called *group anonymity* concepts, and the sets G and T in these definitions are called *anonymity sets*. Thus, group anonymity says that the agent who performs an action or the message which purports an action is not identifiable within a set (group) of agents or messages, respectively.

We want to emphasize that the anonymity of an action which contains messages, such as $sent(A, t)$, should not be confused with the secrecy of t. The anonymity of $sent(A, t)$ within G w.r.t. X means that X is not sure whether A sent the message t because he was able to deduce that any member of G sent at some point in the protocol the message t (although X might knew the message t).

Example 2: Figure 2 presents a sequence of inferences in the state s of the protocol in Figure 1, obtained by playing all actions of the protocol (the right hand side column indicates the inference process). It is not difficult to see that:

1.	$shared\_key(S, S, A, K)$	$\in s_S$
2.	$shared\_key(S, S, B, K)$	$\in s_S$
3.	$rec(S, \{\{t\}_K, \{t'\}_K\}_{K_{SH}}, (H, I))$	$\in s_S$
4.	$rec(S, \{\{t\}_K, \{t'\}_K\}_{K_{SH}})$	$3, R1$
5.	$rec(S, \{t\}_K)$	$4, R5$
6.	$rec(S, \{t'\}_K)$	$4, R5$
7.	$\neg gen(S, \{t\}_K, A)$	$\in s_S$
8.	$\neg gen(S, \{t\}_K, B)$	$\in s_S$
9.	$\neg gen(S, \{t'\}_K, A)$	$\in s_S$
10.	$\neg gen(S, \{t'\}_K, B)$	$\in s_S$
11.	$gen(A, \{t\}_K, S)$	$5, 1, 7, (RG')_S$
12.	$gen(B, \{t\}_K, S)$	$5, 2, 8, (RG')_S$
13.	$gen(A, \{t'\}_K, S)$	$6, 1, 9, (RG')_S$
14.	$gen(B, \{t'\}_K, S)$	$6, 2, 10, (RG')_S$
15.	$sent(A, \{t\}_K, S)$	$5, 11, (RGS)_S$
16.	$sent(B, \{t\}_K, S)$	$5, 12, (RGS)_S$
17.	$sent(A, \{t'\}_K, S)$	$6, 13, (RGS)_S$
18.	$sent(B, \{t'\}_K, S)$	$6, 14, (RGS)_S$

Fig. 2. Examples of inferences by the rules in Table I

- $sent(A, t)$ is anonymous within $\{A, B\}$ w.r.t. S (that is, S cannot clearly identify whether A or B sent t);
- $sent(A, t)$ is anonymous within $\{t, t'\}$ w.r.t. S (that is, S cannot clearly identify whether A sent t or t');
- $sent(A, t)$ is role interchangeable within $\{A, B\} \times \{t, t'\}$ w.r.t. S (that is, from S's point of view, A could have send t and B could have send t', or vice versa).

IV. Relating anonymity concepts

A. Basic properties of group anonymity

An action act of a security protocol is called a *basic-term action* if all terms in the action are basic terms. For instance, $sent(A, N_A, B)$, where N_A is a nonce, is a basic-term action, whereas the action $sent(A, \{N_A\}_K, B)$ is not. From definitions we obtain:

Lemma 3: For any basic-term action act, any agent X, and any protocol states s and s', the following property holds: if $s' \sim^X s$ then $s'_X \models act$ if and only if $s_X \models act$.

Proposition 4: A basic-term action $act(x)$ is anonymous within a group G of basic terms w.r.t. X in a protocol $\mathcal{P}$ if and only if, for any reachable state s in $\mathcal{P}$, $s_X \models act(x)$ implies $(\forall y \in G)(s_X \models act(y))$.

Proof. Assume that $act(x)$ is anonymous within a group G w.r.t. X in $\mathcal{P}$, and let s be a reachable state in $\mathcal{P}$ such that $s_X \models act(x)$.

The anonymity of $act(x)$ within G leads to the fact that for any $y \in G$ there exists a reachable state s', observationally equivalent to s w.r.t. X, such that $s'_X \models act(y)$. Lemma 3 leads then to $s_X \models act(y)$. As a conclusion, we obtain

$$s_X \models act(x) \;\Rightarrow\; (\forall y \in G)(s_X \models act(y))$$

The converse is obtained in a similar way. ■

Anonymity highly depends on the intruder type, passive or active. This is shown by Theorem 5 below: there are group anonymous actions in protocols under passive intruders which are not group anonymous if the intruder is active, and vice versa, there are group anonymous actions in protocols under active intruders which are not group anonymous if the intruder is passive.

Theorem 5:

1) There are protocols $\mathcal{P}$, actions $act(x)$, groups G of agents or message terms, and observers X such that $act(x)$ is anonymous within G w.r.t. X in $\mathcal{P}$ under a passive intruder, but $act(x)$ is not anonymous within G w.r.t. X in $\mathcal{P}$ under an active intruder.
2) There are protocols $\mathcal{P}$, actions $act(x)$, groups G of agents or message terms, and observers X such that $act(x)$ is anonymous within G w.r.t. X in $\mathcal{P}$ under an active intruder, but $act(x)$ is not anonymous within G w.r.t. X in $\mathcal{P}$ under a passive intruder.
3) For any protocol $\mathcal{P}$, basic-term action $act(x)$, group G of agents or basic terms, and observer X, if $act(x)$ is anonymous within G w.r.t. X in $\mathcal{P}$ under an active intruder, then $act(x)$ is anonymous within G w.r.t. X in $\mathcal{P}$ under a passive intruder.

The following results relate the group anonymity concepts for various actions.

Theorem 6: Let $\mathcal{P}$ be a security protocol with the property that for any agents $A, X \in \mathcal{A} - \{I\}$, any message t, and any reachable state s, if $s_X \models (rec(t) \wedge auth(A,t))$ then there exists $B \in \mathcal{A} - \{A, I\}$ such that $s_X \models sent(A,t,B)$. Then, the following properties hold in $\mathcal{P}$ (G is a set of agents, T is a set of messages, and X is an observer):

1) If $\bigwedge_{B \in \mathcal{A}-\{A,I\}} \psi(sent(A,t,B),G,X)$ holds in $\mathcal{P}$, then $\psi(sent(A,t),G,X)$ holds in $\mathcal{P}$;
2) If $\bigwedge_{t \in \mathcal{T}} \psi(sent(A,t,B),G,X)$ holds in $\mathcal{P}$, then $\psi(sent(A,B),G,X)$ holds in $\mathcal{P}$;
3) If $\bigwedge_{B \in \mathcal{A}-\{A,I\}} \psi(sent(A,B),G,X)$ holds in $\mathcal{P}$, then $\psi(sent(A),G,X)$ holds in $\mathcal{P}$;
4) If $\bigwedge_{t \in \mathcal{T}} \psi(sent(A,t),G,X)$ holds in $\mathcal{P}$, then $\psi(sent(A),G,X)$ holds in $\mathcal{P}$;
5) If $\bigwedge_{B \in \mathcal{A}-\{I\}} \psi(sent(t,B),T,X)$ holds in $\mathcal{P}$, then $\psi(sent(t),T,X)$ holds in $\mathcal{P}$.

The hypothesis in Theorem 6 is quite natural: if an agent X receives a message authenticated by some agent A, then he draw the conclusion that A sent that message to some other agent B.

Figure 3 pictorially represents the implications in Theorem 6. Moreover, it is not difficult to find examples of protocols

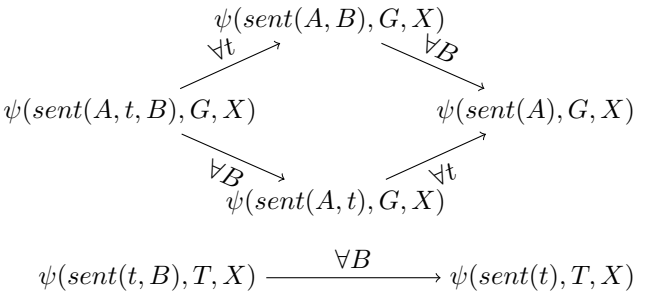


Fig. 3. Relationships between group anonymity concepts

where $\psi(sent(A,B),G,X)$ holds but $\psi(sent(A,t),G,X)$ does not hold, and vice versa. That is, $\psi(sent(A,B),G,X)$ and $\psi(sent(A,t),G,X)$ are incomparable.

Under the *sender identifiability restriction*, Theorem 6 holds for rec-actions too.

Definition 7: Let $\mathcal{P}$ be a security protocol. An action $rec(A,t)$ ($rec(A)$, $rec(t)$, resp.) is *sender identifiable* if, for any X and reachable state s of $\mathcal{P}$ with $s_X \models rec(A,t)$ ($s_X \models rec(A)$, $s_X \models rec(t)$, resp.) there exists B such that $s_X \models rec(A,t,B)$ ($s_X \models rec(A,B)$, $s_X \models rec(t,B)$, resp.).

It is obvious that $rec(A,t)$, $rec(A)$, and $rec(t)$ are all sender identifiable in any protocol $\mathcal{P}$ under a passive intruder. Similar to Theorem 6 we obtain the following result.

Theorem 8: The following properties hold in any security protocol $\mathcal{P}$ (G is a set of agents, T is a set of messages, and X is an observer but not I):

1) If $\bigwedge_{B \in \mathcal{A}-\{A,I\}} \psi(rec(A,t,B),G,X)$ holds in $\mathcal{P}$ and $rec(A,t)$ is sender identifiable, then $\psi(rec(A,t),G,X)$ holds in $\mathcal{P}$;
2) If $\mathcal{P} \models \bigwedge_{t \in \mathcal{T}} \psi(rec(A,t,B),G,X)$ holds in $\mathcal{P}$, then $\psi(rec(A,B),G,X)$ holds in $\mathcal{P}$;
3) If $\bigwedge_{B \in \mathcal{A}-\{A,I\}} \psi(rec(A,B),G,X)$ holds in $\mathcal{P}$ and $rec(A)$ is sender identifiable, then $\psi(rec(A),G,X)$ holds in $\mathcal{P}$;
4) If $\bigwedge_{t \in \mathcal{T}} \psi(rec(A,t),G,X)$ holds in $\mathcal{P}$ and $rec(A)$ is sender identifiable, then $\psi(rec(A),G,X)$ holds in $\mathcal{P}$;
5) If $\bigwedge_{B \in \mathcal{A}-\{I\}} \psi(rec(t,B),T,X)$ holds in $\mathcal{P}$ and $rec(t)$ is sender identifiable, then $\psi(rec(t),T,X)$ holds in $\mathcal{P}$.

B. Minimal and group anonymity

An action act of a protocol $\mathcal{P}$ is *minimally anonymous w.r.t.* X [24], [26] if $\mathcal{P} \models (act \Rightarrow \neg \mathrm{K}_X act)$. Using a multi-agent framework, it has been shown [11] that any *exclusive action* which is anonymous within a group of agents is also minimally anonymous (w.r.t. the same observer). The exclusiveness of an action means that no two different agents can perform the action. This result holds in our framework too.

Definition 9: Let $\mathcal{P}$ be a security protocol and A an honest agent. An action $act(A)$ performed by A is *locally exclusive* if $s_B \models \neg(act(A) \wedge act(A'))$, for any reachable state s of $\mathcal{P}$, any honest agent B, and any honest agent $A' \neq A$.

Proposition 10: If a locally exclusive action $act(A)$ of a security protocol $\mathcal{P}$ is anonymous within G w.r.t. an honest agent H and $|G| \geq 3$, then $act(A)$ is minimally anonymous w.r.t. H.

If the agent H in Proposition 10 is replaced by the intruder, then Proposition 10 might not hold. This is because if an action $act(A)$ is in some state of H, then the action still may be anonymous within some set G w.r.t. H, but it is definitely not minimally anonymous w.r.t. H.

If the local exclusiveness of an action fails to hold, then the conclusion of Proposition 10 may fail too.

C. Role interchangeability and group anonymity

As it was remarked in [12], [20] using a multi-agent framework, role interchangeability implies group anonymity, under certain conditions. We recall this result here in our security protocol framework.

Definition 11: Let $\mathcal{P}$ be a security protocol, $G \subseteq \mathcal{A} - \{I\}$ a set of agents, and $T \subseteq \mathcal{T}$ a finite set of message terms, and X an observer. We say that an action $sent(A,t)$ is $(G \times T)$-*observable w.r.t.* X if the following property holds

$$s_X \models (sent(A,t) \;\Rightarrow\; \bigwedge_{C \in G} \bigvee_{t' \in T} sent(C,t'))$$

for any reachable state s in $\mathcal{P}$.

One can easily prove now the following result:

Proposition 12: Let $\mathcal{P}$ be a security protocol, $G \subseteq \mathcal{A} - \{I\}$ a set of agents, $T \subseteq \mathcal{T}$ a finite set of message terms, and X an observer. If X is not in G and $sent(A,t)$ is role interchangeable within $G \times T$ and $(G \times T)$-observable w.r.t. X, then $sent(A,t)$ is anonymous within G w.r.t. X.

Role interchangeability can similarly be formulated for actions like $sent(A,B)$, $rec(A,t)$, or $rec(A,B)$.

V. Deciding Group Anonymity

In this section we establish several undecidability results for the anonymity concepts defined so far. The proofs are based on the undecidability of the halting problem for counter machines and various reduction techniques.

Each action has a *type* which is a tuple. For instance, $sent(A,t,B)$ has type (s,a,m,a) and $rec(A,t)$ has type (r,a,m), where s stands for "$sent$", r stands for "rec", a for "agent", and m for "message".

Each action type τ induces two decision problems:

1) the *group anonymity problem for type τ actions w.r.t. an honest agent*, abbreviated $GAP(\tau)$, which is the problem to decide, given a security protocol $\mathcal{P}$, a type τ action act, a non-empty set G of honest agents or messages, and an honest agent H not in G, whether act is anonymous within G w.r.t. H in $\mathcal{P}$;
2) the *group anonymity problem for type τ actions w.r.t. the intruder*, abbreviated $GAP_I(\tau)$, which is the problem to decide, given a security protocol $\mathcal{P}$, a type τ action act, a non-empty set G of honest agents or messages, whether act is anonymous within G w.r.t. the intruder.

Now, we can prove the following theorem.

Theorem 13:
1) *$GAP(\tau)$ is undecidable in unrestricted security protocols, for any action type τ.*
2) *$GAP_I(\tau)$ is undecidable in unrestricted protocols, for any $sent$-action type τ.*

In Section IV-B it has been shown that group anonymity implies minimal anonymity in case of exclusive actions. However, exclusiveness is undecidable.

Theorem 14: Local exclusiveness problems is undecidable in unrestricted security protocol.

Role interchangeability is an undecidable problem too. It is the problem to decide, given a security protocol $\mathcal{P}$, an action $sent(A,t)$, a group $G \subseteq \mathcal{A} - \{I\}$ of agents, a finite set $T \subseteq \mathcal{T}$ of message terms, and an honest observer H, whether $sent(A,t)$ is role interchangeable within $G \times T$ w.r.t. H.

Theorem 15: Role interchangeability is undecidable in unrestricted security protocols.

VI. Complexity of Group Anonymity

The group anonymity problem is undecidable in unrestricted security protocols. Clearly, if we focus on bounded security protocols then group anonymity is decidable. In this section we study the complexity of this problem. Recall first a few concepts regarding bounded protocols [30].

Let $\mathcal{P} = (\mathcal{S}, \mathcal{C}, w)$ be a security protocol, $T \subseteq \mathcal{T}_0$ a finite set, and $k \geq 1$. A (T,k)-*run* of $\mathcal{P}$ is any run with the property that all terms in the run are built up upon T and all messages communicated in the course of the run have length at most k. When for $\mathcal{P}$ only (T,k)-runs are considered we say that it is a *protocol under (T,k)-runs* or a (T,k)-*bounded protocol*, and denote this by $(\mathcal{P},T,k)$. A *bounded protocol* is a (T,k)-bounded protocol, for some finite set $T \subseteq \mathcal{T}_0$ and $k \geq 1$.

A *1-session (T,k)-run* of $\mathcal{P}$ is any (T,k)-run of $\mathcal{P}$ obtained by applying each role at most once (not necessarily in its entirety), under the same substitution (i.e., all its events are defined by using the same substitution). Therefore, any 1-session (T,k)-run has length at most $|w|$. When for the protocol $\mathcal{P}$ only 1-session (T,k)-runs are considered we say that it is a *1-session (T,k)-bounded protocol*. A *1-session bounded protocol* is a 1-session (T,k)-bounded protocol, for some finite set $T \subseteq \mathcal{T}_0$ and $k \geq 1$.

In [30] it has been shown that the number of distinct events in a (T,k)-run of a protocol $\mathcal{P}$ is exponential in $poly(size(\mathcal{P}))$, where $size(\mathcal{P}) = |w| + k \log |T|$ and $poly$ is a polynomial. For 1-session (T,k)-runs, the number of events

in each such run is at most the length of the protocol's body. Although the term $log|T|$ is not necessary to define the size of a 1-session (T,k)-bounded protocol, we will use the same protocol size as defined above just for the sake of uniformity with the results in [30].

The following technical lemma [26] will be very useful in estimating the time complexity of our algorithms.

Lemma 16 ([26]): Let $\mathcal{P}=(\mathcal{S},\mathcal{C},w)$ be a (T,k)-bounded protocol, s be the last state of some run of length n of $\mathcal{P}$, A an agent, t be a message of length at most k over T, and φ a fact whose terms have length at most k. Then,

1) it is decidable in $\mathcal{O}(nk^2)$ time whether t is derivable from $s_{A,m}$ (i.e., $t\in\overline{s_{A,m}}$);
2) it is decidable in $\mathcal{O}(n^3k^6)$ time whether $\varphi\in Analz(s_A)$;
3) it is decidable in $\mathcal{O}(n^3k^6|w|)$ time whether $\varphi\in Analz(s_B)$ for some agent B.

The state space of a bounded security protocol is finite and so we are able to decide whether an action $act(x)$ is anonymous within some group G w.r.t. some observer X. An obvious algorithm for deciding this would search the state space twice: first, the algorithm detects a state s with $s_X\models act(x)$ and then, for each $y\in G$, the algorithm searches for a state s' with $s'\sim^X x$ and $s'_X\models act(y)$. As the number of events of a bounded security protocol is exponential w.r.t. the size of the protocol [30], this algorithm has a very high time complexity (w.r.t. the size of the protocol).

If we restrict the group anonymity problem to basic-term actions (Section IV-A) then Proposition 4 shows that only one search through the state space would suffice.

Theorem 17: $GAP(\tau)$ and $GAP_I(\tau)$ are in $NEXPTIME$ for any τ if they are restricted to basic-term actions of type τ and bounded security protocols. Moreover, except for $GAP_I(\tau)$ where τ is a rec-action type, all the other group anonymity problems restricted as above are complete for $NEXPTIME$.

If we restrict more bounded security protocols by allowing only 1-session runs, then we obtain the following results.

Theorem 18: $GAP(\tau)$ and $GAP_I(\tau)$ are in NP for any τ if they are restricted to basic-term actions of type τ and 1-session bounded security protocols. Moreover, except for $GAP_I(\tau)$ where τ is a rec-action type, all the other group anonymity problems restricted as above are complete for NP.

VII. Conclusions

Employing the epistemic logic framework developed in [24], [26], this paper proposes an approach to group anonymity in security protocols. This is formulated with respect to an honest agent and with respect to the intruder. A large spectrum of relationships between, and properties of, anonymity concepts were provided. One of the most interesting properties states that a group anonymous action in a security protocol under a passive intruder might not be group anonymous in the same security protocol if the intruder is active, and vice-versa.

It is shown that group anonymity is undecidable in unrestricted security protocols. Clearly, it becomes decidable in bounded security protocols. More precisely, group anonymity is complete for NEXPTIME if it restricted to basic-term actions and bounded security protocols, and it is complete for NP if it restricted to basic-term actions and 1-session bounded security protocols. These results show how difficult is to prove group anonymity for bounded security protocols. In practice, one has to design decision tools for group anonymity to work on restricted classes of protocols in order to obtain feasible results. We are not aware of any approaches for "practical classes of security protocols". However, there are tools such as MCMAS [33], [34] and PRISM [35] capable to check epistemic formulas on "not very complex" security protocols met in practice.

References

[1] D. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84–88, Feb 1981. doi: 10.1145/358549.358563. [Online]. Available: http://dx.doi.org/10.1145/358549.358563

[2] ——, "Security without identification: Transaction systems to make big brother obsolete," *Communications of the ACM*, vol. 28, no. 10, pp. 1030–1044, Oct 1985. doi: 10.1145/4372.4373. [Online]. Available: http://dx.doi.org/10.1145/4372.4373

[3] ——, "The dining cryptographers problem: Unconditional sender untraceability," *Journal of Cryptology*, vol. 1, no. 1, pp. 65–76, 1988. doi: 10.1007/BF00206326. [Online]. Available: http://dx.doi.org/10.1007/BF00206326

[4] S. Schneider and A. Sidiropoulos, "CSP and anonymity." in *4th European Symposium on Research in Computer Security (ESORICS'96)*, ser. Lecture Notes in Computer Science, E. Bertino, H. Kurth, G. Martella, and E. Montolivo, Eds., no. 1146, 1996. doi: 10.1007/3-540-61770-1_38 pp. 198–218. [Online]. Available: http://dx.doi.org/10.1007/3-540-61770-1_38

[5] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for web transactions," *ACM Trans. Inf. Syst. Secur.*, vol. 1, pp. 66–92, Nov 1998. doi: 10.1145/290163.290168. [Online]. Available: http://dx.doi.org/10.1145/290163.290168

[6] P. F. Syverson and S. G. Stubblebine, "Group principals and the formalization of anonymity," in *World Congress on Formal Methods'99*, 1999. doi: 10.1007/3-540-48119-2_45 pp. 814–833. [Online]. Available: http://dx.doi.org/10.1007/3-540-48119-2_45

[7] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," *In 13th USENIX Security Symposium*, 2004.

[8] D. Hughes and V. Shmatikov, "Information hiding, anonymity and privacy: A modular approach," *Journal of Computer Security*, vol. 12, pp. 3–36, Jan 2004. doi: 10.3233/JCS-2004-12102. [Online]. Available: http://dx.doi.org/10.3233/JCS-2004-12102

[9] S. Mauw, J. Verschuren, and E. P. de Vink, "A formalization of anonymity and onion routing," in *In Proceedings of the 9th European Symposium on Research in Computer Security (ESORICS 2004)*. Springer, 2004. doi: 10.1007/978-3-540-30108-0_7 pp. 109–124. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30108-0_7

[10] F. D. Garcia, I. Hasuo, W. Pieters, and P. van Rossum, "Provable anonymity," in *Proceedings of the 2005 ACM Workshop on Formal Methods in Security Engineering (FMSE'05)*, 2005. doi: 10.1145/1103576.1103585 pp. 63–72. [Online]. Available: http://dx.doi.org/10.1145/1103576.1103585

[11] J. Y. Halpern and K. R. O'Neill, "Anonymity and information hiding in multi-agent systems," *Journal of Computer Security*, vol. 13, no. 3, pp. 483–514, 2005. doi: 10.1109/CSFW.2003.1212706. [Online]. Available: http://dx.doi.org/10.1109/CSFW.2003.1212706

[12] K. Mano, Y. Kawabe, H. Sakurada, and Y. Tsukada, "Role interchangeability and verification of electronic voting," in *The 2006 Symposium on Cryptography and Information Security*, Hiroshima, Japan, Jan 2006.

[13] T. Chothia, S. Orzan, J. Pang, and M. T. Dashti, "A framework for automatically checking anonymity with μcrl," in *In Proceedings TGC06, LNCS*, 2007. doi: 10.1007/978-3-540-75336-0_19 pp. 301–318. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-75336-0_19

[14] J. Feigenbaum, A. Johnson, and P. Syverson, "A model of onion routing with provable anonymity," in *In Proceedings of the 11th Financial Cryptography and Data Security Conference*. Springer-Verlag, 2007. doi: 10.1007/978-3-540-77366-5_9 pp. 57–71. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-77366-5_9

[15] Y. Kawabe, K. Mano, H. Sakurada, and Y. Tsukada, "Theorem-proving anonymity of infinite-state systems," *Inf. Process. Lett.*, vol. 101, pp. 46–51, Jan 2007. doi: 10.1016/j.ipl.2006.06.016. [Online]. Available: http://dx.doi.org/10.1016/j.ipl.2006.06.016

[16] X. Sun, H. Wang, and J. Li, "On the complexity of restricted k-anonymity problem," in *Proceedings of the 10th Asia-Pacific Web Conference on Progress in WWW Research and Development*, ser. APWeb'08. Berlin, Heidelberg: Springer-Verlag, 2008. ISBN 3-540-78848-4, 978-3-540-78848-5 pp. 287–296.

[17] M. Edman and B. Yener, "On anonymity in an electronic society: A survey of anonymous communication systems," *ACM Computing Surveys*, vol. 42, no. 1, 2009. doi: 10.1145/1592451.1592456. [Online]. Available: http://dx.doi.org/10.1145/1592451.1592456

[18] J. F. Groote and S. Orzan, "Parameterised anonymity," in *Formal Aspects in Security and Trust*, P. Degano, J. Guttman, and F. Martinelli, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 177–191. ISBN 978-3-642-01464-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-01465-9_12

[19] H. Comon-Lundh, Y. Kawamoto, and H. Sakurada, "Computational and symbolic anonymity in an unbounded network," *JSIAM Letters*, vol. 1, pp. 28–31, 2009. doi: 10.14495/jsiaml.1.28. [Online]. Available: http://dx.doi.org/10.14495/jsiaml.1.28

[20] Y. Tsukada, K. Mano, H. Sakurada, and Y. Kawabe, "Anonymity, privacy, onymity, and identity: A modal logic approach," in *2009 International Conference on Computational Science and Engineering*, 2009. doi: 10.1109/CSE.2009.251 pp. 42–51. [Online]. Available: http://dx.doi.org/0.1109/CSE.2009.251

[21] C. A. Ardagna, S. Jajodia, P. Samarati, and A. Stavrou, "Providing mobile users' anonymity in hybrid networks," in *European Symposium on Research in Computer Security (ESORICS 2010)*, ser. Lecture Notes in Computer Science, D. Gritzalis, B. Preneel, and T. Theoharidou, Eds., vol. 6345, 2010. doi: 10.1007/978-3-642-15497-3_33 pp. 540–557. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15497-3_33

[22] M. Backes, G. Doychev, M. Dürmuth, and B. Köpf, "Speaker recognition in encrypted voice streams," in *European Symposium on Research in Computer Security (ESORICS 2010)*, ser. Lecture Notes in Computer Science, D. Gritzalis, B. Preneel, and T. Theoharidou, Eds., vol. 6345, 2010. doi: 10.1007/978-3-642-15497-3_31 pp. 508–523. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15497-3_31

[23] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf, Aug 2010.

[24] F. L. Ţiplea, L. Vamanu, and C. Vârlan, "Complexity of anonymity for security protocols," in *European Symposium on Research in Computer Security (ESORICS 2010)*, ser. Lecture Notes in Computer Science, D. Gritzalis, B. Preneel, and T. Theoharidou, Eds., vol. 6345, 2010. doi: 10.1007/978-3-642-15497-3_34 pp. 558–572. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15497-3_34

[25] I. Goriac, "An epistemic logic based framework for reasoning about information hiding," in *Availability, Reliability and Security, International Conference on*. Los Alamitos, CA, USA: IEEE Computer Society, 2011. doi: 10.1109/ARES.2011.49 pp. 286–293. [Online]. Available: http://dx.doi.org/10.1109/ARES.2011.49

[26] F. L. Ţiplea, L. Vamanu, and C. Vârlan, "Reasoning about minimal anonymity in security protocols," *Future Generation Computer Systems*, vol. 29, pp. 828–842, March 2013. doi: 10.1016/j.future.2012.02.001. [Online]. Available: http://dx.doi.org/10.1016/j.future.2012.02.001

[27] S. Kramer, "Cryptographic protocol logic: Satisfaction for (timed) Dolev-Yao cryptography," *The Journal of Logic and Algebraic Programming*, vol. 77, pp. 60–91, Sep 2008. doi: 10.1016/j.jlap.2008.05.005. [Online]. Available: http://dx.doi.org/10.1016/j.jlap.2008.05.005

[28] R. Ramanujam and S. P. Suresh, "A decidable subclass of unbounded security protocols," in *Workshop on Issues in the Theory of Security (WITS'03)*, 2003, pp. 11–20.

[29] F. L. Ţiplea, C. Enea, and C. V. Bîrjoveanu, "Decidability and complexity results for security protocols," in *Verfication of Infinite-state Systems with Applications to Security (VISSAS 2005)*, E. Clarke, M. Minea, and F. Tiplea, Eds. IOS Press, 2005, pp. 185–211.

[30] F. L. Ţiplea, C. Enea, C. V. Bîrjoveanu, and I. Boureanu, "Secrecy for bounded protocols with freshness check is NEXPTIME-complete," *Journal of Computer Security*, vol. 16, pp. 689–712, Dec 2008. doi: 10.3233/JCS-2007-0306. [Online]. Available: http://dx.doi.org/10.3233/JCS-2007-0306

[31] D. Dolev and A. Yao, "On the security of public-key protocols," *IEEE Transactions on Information Theory*, vol. 29, no. 2, pp. 198–208, 1983. doi: 10.1109/TIT.1983.1056650. [Online]. Available: http://dx.doi.org/10.1109/TIT.1983.1056650

[32] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi, *Reasoning About Knowledge*. The MIT Press, 2003.

[33] A. Lomuscio, H. Qu, and F. Raimondi, "MCMAS: A model checker for the verification of multi-agent systems," in *Computer Aided Verification*, ser. Lecture Notes in Computer Science, A. Bouajjani and O. Maler, Eds. Springer Berlin Heidelberg, 2009, pp. 682–688. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02658-4_55

[34] ——, "MCMAS: An open-source model checker for the verification of multi-agent systems," *International Journal on Software Tools for Technology Transfer*, pp. 1–22, 2015. doi: 10.1007/s10009-015-0378-x. [Online]. Available: http://dx.doi.org/10.1007/s10009-015-0378-x

[35] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM 4.0: Verification of probabilistic real-time systems," in *Proceedings of the 23rd International Conference on Computer Aided Verification*. Springer-Verlag, 2011. doi: 10.1007/978-3-642-22110-1_47 pp. 585–591. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-22110-1_47

3rd Workshop on Constraint Programming and Operation Research Applications

THE aim of the CPORA-Workshop on Constraint Programming and Operation Research Applications is to bring together interested researchers from constraint programming/constraint logic programming (CP/CLP), operations research (OR) and artificial intelligence (AI) to present new techniques or new applications in decision support, combinatorial optimization, modeling and control processes arising in manufacturing, transportation, telecommunication, computer networks, logistic systems etc. and to provide an opportunity for researchers in one area to learn about techniques in the others. The aim of this workshop is share ideas, projects, researches results, models, experiences etc. associated with CP/CLP/OR/AI and to give researchers the opportunity to show how the integration of techniques from different fields can lead to interesting results on large and complex problems. Additionally, we would like to stimulate the communication between researchers working on different fields and practitioners who need reliable and efficient modelling and computational methods for industrial and business processes.

Contributions containing of both: the theoretical and practical results obtained in this area are welcome.

TOPICS

- Constraint programming/Constraint logic programming,
- Mathematical programming,
- Constraint Satisfaction Problem,
- Logic programming,
- Hybrid methods,
- Network programming,
- Petri-Nets,
- Knowledge methods,
- Soft computing (FL, GA, NN etc.),
- Answer Set Programming (ASP),
- The boolean satisfiability problem (SAT).
- Manufacturing,
- Multimodal processes management,
- Project management,
- Supply chain management,
- Modeling and planning production flow,
- Production scheduling,
- Multimodal social networks,
- Intelligent transport and passenger routing,
- Network knowledge modeling,
- Transportation networks.

EVENT CHAIRS

- **Bocewicz, Grzegorz,** Koszalin University of Technology, Poland
- **Sitek, Pawel,** Kielce University of Technology, Poland

PROGRAM COMMITTEE

- **Banaszak, Zbigniew,** Warsaw University of Technology, Poland
- **Burduk, Anna,** Wrocław University of Science and Technology, Poland
- **Bzdyra, Krzysztof,** Koszalin University of Technology, Poland
- **Gola, Arkadiusz,** Lublin University of Technology, Poland
- **Janardhanan, Mukund Nilakantan,** University of Leicester, United Kingdom
- **Nielsen, Peter,** Aalborg University, Denmark
- **Nielsen, Izabela Ewa,** Aalborg University, Denmark
- **Ratnayake, Chandima,** University of Stavanger, Norway
- **Terkaj, Walter,** ITIA-CNR, Italy
- **Turkyilmaz, Ali,** NAZARBAYEV UNIVERSITY, Kazakhstan
- **Wikarek, Jarosław,** Kielce University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 419–424
DOI: 10.15439/2018F264
ISSN 2300-5963 ACSIS, Vol. 15

Visualization of logical formulas

Radosław Klimek
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Krakow, Poland
Email: rklimek@agh.edu.pl

***Abstract*—Visualization of logical formulas for classical propositional calculus gains a new meaning in the context of a huge development of SAT solvers which find solutions of satisfiability problems of bigger and more complex tasks including the ones of industrial meaning. The aim of this work is to create, using modern IT tools, a coherent and widely accessible system in a form of web application, which will be able to provide a coopertive system, as well as will help to visualize, analyze and transform logical formulas. The system itself is now in the initial, however, mature implementation phase. It is open for new ideas and methods. New functionalities have already been introduced but there are plans to create the new ones.**

***Index Terms*—SAT; visualization; graph.**

I. Introduction

SATISFIABILITY problem is the classical and fundamental problem of theoretical computer science in which there can be encoded many other problems [1]. Satisfiability problem is a testing whether exists of a logical value assignment (true/1, false/0) to variables in a particular formula which will satisfy it, namely its logical value will be true. Those types of problems can be applied to many branches and are commonly used, for example in combinatorial optimization [2], graph theory, automated theorem proving, verification of software models [3], [4], [5], [6], and artificial/ambient intelligence [7], [8], [9].

Tools which solve satisfiability problem are called SAT solvers. Their rapid development and routine solving of tasks, which consist of many thousands of variables and clauses, enables practical implementation of reasoning engines which are quite common and easily accessible [10], [11]. There was created a standard of input files recording, known as DIMACS CNF format or simply DIMACS, where the base DIMACS relates to the name of the research center where it was created, namely The Center of Discrete Mathematics and Theoretical Computer Science, and CNF is a reference to the conjunctive form of the normal formula (Conjunctive Normal Form). This form is used to save logical formula in leading solvers based on CDCL method where problem is enclosed in this form in a particular file having DIMACS form. Thanks to using one format it is easy to compare existing SAT solvers and to see new possibilities of creating new, interesting tools based on the same input files which can be related to SAT or similar concepts.

Nowadays, SAT solvers are commonly used in a formal verification of designed systems, especially embedded systems, cyber-physical systems and software drivers. Solvers become more and more popular. Their widespread use is visible in the processes described as Electronic Design Automation (EDA) because designing processes of those sets, especially embedded sets, need to be analyzed carefully as the errors done during their design and production can have detrimental effects and be very costly. The similar situation appears when we talk about drivers which unstable work can interfere with the working of the system core.

Visualization of logical formula can also be a useful, and to some extend helpful, tool in solving satisfiability problem as well as in analysis and preprocessing of formulas as input data for SAT solvers. The graphic form of formula, its shape, regularity or irregularity, consistency or inconsistency can provide us with many valuable pieces of information about the encoded problem, see for example [9, Table 11]. It can also suggest recommended formula preprocessing methods which can speed up the process of searching for a satisfiable solution. That is why visualization helps us to understand the encoded problem better and gives us an opportunity of carrying an additional and quicker analysis of the problem before our attempt to solve it. Last but not least, visual image of a formula can have an aesthetic meaning. The generated visualizations create interesting visual effects which may be attractive for people who have not dealt with SAT topic so far.

The aim of this work is to create widely accessible service in form of web application, built on the basis of modern IT tools, easy to handle and enabling the visualization of logical formulas of propositional calculus which are the input data for SAT solvers based on CDCL method (Conflict Driven Clause Learning). The present system is based on work [12], however, it has been adjusted to the web requirements[1].

The development of solvers which work on the basis of this method needs to be acknowledged as a spectacular one. Every formula can be visualized and analyzed before we start looking for its solution. Although in this work there were presented the well-known visualization methods but there also were proposed new functionalities of the system. The existing program called *DPvis* created by Carsten Sinz, see seminal and fundamental work [13], enables crating different types of graphs on the basis of formulas describing SAT problems. However, this program is hardly accessible and not compatible with demands of the modern and dynamic IT market which can discourage the present and potential users. Under those circumstances, creating the fully web application, see [12], which enables cooperative work in client-server

[1]The planned website for this service is http://forvis.agh.edu.pl.

architecture, built on the basis of modern IT tools and offering SAT problems visualizations, may encourage new users or specialists and provide them with tools which simplify their everyday work.

In the future the service will be extended to provide fluent transition between different methods of formula visualizations or different methods of encoding logical formulas. There are also plans to open the service for related problems according to classical SAT, namely *MaxSAT* (finding the maximal number of clauses, if not all, of the entire formula to be satisfied) and *weighted MaxSAT* (finding an assignment that maximizes the total weight of the satisfied clauses), *Partial MaxSAT*(some clauses are deemed hardinfinite weights, clauses with finite weight are soft), and other problems, which will enable the analysis of fragmentary problem solutions in the context of the whole problem.

II. Preliminaries

Supported file format is a commonly recognized DIMACS CNF format. System should offer a possibility of storing saved files and exporting created visualizations to graphic files within a wide range of formats.

Logical formula F is in CNF form because it is built from conjunction of clauses, which can be presented as $F = \bigwedge_{i=1}^{m} C_i$ that is a conjunction of clauses C_1, C_2, ..., C_m and every clause is built from disjunction of literals $C_i = \bigvee_{j=1}^{n} l_i, j$ where literal l can be a variable x or its negation $\neg x$. The example of logical formula in CNF form is the formula:

$$(x \vee \neg y) \wedge (z \vee y \vee \neg x)$$

DIMACS CNF file is saved as ASCII text file which enables easy transmission between different operational systems or working environments. The file can have lines of comments, so-called line of problem, but its most important part are the following clauses, whereby every of them finishes with number 0. Every clause is built from a sequence of natural numbers where every number clearly identifies one variable (can be from thesaurus of variables). Moreover, every number may be predeceased by minus or not be predeceased, where symbol of minus signifies negation of this variable. Therefore, the particular numbers signify literals. The following file is an example of that:

```
c
p cnf 212 118098
-169 -177 -184 -182 -174 -187 -196 -201 -199
-191 -17 -26 -32 -28 -23 -35 -41 -48 -46 -38
-53 -60 -65 -62 -56 -68 -77 -83 -79 -73 -86
-93 -100 -97 -89 -103 -109 -117 -113 -108
-122 -126 -132 -129 -125 -139 -144 -151 -146
-142 -154 -162 -166 -164 -157 12 13 14 15 0
-169 -177 -184 -182 -174 -187 -196 -201 -199
-191 -17 -26 -32 -28 -23 -35 -41 -48 -46 -38
-53 -60 -65 -62 -56 -68 -77 -83 -79 -73 -86
-93 -100 -97 -89 -103 -109 -117 -113 -108
-122 -126 -132 -129 -125 -139 -144 -151 -146
-142 -154 -162 -166 -164 -157 12 13 14 15
170 171 172 173 178 179 184 185
181 182 175 176 0
```

The size of logical formula within a file does not prove difficulty of the problem itself but, of course, in practice, the time spent on searching for the solutions of the big formulas is longer, sometimes very long. There are tools which aim is to minimize logical formulas which can reduce time of solving the problem significantly. One of the better known tools is *SatELite* program which implements the following techniques [14]:

1) elimination of variables by resolutions – deleting unnecessary variables and equivalent literals;
2) fast subsumption – C_1 clause subsumes C_2 when $C_1 \subseteq C_2$, therefore C_1 may be deleted;
3) self-subsumption – C_1 clause almost fully subsumes C_2 with exception of one literal x which appears in C_2 with a different value. For example, $C_1 = \{x, a, b\}$, and $C_2 = \{\neg x, a\}$. Resolution for x gives $C_1' = \{a, b\}$, which subsumes C_1, therefore C_1 can be added to logical formula and C_1 can be removed.

III. Visualization methods

There are used three basic methods of visualization in this work . It does not cover all planned implementation methods. The present set should be interpreted as an initial proposition for system functionalities. The currently used methods are [13]:

- factor graph,
- interaction graph,
- resolution graph.

Undirected factor graph $G_F = (V, E)$ is a graph where a set of vertices V is created by variables and clauses, therefore $V = X \cup C$, where X is a set of variables and C is a set of clauses. E symbol means a set of edges. The edge is drawn between variable x and clause c when $x \in C$ or $\neg x \in C$. Provided that there exists a problem described by formula: $(\neg x \vee \neg y \vee z) \wedge (y \vee z) \wedge (x \vee \neg z \vee u)$ it is possible to determine the following clauses:

1) $C_1 = (x \vee \neg y)$
2) $C_2 = (\neg x \vee z \vee u)$
3) $C_3 = (x \vee z)$

Construction of factor graph for the problem presented above should start from drawing variables and clauses in form of vertices, see Fig. 1. Later on we draw edges which connect variables with clauses in which they appear, starting from clause C_1. Finally, the graph looks like the last structure in Fig. 1.

In contrast to undirected factor graph, directed factor graph includes the difference between a variable in positive and negative form. It is usually presented using colors, where green means not negated variable and red a negated one. For the analyzed problem, the graph looks like the following one: Fig. 2.

Variable interaction graph $G_I = (V, E)$ is a graph in which a set of vertices V is equal to a set of variables X which is equivalent to $V = X$. E means a set of edges. The edge

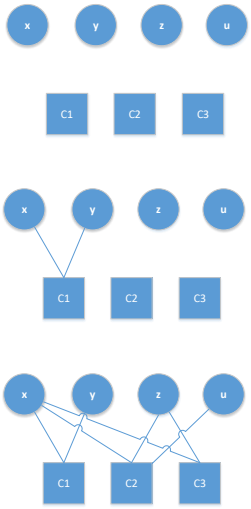


Fig. 1. Structure of a factor graph, and successive construction steps

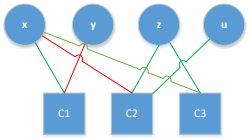


Fig. 2. Directed factor graph

connects two variables: x and y when they appear together in at least one clause ($x \in C_i$ and $y \in C_i$). On the basis of the example from the previous section, there are determined the following clauses:

1) $C_1 = (x \vee \neg y)$
2) $C_2 = (\neg x \vee z \vee u)$
3) $C_3 = (x \vee z)$

The first step to create interaction graph is drawing variables in the form of vertices, see Fig. 3. Later on, starting from C_1 clause, the appearing variables are connected together. When analyzing next clauses, if the particular edge already exists for a pair of variables, it is omitted. Finally, the graph looks like the last structure in Fig. 3.

Resolution graph $G_R = (V, E)$ is a graph in which a set of vertices V is equal to a set of clauses, therefore $V = C$. E means a set of edges. The edge connects two variables C_1 and C_2 only if there is a variable x where $x \in X$, in which X means a set of variables where $x \in C_1$ and $\neg x \in C_2$. Using the previously presented formula and having already determined clauses:

1) $C_1 = (x \vee \neg y)$
2) $C_2 = (\neg x \vee z \vee u)$
3) $C_3 = (x \vee z)$

Construction of the resolution graph starts from drawing clauses in form of vertices, see Fig. 4. Later on the following variables which appear in both C_1 and C_2 clauses are analyzed. Because $x \in C_1$ and $x \in C_2$ there can be drawn an

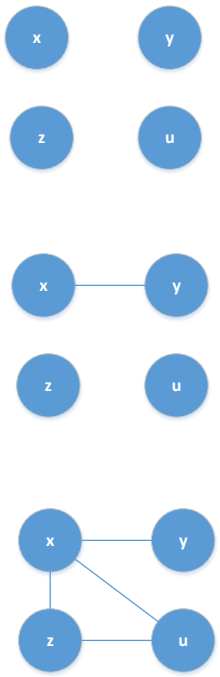


Fig. 3. Structure of interaction graph, and the succseeding, and successive construction steps

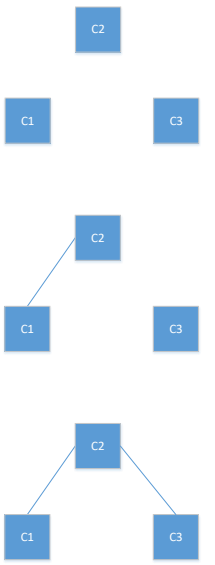


Fig. 4. Structure of resolution graph, and the succseeding, and successive construction steps

edge between the clauses. The redundant edges are omitted. Finally, the graph looks like the last structure in Fig. 4.

The methods of visualization presented above are the basic set of methods. In the future there are plans to introduce the new ones taking into consideration needs and nature of SAT-related problems.

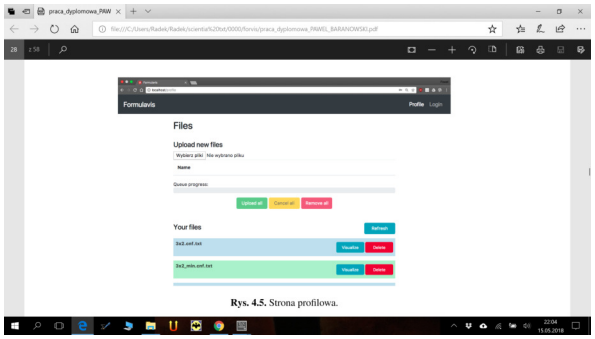


Fig. 5. The screenshot of the system – formula management

IV. Design of visualization system

The implemented system, see Fig. 5, is breifly described. Server application was made using *Python* language together with frame of *Django REST Framework* application [15]. For operation of asynchronous tasks, *Celery* task queue [16], together with *RabbitMQ* [17] message broker, are responsible. Data is stored in *PostgreSQL* database server. Client application, as well as user's interface, was created on the basis of *TypeScript* language using the frame of *Angular 4* application. There was used an effective visualization library based on *JavaScript* language named *vis.js* [18]. Server proxy *Nginx* [19] is responsible for managing the movement between user's graphic interface and server application. The whole project is containerized using *Docker* technology [20]. *Nginx* plays the role of Reverse-Proxy or Forward-Proxy server, see Fig. 6. Reverse-Proxy server helps in:

- hiding the current system which is behind proxy server,
- distributing the movement between application instances of the server,
- pointing the movement towards proper applications,
- compressing the content of data flow,
- manipulating with requests and answers.

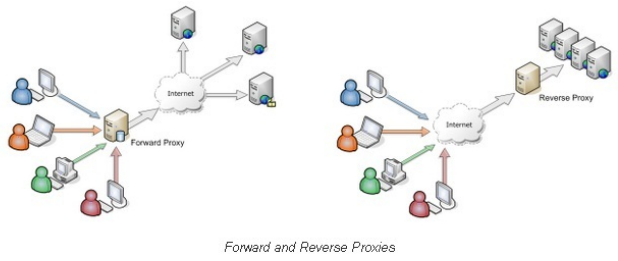


Fig. 6. Difference between Forward-Proxy i Reverse-Proxy server. Source: Vivek Srivastav: Proxy Pass. [In:] http://viveksrivastv.blogspot.com/p/apache-administration.html. Accessed on 1 May 2018.

Celery is a asynchronous task queue based on task distribution using messages. It concentrates on operations performed in the shortest possible time but also it supports planned tasks. Tasks are performed at the same time using one or many executive instances – workers and with the use of multiprocess transformation. Those tasks can be performed asynchronously (it enables the ordering application to perform further work) or synchronically (the ordering application waits for the result). In order to use *Celery*, it is required to use message broker which provides data for workers. The message broker recommended by *Celery* authors is *RabbitMQ*. It is an open source and easy to implement system in *Erlang* language. It supports and monitors asynchronous message sending and their deployment using clusters which enables further development of the system.

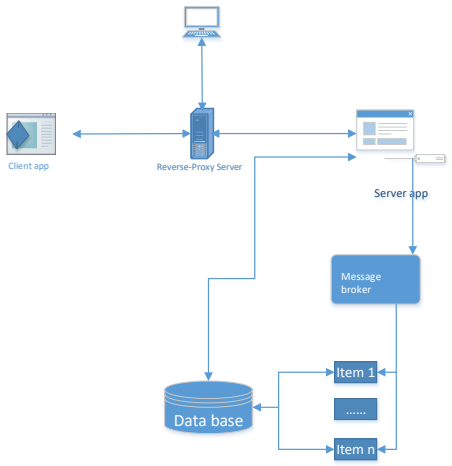


Fig. 7. Structure of logical formula visualization system

The created client application is an interface of the whole system because a target user has an access to it. Thanks to that, the user can log in and log out, upload and delete data from server and, most importantly, display visualizations and save them to file. Implemented application is in the form of browser application. Server application, implemented separately, organizes the whole data processing according to user's expectations. Detailed analysis of this issue exceeds the aim and size of this work, see also Fig. 7.

V. Examples of visualization

There were examined many formulas in terms of possibilities for system work as well as visualization effects. All figures (from 8 to 13) are provided by the designed system, see also [12]. Presentation starts from simple formulas and continues to the most complicated ones. We will start from a very simple example, see also Fig. 8:

```
p cnf 3 2
1 -3 0
2 3 -1 0
```

In interaction graph:

- blue vertices – represent variables,
- edge, shade of edge signifies number of connections between variables (the more of them, the darker the shade is). Interaction graph presents the structure of variables neighborhood within the examined logical formula. On the basis of that there can be performed the analysis if a

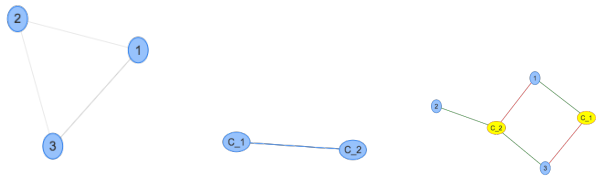


Fig. 8. First example: interaction, resolution and factor graphs

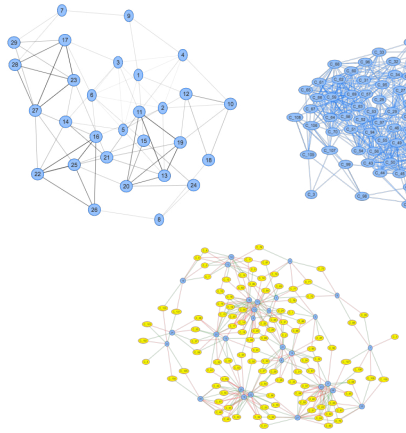


Fig. 9. Second example, interaction, resolution and factor graphs

problem is integral or if it can be divided into parts. What is more, the edges brightness can help to find frequent neighborhoods of variables.

In resolution graph:

- blue vertices – represent clauses.

Resolution graph visualizes the structure of clause dependencies. Vertices are connected by edge if they have one (or more) literals of a different logical value. In directed factor graph:

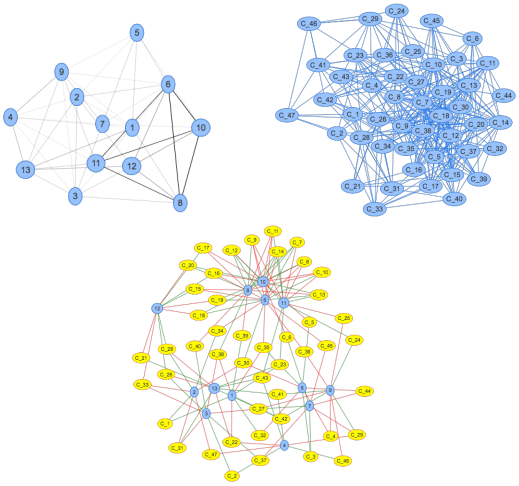


Fig. 10. Second example, after minimalization: interaction, resolution and factor graphs

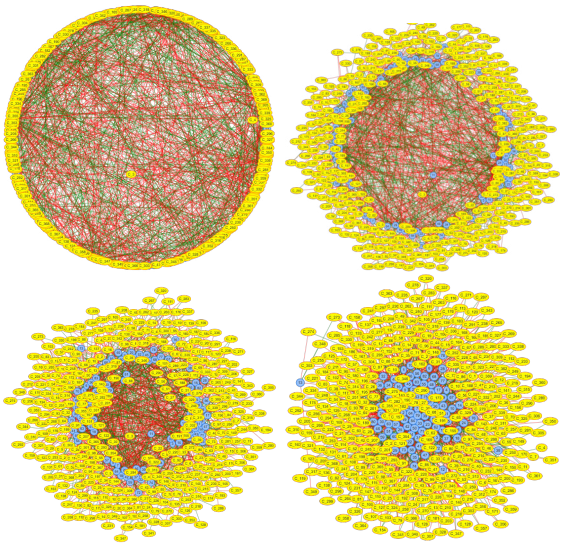


Fig. 11. Third example: stabilization, factor graphs – before stabilization, first screenshot, second screenshot, after stabilization

- yellow vertices – represent clauses,
- blue vertices – represent variables,
- red edge – represents a variable with negation,
- green edge – represents a variable without negation.

Factor graph shows clause dependencies and variables with their logical values.

The next analyzed formula has 29 variables and 109 clauses, and is shown in Fig. 9. After preprocessing, the minimization was performed which gave 13 variables and 47 clauses, see Fig. 10.

In a created system there was implemented graph/formula *stabilization* operation. It means, that the graph was rebuilt in such a way that the edges got a total minimal length. Stabiliza-

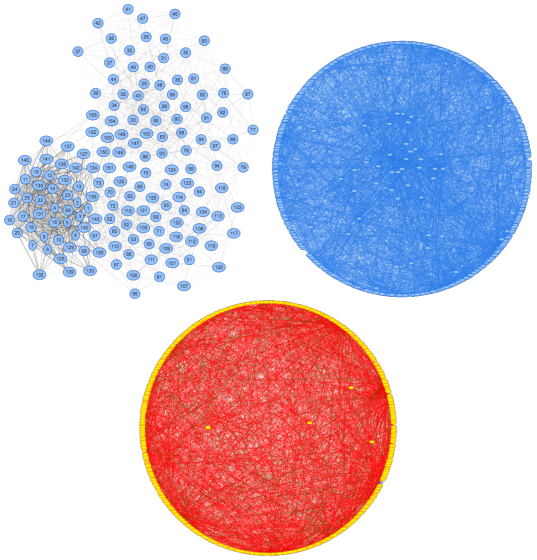


Fig. 12. Fourth example: interaction, resolution and factor graphs

tion helps to get more compact visualizations. The stabilization process can be observed in the case of logical formula of 83 variables and 369 clauses. The next images present the following visualization stages, until its accomplishment, see Fig. 11.

The fourth example: logical formula – 155 variables, 1135 clauses, see Fig. 12.

And the last example: logical formula after minimalization – 42 variables, 133 clauses, see Fig. 13.

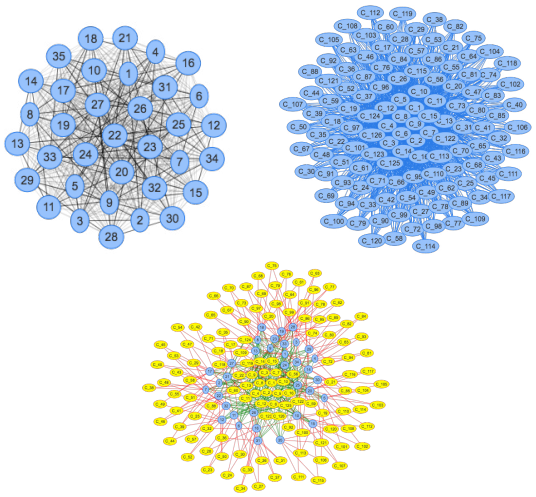


Fig. 13. The last example: interaction, resolution and factor graphs

VI. Conclusion

Visualization project of logical formulas has to be considered as a successful one, however, it is now at its initial stage and it will be continued. The system enables the elementary transformation of formulas and their preprocessing. The aim of the whole project was to create widely accessible tool helping to understand, analyze and examine the structures of problems presented with the use of logical formulas.

The system also offers storing files on the server which requires creating log-in system and the user's profile. It also offers the possibility of interaction with generated graphs, minimization of uploaded formulas and export of visualizations to the selected graphic formats. Its design and implementation aspects are very modern as a result of using the open approach as well as modern software technologies.

There are plans of a further system development by introducing new functionalities, another visualization methods and opening it for other SAT problems, for example MaxSAT.

References

[1] A. Biere, M. Heule, H. van Maaren, and T. Walsh, *Handbook of Satisfiability: Volume 185 Frontiers in Artificial Intelligence and Applications*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2009.

[2] E. Kucharska, K. Grobler-Debska, and K. Raczka, "ALMM-based methods for optimization makespan flow-shop problem with defects," in *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology ISAT 2016 - Part I, Karpacz, Poland, September 18-20, 2016*, 2016, pp. 41–53. [Online]. Available: https://doi.org/10.1007/978-3-319-46583-8_4

[3] R. Klimek, "Towards formal and deduction-based analysis of business models for SOA processes," in *Proceedings of 4th International Conference on Agents and Artificial Intelligence (ICAART 2012), 6–8 February, 2012, Vilamoura, Algarve, Portugal*, J. Filipe and A. Fred, Eds., vol. 2. SciTePress, 2012, pp. 325–330.

[4] R. Klimek and P. Szwed, "Verification of archimate process specifications based on deductive temporal reasoning," in *Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS 2013), 8–11 September 2013, Kraków, Poland*. IEEE Xplore Digital Library, 2013, pp. 1131–1138.

[5] R. Klimek, "From extraction of logical specifications to deduction-based formal verification of requirements models," in *Proceedings of 11th International Conference on Software Engineering and Formal Methods (SEFM 2013), 25–27 September 2013, Madrid, Spain*, ser. Lecture Notes in Computer Science, R. M. Hierons, M. G. Merayo, and M. Bravetti, Eds., vol. 8137. Springer Verlag, 2013, pp. 61–75.

[6] P. Wiśniewski, K. Kluza, A. Ligeza, and A. Suchenia, "Generation of synthetic business process traces using constraint programming," in *Proceedings of Federated Conference on Computer Science and Information Systems (FedCSIS 2018), 9–12 September 2018, Poznań, Poland*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. IEEE Xplore Digital Library, 2018, pp. 441–449.

[7] R. Klimek, "Behaviour recognition and analysis in smart environments for context-aware applications," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015), October 9–12, 2015, City University of Hong Kong, Hong Kong*. IEEE Computer Society, 2015, pp. 1949–1955.

[8] R. Klimek and L. Kotulski, "Towards a better understanding and behavior recognition of inhabitants in smart cities. a public transport case," in *Proceedings of 14th International Conference on Arificial Inteligence and Soft Computing (ICAISC 2015), 14–18 June, 2015, Zakopane, Poland*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 9120. Springer Verlag, 2015, pp. 237–246.

[9] R. Klimek, "Exploration of human activities using message streaming brokers and automated logical reasoning for ambient-assisted services," *IEEE Access*, vol. 6, pp. 27 127–27 155, 2018.

[10] C. P. Gomes, H. A. Kautz, A. Sabharwal, and B. Selman, "Satisfiability solvers," in *Handbook of Knowledge Representation*, ser. Foundations of Artificial Intelligence, F. van Harmelen, V. Lifschitz, and B. W. Porter, Eds. Elsevier, 2008, vol. 3, pp. 89–134.

[11] D. E. Knuth, *The Art of Computer Programming, Volume 4, Fascicle 6: Satisfiability*, 1st ed. Addison-Wesley Professional, 2015.

[12] P. Baranowski, "System for visualization of logical formulas, Engineering diploma thesis, supervisor: Radosław Klimek, AGH University of Science and Technology," 2018.

[13] C. Sinz, "Visualizing SAT Instances and Runs of the DPLL Algorithm," *Journal of Automated Reasoning*, vol. 39, no. 2, pp. 219–243, Aug. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10817-007-9074-1

[14] N. Eén and A. Biere, "Effective preprocessing in sat through variable and clause elimination," in *Theory and Applications of Satisfiability Testing*, F. Bacchus and T. Walsh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 61–75.

[15] T. Christie, "Website: Django rest framework," 2018, accessed on 6-Jan-2018. [Online]. Available: http://www.django-rest-framework.org/

[16] Celery Development Team, "Website: Celery: Distributed task queue," 2018, accessed on 6-Jan-2018. [Online]. Available: http://www.celeryproject.org/

[17] Rabbit Technologies Ltd., "Website: RabbitMQ documentation," 2018, accessed on 7-Feb-2018. [Online]. Available: https://www.rabbitmq.com/documentation.html

[18] vis.js Development Team, "Website: vis.js. dynamic browser based visualization library," 2018, accessed on 6-Jan-2018. [Online]. Available: vhttp://visjs.org/

[19] NGINX Development Team, "Website: Nginx products," 2018, accessed on 6-Jan-2018. [Online]. Available: https://www.nginx.com/resources/wiki/

[20] Docker Development Team, "Website: Docker - software containeraization platform," 2018, accessed on 6-Jan-2018. [Online]. Available: http://www.docker.com/

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 425–435
DOI: 10.15439/2018F114
ISSN 2300-5963 ACSIS, Vol. 15

Siphon-based deadlock detection in Integrated Model of Distributed Systems (IMDS)

Wiktor B. Daszczuk
Institute of Computer Science,
Warsaw University of Technology,
Nowowiejska Str. 15/19, 00-665 Warsaw, Poland
e-mail wbd@ii.pw.edu.pl

***Abstract*—Integrated Model of Distributed Systems (IMDS) is a formalism for specification and verification of distributed systems, especially following IoT (Internet of Things) paradigm. The formalism emphasizes such features as asynchrony of actions and communication, locality of decisions, and autonomy in executing actions. In conjunction with model checking, IMDS allows to analyze such features of distributed systems as deadlocks or distributed termination. However, the nature of model checking allows to find one deadlock in a single run of the verifier, which produces a counterexample.**

The conversion of IMDS specification to a Petri net is used to identify multiple deadlocks in one verification, using siphons. Model checking is used to verify if a siphon can become empty, which denotes a true deadlock in a purely cyclic system, like FMS (Flexible Manufacturing Systems). The extension of the verification by temporal checking allows to cover systems with any structure: cyclic, terminating, or with a more complex scheme. In addition, the proposed procedure allows to easily identify processes participating in partial deadlocks. Two types of deadlock can be identified: communication deadlocks and resource deadlocks.

I. Introduction

IMDS (Integrated Model of Distributed Systems [1][2]) is a formalism for describing the behavior of distributed systems, especially for finding deadlocks. A system is modeled as a set of actions, having servers' states and agents' messages on input and on output. In IMDS, a communication dualism is exploited, since the modeled system is represented as server processes that communicate by messages, or alternatively by travelling processes (agents) that communicate by means of servers' states. A model of a distributed system is uniform (that is, it has a single form), but it can be decomposed ("cut") to a set of server processes or a set of agent processes. System actions are combined in sequences to form the processes. An action has a current server's state and an agent's current message on input, and it produces a similar pair (a new server's state and a new agent's message) on output.

The two views of a system (server view and agent view) are obtained by the two possible groupings of a set of all actions into sequences. In the server view, actions of an individual server are grouped into a process (the definition of processes is included in Sect. IIIB). The server's states are the carrier of the server process, and the messages are the communication means between server processes. In the agent view, actions concerning an individual agent conform a process. Messages are internal to a process: they are the carrier of the process. The agent processes communicate via servers' states.

The IMDS formalism was used, together with model checking technique [3], to develop the Dedan program which finds various kinds of deadlock in a verified system [4]. These are: communication deadlock (in the server view), resource deadlock (in the agent view), partial deadlock (in which a subset of system's processes participate) and total deadlock (concerning all processes). A counterexample is generated if a deadlock is found. A counterexample is a path leading from the start of the system to the deadlock.

In Dedan, automatic conversion between the server view and the agent view is performed. Also, observation of a global transition graph and simulation on this graph are possible.

Dedan is built in such a way that the specification of temporal formulas and temporal verification are hidden to a user. The reason is that model checking techniques are seldom known by the engineers. Therefore, the program is constructed in such a way that a user specifies a system and simply "pushes the button" to check for the existence of a deadlock.

The model checking technique has a disadvantage: the evaluation of temporal formula consists in finding a single global configuration (will be defined in Section IIIA) which causes the *false* result, which denotes a deadlock. A *counterexample* is a sequence leading from initial configuration to the deadlock. The designer may repair the erroneous specification and run the verification again. The scheme should be repeated multiple times, until all deadlocks are found and repaired.

The other technique of deadlock identification is finding siphons in a Petri net corresponding to a verified IMDS specification. A siphon is a Petri subnet, which cannot restore tokens if it is emptied [5][6][7]. If an empty siphon is

This work was not supported by any organization

reachable, it denotes a deadlock. The deadlock concerns the processes (server processes and/or agent processes) that take part in the siphon. Therefore, it may be total or partial deadlock. Siphon analysis can find multiple deadlocks in a system, because multiple siphons may exist in a net and the algorithms find all siphons in a single run [7].

Siphon-based deadlock detection is used in some purely cyclic classes of Petri nets, used to model FMS (Flexible Manufacturing Systems) [8][9]. However, many systems cannot be modeled as a cyclic Petri net. Some examples of such systems, for example an IoT (Internet of Things) distributed systems with multiple terminating processes, are mentioned in Section VI.

The previous paper [10] concerned identification of deadlocks in IMDS specifications which correspond to purely cyclic systems, like a class of FMS systems. The contribution of this paper is an application of siphon-based deadlock detection to systems of arbitrary schemes: cycling, terminating or intermediate (some processes are cyclic while other ones terminate). For this purpose an IMDS model is converted to a Petri net. Siphon detection is done in the Petri net, while identification of deadlocks and finding processes involved (partial deadlocks and total deadlocks are identified) is performed using reachability verification and temporal analysis in IMDS specification.

As a siphon may be emptied in different ways, thus it may lead to more than one deadlock. Model checking identifies one example of siphon emptying in a single run. Therefore our procedure does not guarantee identification of all deadlocks in a single run, one deadlock is found per reachable empty siphon. Still, a possibility of identification of multiple deadlocks (one for every emptied siphon) in a single procedure is a benefit. Additionally, the described procedure liberates from constraining siphon-based deadlock detection from purely cyclic systems only.

The described procedure gives a possibility of identification of multiple deadlocks in distributed systems specified in IMDS formalism, preserving communication duality, locality and autonomy of distributed components, and asynchrony of actions and communication.

In this paper, a background of static deadlock detection methods is given in Section II. A definition of IMDS is given in Section III. The definition is formulated differently from the paper [10], where a distributed system was defined using four basic sets: servers, state values, services and agents. The present definition is much easier for readers, because it uses two basic sets: states and messages. The previously used four sets are used in IMDS implementation, mentioned in Section IV, where an example of a bounded buffer is presented. The conversion of IMDS specification to a Petri net, and deadlock detection using siphons and reachability is described in Section V. Section VI presents the application of proposed method to systems with various structures, including acyclic and hybrid ones. An example of a not purely cyclic system containing deadlock siphons and no-deadlock siphons is described in Section VII. A practical example is presented in Section VIII. Section IX concludes the paper.

II. Related Work on Deadlock Detection

Many deadlock detection techniques are described in the literature. Dynamic methods typically use some kind of wait-for graph [11] to discover a deadlock (ant typically to prevent a deadlock or to escape from it).

Static methods use a model of a system and explore it to find deadlocks. Model checking techniques are based on temporal reachability space verification. The activities of the system are expressed in terms of local features of its components, and the global reachability space of the system is constructed. The features of system components are given as temporal formulas and verified by the evaluation of them. Model checkers are often equipped with automatic deadlock detection procedures. Typically, deadlock is identified as "a state with no future", i.e., a strongly connected subgraph containing one state only: the deadlock itself [12]. Deadlock freeness is checked by a CTL temporal formula **AG EX** *true* (for any state a next state exists) [13]. Yet, total termination seems to be analogous state: no future exists. In cyclic system, where termination is not expected, the above formula identifies a deadlock. In terminating systems a deadlock should be distinguished from termination.

Temporal formulas can also be used to check partial deadlocks, in which some processes are involved in a deadlock, but other processes continue their run. Generally, in the case of partial deadlock detection, temporal formulas are based on the structure of verified models to identify deadlocks in individual processes [14]. The disadvantage of such an approach is that temporal formulas need to be developed individually for each analyzed system, using its specific features.

Some other approaches to partial deadlock detection use temporal formulas that are not related to the structure of a verified model. However, such model-unrelated formulas require the system to have specific properties [15][16]. If a system is non-terminating (cycling), a discontinuation of a process is obviously a deadlock [3]. Conversely, a method [17] may be ascribed to terminating processes only. Some detection methods are used for specific architectures of systems. For example, WickedXmas approach uses nodes communicating by queues [18].

Other set of static methods concern Petri nets. Some of them are based on analysis of reachability graph of a Petri net [19]. Total deadlock is a leaf in reachability graph – no outgoing transition is present. Thus, reachability graph analysis is similar to model checking techniques, and typically they are combined as temporal analysis of the graph. In both approaches it is hard to distinguish a deadlock from distributed termination: these methods are addressed to endlessly cycling systems [20].

Alternatively, structural analysis of Petri net can be used. Structural analysis determines properties of models on the basis of their structure, so no exploration of the reachability

space is needed. Structural analysis of deadlocks is based on subnets called siphons [6][7]. It can be shown that if a model is deadlocked, the unmarked places constitute a siphon. Structural analysis of deadlocks systematically finds elementary siphons. Elementary siphons are ones from which other siphons are composed. After siphons identification, they are checked for unmarking possibility. The advantage of these methods is that multiple deadlocks are found in a single verification and both total and partial deadlocks are identified.

The deadlock detection procedure presented in this paper, based on combining siphon identification with temporal analysis, joins the advantages of the two methods and frees from their disadvantages:

- Identification of multiple deadlocks in single verification run.
- Finding both total and partial deadlocks.
- Distinguishing deadlocks from termination.
- Distinguishing between communication deadlocks and resource deadlocks.
- Automated verification, as deadlocks are expressed as formulas not related to specific features of a verified system.
- Verification of systems having arbitrary shape, without limitation to cycling, terminating or other schemes.

III. Integrated Model of Distributed Systems (IMDS)

A. Basic definition

IMDS is defined in [1][2]. In the present paper, the simplified version of IMDS is used, without dynamic process creation, which is suitable for static model checking. The formalism is founded on a basic observation: nodes on a distributed system (which are called *severs* in IMDS) receive messages, execute some actions changing their *states* upon accepted messages and finally send consecutive *messages*. Thus a distributed system may be defined as a relation between a finite set states of the servers $P=\{p_1,p_2,\ldots\}$ and a finite set of messages $M=\{m_1,m_2,\ldots\}$. The relation Λ defining the actions is:

$\Lambda \subset (M \times P) \times (M \times P)$

For an action $\lambda\in\Lambda$, $\lambda=((m,p),(m',p'))$ the first pair (m,p) is its *input* while the second pair (m',p') is its *output*.

A *configuration* T of a distributed system is a set of *current* states and *pending* messages. The messages in a configuration are current as well, but they are called pending to emphasize the fact that an action extracts from the configuration exactly one message, replacing it with a next message, and all other messages addressed to the action's server remain pending at the server. The system starts from its *initial configuration* T_0, containing initial set of states and messages.

Every action $\lambda=((m,p),(m',p'))$ converts a configuration $T_{inp}(\lambda)$ to a new configuration $T_{out}(\lambda)$ by replacing $\{m,p\}\subset T_{inp}(\lambda)$ with $\{m',p'\}\subset T_{out}(\lambda)$. Behavior of a distributed system is described by a Labeled Transition System LTS [21], containing all executions of the system. *Nodes* of the LTS (not called *states* for unambiguousness) are configurations and *transitions* are actions:

$LTS = \langle Q,q_0,W \rangle \mid$
$Q = \{T_0,T_1,\ldots\}$ (nodes);
$q_0 = T_0$ (initial node);
$W = \{(T,\lambda,T') \mid \lambda \in \Lambda,\ T=T_{inp}(\lambda),\ T'=T_{out}(\lambda)\}$
(transitions)

The interleaving way of executing the action is assumed (one action at a time [22]). Since all transitions in the LTS are instantaneous, it is assumed that message passing and actions take zero time. A timed version of IMDS, in which message passing and actions execution takes some periods of time, is also developed. This feature goes beyond the scope of this paper.

To differentiate between messages sent in different purposes (in a context of separate distributed computations), the autonomous sequential computations in a distributed system are extracted. The messages passed in a context of a given computation conform an *agent*. Thus the set of states P is split into subsets for the servers $1..n$ and the set of messages M into subsets for the agents $1..k$:

$P=\bigcup_{i=1..n} P_i$, $M=\bigcup_{j=1..k} M_j$

The subsets are pairwise disjoint:

$i\neq j \Rightarrow P_i \cap P_j=\varnothing$, $i\neq j \Rightarrow M_i \cap M_j=\varnothing$

The initial configuration contains *initial states* of all servers, one state for every server, and *initial messages* of all agents, one message for every agent:

$T_0\cap P_i=\{p_{0i}\}$, $T_0\cap M_i=\{m_{0i}\}$

The input and output state of an action concern the same server and the input and output message of an action concern the same agent:

$\lambda=((m,p),(m',p'))$, $\{m,m'\}\subset M_i$, $\{p,p'\}\subset P_j$

Agents may terminate in special actions of the form $\lambda=((m,p),(p'))$, where an output message is absent.

B. Processes

Is it useful to identify processes in a system, especially for verification purposes. Two "classical" models of distributed processes are used: client-server and RPC [23]. In the former model servers communicate by messages while in the latter one processes migrate forth and back. IMDS contains both models in a single specification, the models are extracted as the two perspectives: *server view* and *agent view*. A server process B in the server view is a sequence of actions connected by states of a server, as input and output states of an action concern the same server. Some actions may be unreachable and thus they would become "orphaned" (not included in any process), so the definition is extended to a set of all actions of a given server (rather than a sequence):

$B_i=\{\lambda\in\Lambda \mid \lambda=((m,p),(m',p')) \vee \lambda=((m,p),(p')),\ p,p'\in P_i\}$

In the agent view, an agent process C is a sequence of actions connected by messages of an individual agent, because input and output messages of an action concern the same

agent. As for server processes, the definition is extended to a set of all actions of a given agent:

$C_j=\{\lambda\in\Lambda \mid \lambda=((m,p),(m',p')) \vee \lambda=((m,p),(p')), m,m'\in M_j\}$

As a result, a distributed system may be decomposed to server processes or to agent processes, giving the server decomposition **B** and agent decomposition **C**:

$\mathbf{B}=\{B_i \mid i=1..n\}$

$\mathbf{C}=\{C_j \mid j=1..k\}$

C. Locality, Asynchrony and Autonomy

An important feature of a distributed system is locality. In IMDS, locality means that no message may cause actions in distinct servers (for a message $m\in M$ and two actions $\lambda_1,\lambda_2\in\Lambda$, $\lambda_1=((m,p_1),(m_1',p_1'))$, $\lambda_2=((m,p_2),(m_2',p_2'))$, $p_1,p_2\in P_i$). Thus, a function *target*: $M\rightarrow S$ assigns a target server for every message. A server component of a message and a state in the input pair of an action must match: $target(m)=s_i$, $p\in P_j$, $i=j$. A configuration contains exactly one state for every server, as this is required for initial configuration, and every action replaces its input state with a next state of the same server. Yet, multiple messages may be pending at given server in a configuration, which is natural. A set of actions in a distributed system determines which messages may be accepted in individual states. If a server allows a message to be accepted, an action is defined for this message together with the current state of the server. If the acceptance of a message is prohibited in a given state, no action is defined for this pair.

Note that every server performs its action autonomously (only the current server's state and the messages pending at this server are considered). Also, the communication is asynchronous: a server process sends a message to some other server process (or in the agent view, an agent sets the server's state for some other agent) regardless of the current situation of a process with which it communicates (and every other process). As a result, the process may be called *autonomous* and *asynchronous*.

D. Deadlocks in IMDS

A deadlock in IMDS is defined as a discontinuation of a process (with an exception of process termination). As there are two views of processes, different type of deadlocks concern server processes communicating by messages and agent processes communicating by states of servers. There may be a communication deadlock that is not a resource deadlock [2].

- a *communication deadlock* of a server process – when there are messages pending at the server, but no matching pair of any message with a current server state will occur;
- a *resource deadlock* of an agent process – when an agent's message is pending at a server but it will never match any current or future state of this server.

For the identification of deadlocks, universal temporal formulas were elaborated [2]. Universality of the formulas means that they are independent on a structure of a given distributed system – only the two facts are concerned: if an action in a process is enabled and if it is executed. Therefore, temporal logic is built inside the verification tool and the user need not know temporal logic nor model checking technique.

The paper [2] presents a terminating distributed system in which two servers, every one containing a semaphore, are used by two agents (the third agent performs some other work to show a detection of a partial deadlock). This shows a communication deadlock in the server view and a resource deadlock in the agent view. The system is presented briefly in Sect VII. Another example [24] is the Automatic Vehicle Guidance System: in the server view the cooperation of the road segment controllers during the piloting of a vehicle is shown, while in the agent view the traffic from the vehicles perspective is presented. The deadlocks in both views are shown. Similar system is mentioned in Sect. VIII. In the IMDS specification of Karlsruhe Production Cell [25], the controllers of individual devices are modeled as servers and the metal plates traveling through the cell are agents. Additional agents server for performing some actions without the plates, for example return to a rest position.

IV. Example – Bounded Buffer

An example of IMDS system is a buffer with producer and consumer agents (each of them starting at its own server). In IMDS notation, sets of servers *S,* agents *A,* state values *V* and services *R* are introduced explicitly. The services are used to distinguish between messages sent in given purposes to the servers (like operations *wait* and *signal* on a semaphore). The messages are triples (a,s,r), $a\in A$, $s\in S$, $r\in R$. The states of servers are pairs (s,v) , $s\in S$, $v\in V$, where v is a value that represents a given state. Thus an action $(m,p)\Lambda(m',p')$ is denoted $((a,s,r),(s,v))\Lambda((a,s',r'),(s,v'))$. Note that the same agent a is used in an input and output message, in such a way a computation is continued. Likewise, the same server s is used in an input and output state. Also, the same server s is in input state and input message, which models an acceptance of a message on a server with its current state.

The server view of the system in Dedan notation is presented below. As in typical programming language, server types are introduced (lines 3, 12), with formal parameters that specify agents and other servers used in actions. Every server has sets of its states (l.4, 13), services (l.5, 14) and actions (l.7-10, 16-19). An action in Dedan is denoted `{a.s.r, s.v}→{a.s'.r', s.v'}`. Servers and agents are declared as variables (l.21-22, server types are omitted in the declaration because they have names equal to variables in this example). Lastly, actual parameters are passed to the servers and initial states are assigned to every server and initial messages are assigned to every agent (l.24-26).

```
1.  #DEFINE N 2
2.  #DEFINE K 1

3.  server: buf(agents A[N];servers S[N]),
4.  services{put, get},
5.  states{elem0,elem[K]},
6.  actions {
7.  <i=1..N>{A[i].buf.put, buf.elem0}
       -> {A[i].S[i].ok_put, buf.elem[1]},
8.  <i=1..N><j=1..K-1>{A[i].buf.put, buf.elem[j]}
       -> {A[i].S[i].ok_put, buf.elem[j+1]},
9.  <i=1..N><j=2..K>{A[i].buf.get, buf.elem[j]}
       -> {A[i].S[i].ok_get, buf.elem[j-1]},
10. <i=1..N>{A[i].buf.get, buf.elem[1]}
       -> {A[i].S[i].ok_get, buf.elem0}
11. };

12. server: S(agents A;servers buf),
13. services{doSth,ok_put,ok_get}
14. states{neutral,prod,cons}
15. actions {
16. {A.S.doSth, S.neutral}
       -> {A.buf.put, S.prod}
17. {A.S.doSth, S.neutral}
       -> {A.buf.get, S.cons}
18. {A.S.ok_put, S.prod}
       -> {A.S.doSth, S.neutral}
19. {A.S.ok_get, S.cons}
       -> {A.S.doSth, S.neutral}
20. };

21. servers buf,S[N];
22. agents A[N];

23. init  ->{
24. <j=1..N>S[j](A[j],buf).neutral,
25.         buf(A[1..N],S[1..N]).elem0,
26. <j=1..N>A[j].S[j].doSth,
27. }.
```

Obviously, one can expect two deadlocks in the example: two agents both trying to get from an empty buffer and two agents trying to put to a full buffer. In model checking, a verifier typically shows the former deadlock because it requires two *get* operations to be reached. However, the latter deadlock is reached after three *put* operations, which lengthens a counterexample. A model checker searches for first counterexample and then it stops the evaluation. Therefore the latter deadlock may be reported in a second verification run, after a modification of the system to avoid the former deadlock (if the modification does not repair both deadlocks).

V. Deadlock Detection in a Petri Net Equivalent to IMDS Model

The main task of the Dedan program is identification of deadlocks and distributed termination. The Conversion of an IMDS system to a Petri net offers the designer new possibilities:

- identification of some structural properties: structural conflicts, dead code, etc.,
- temporal properties expressed in terms of Petri net,
- observation of the system in a graphical form,
- graphical simulation of a system run.

For this purpose a possibility of export of a model to a Petri net is included in Dedan. A format of ANDL (Abstract Net Description Language [26]) is used, which is the input of Charlie Petri net analyzer [27][28].

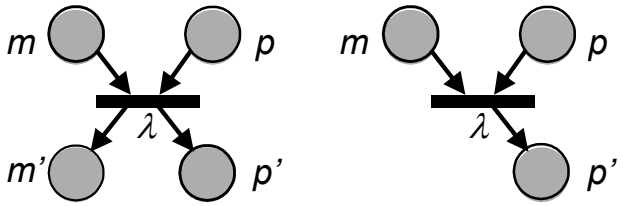


Fig. 1 IMDS actions converted to Petri net transitions: regular action (left) and agent-terminating action (right)

Fig. 1 shows a transition of a Petri net, which corresponds to an IMDS action. The input message and the input state are input places. The output message and the output state are output places (or only the output state in the case of a terminating action, Fig. 1b). The initial marking of the Petri net has tokens in the initial places of server states and initial messages of agents. The graph of reachable markings is equivalent to the LTS of the IMDS system, where states and messages correspond to the places, actions correspond to the transitions and markings correspond to the configurations.

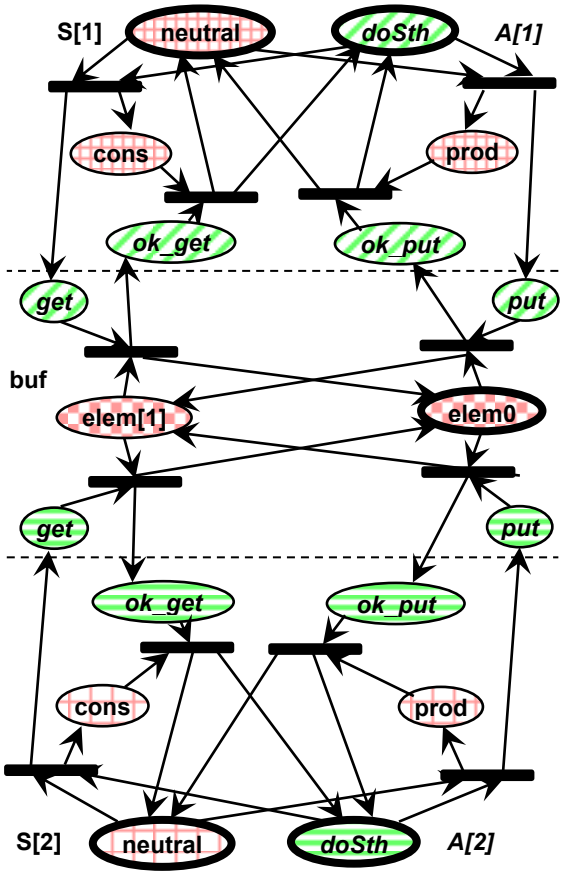


Fig. 2 Petri net representation of the *bounded buffer* system: servers *S*[1..2], *buf*, agents *A*[1..2]

The Petri net corresponding to the *bounded buffer* example is illustrated in Fig. 2. The states and messages in individual servers are grouped and separated by dashed lines. The states of servers are filled red, with patterns individual to every server. The messages are filled green, with patterns individual to every agent. Initial states and initial messages are bold. The servers and their states are in regular font, while the agents and their messages are in italics.

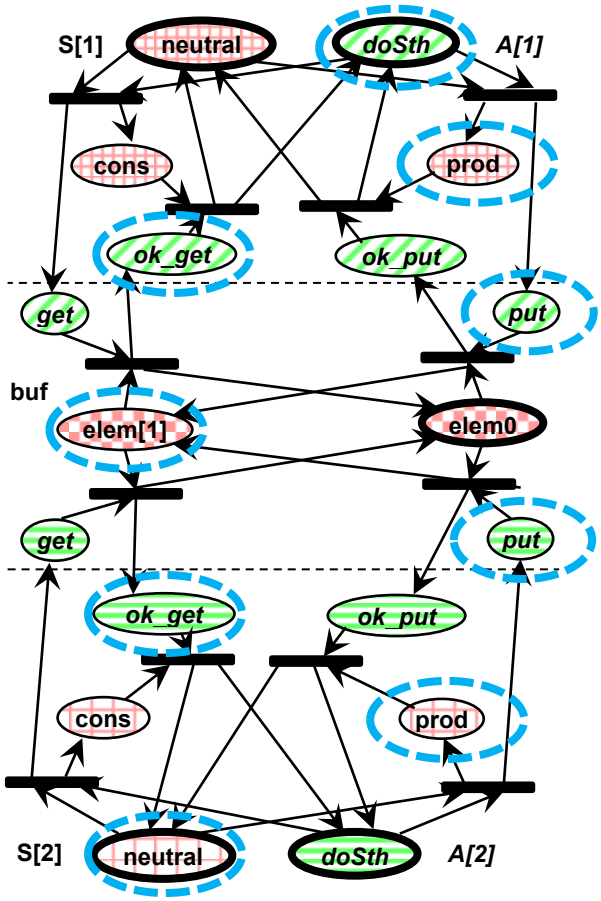


Fig. 3 One of elementary siphons found in the system

As the system falls into a deadlock, there should be siphons that may be emptied. A Petri net may contain a large number of siphons, but some of them are elementary, i.e., they do not contain other siphons [29]. Therefore, only elementary siphons need to be analyzed. The Charlie program reports 49 elementary siphons in the net. Every siphon should be tested for a reachability of its emptying. As an empty state place denotes a state which is absent in a configuration, and an empty message place denotes an absent message, an IMDS configuration should be found in which the siphon's states and messages are absent. A siphon may concern not all of the servers (and/or not all of the agents), in such a way partial deadlocks are found.

One of the siphons found in our example is presented in Fig. 3. It contains states (S[1],*prod*), (*buf*,*elem*[1]), (S[2],*prod*), (S[2],*neutral*) and messages (A[1],S[1],*ok_get*), (A[1],S[1],*doSth*), (A[1],*buf*,*put*), (A[2],*buf*,*put*), (A[2],S[2],*ok_get*).

The siphon emptying is verified by model checking. To do this, the CTL formula **AG** ($\neg$ φ) (it reads: always not φ) is used, where φ is a set of states and messages in a configuration corresponding to a siphon's complement. Often a siphon represents a class of configurations, for example a siphon in Fig. 3 represents all configurations in which server S[1] is not in a state (S[1],*prod*), and thus it may be in one of a subset of states {(S[1],*neutral*), (S[1],*cons*)}. The formula for checking if the siphon cannot be emptied has the form **AG** ($\neg$ ((((S[1],*neutral*)$\vee$(S[1],*cons*)) $\wedge$ (*buf*,*elem0*) $\wedge$ (S[2],*cons*) $\wedge$ ((A[1],S[1],*ok_put*)$\vee$ (A[1],*buf*,*get*)) $\wedge$ ((A[2],S[2],*doSth*)$\vee$(A[2],S[2],*ok_put*)$\vee$ (A[2],*buf*,*get*)))). For verification, internal Dedan model checker is used for typical cases (as it uses explicit state space) and Uppaal [31] for large cases.

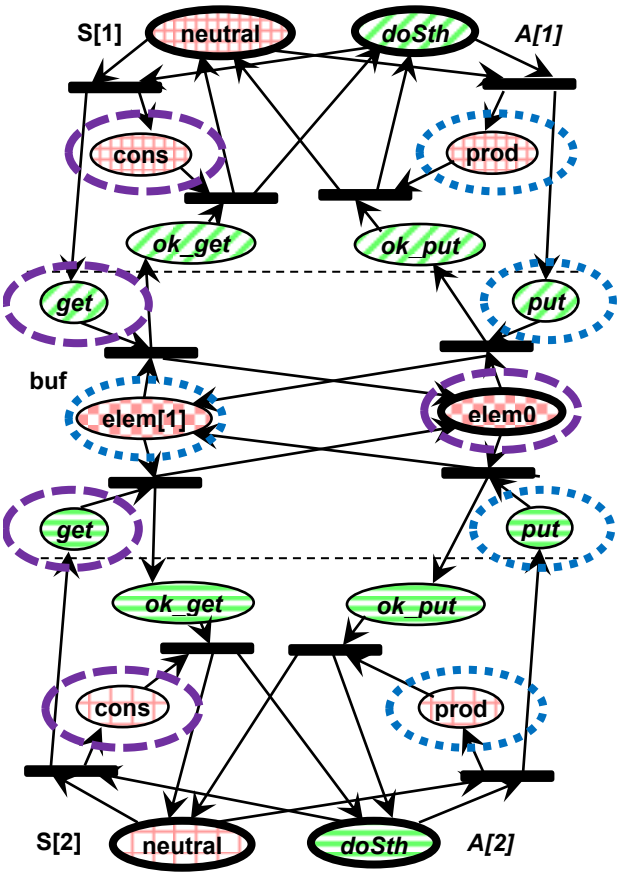


Fig. 4 The two deadlocks identified in the example. The dashed one is associated with the siphon in Fig. 3

The verification results in *false*. This means that the empty siphon is reachable, which denotes a deadlock. States of all three servers S[1],S[2] and *buf* take part in the siphon, so it is a total communication deadlock. Also, both agents A[1] and A[2] take part, denoting a total resource deadlock. Model checker generates a counterexample, in which both agents perform *get* on empty buffer (state (*buf*,*elem0*)). Charlie

reports multiple siphons that may be emptied by two *get* operations on the empty buffer. All these situations constitute a single deadlock (but the counterexamples for individual temporal formulas may differ in the order of issuing *get* by the two agents). The configuration finishing all counterexamples leading to this deadlock is exactly the same. For a partial deadlock, configurations finishing the counterexamples may differ, but only in states/messages of servers/agents not taking part in the deadlock. In the example, all tests for emptying of the siphons either finish in one of the two deadlock configurations, or emptying occurs unreachable (such siphons do not denote a deadlock).

Fig. 4 shows the two possible deadlocks that finish all reachable emptying of siphons. The ovals surround places of the two identified deadlocks: dashed ovals are associated with the siphon in Fig. 3, which it is caused by two *gets* on an empty buffer. The other deadlock (dotted ovals) is caused by two *puts* on a full buffer. Distinguishing between the two deadlocks is based on the two configurations finishing the counterexamples.

VI. Verification of Systems with Various Structures

Various systems can be modeled in IMDS, not only those having a shape of purely cyclic FMS. Fig. 5 shows some examples of shapes of not purely cyclic systems. In the figure "Ending Strongly Connected Subgraph" is a cycle from which no escape is possible. The pictures are schematic, showing general shape of a system. In IMDS specification every transition has two input places and two output places (or one in the case of agent-terminating action), see Fig. 1.

- "linear" systems (like the example of "two semaphores" described in [2]: users issue *wait* on two semaphores, then they issue *signal*),
- a system with a "*leader*" (initial part) and a main loop, sometimes called "lasso-shaped" [32],
- terminating system with a main loop, for example WF-net system [33],
- similar to lasso-shaped, but with an initial loop.

In some cases, for example in the acyclic system in Fig. 5 (on the top), the system may be easily converted to a cyclic one by connecting initial and terminating places. However in the analysis of distributed systems, especially those following the IoT paradigm, in which autonomous nodes agree their coordinated behavior. Such system may have multiple leaders (for every node) and multiple terminating places, where the nodes reach their goals. An example is Automatic Vehicle Guidance System presented in [24]. In IMDS model, servers implement road segment controllers while agents implement the vehicles. Such a system may be additionally complicated if endlessly-looping nodes are added, for example charging stations where serving agents run in cycles.

It is obvious that most of the leaders contain siphons that may be emptied. Yet, emptying of such a siphon does not denote a deadlock because the system runs further. This is the main difference in deadlock detection between purely cyclic and differently shaped systems.

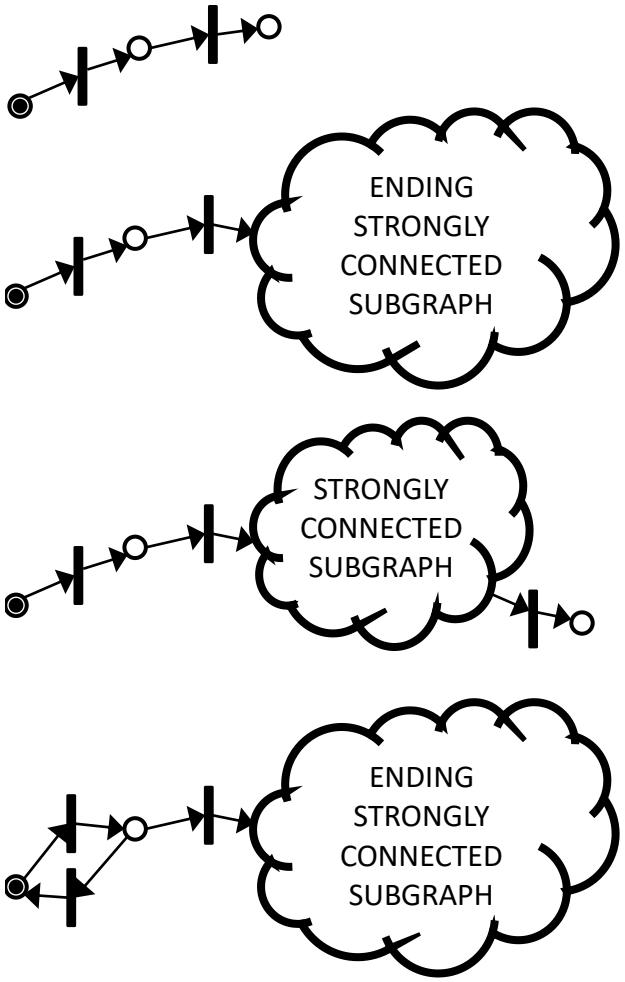


Fig. 5 Examples of acyclic and not purely cyclic systems containing emptyable and reachable siphons which are not deadlocks

A solution of this problem is quite simple: an emptied siphon should be verified if it is a real deadlock or not. This is done using an additional temporal formula, which uses the same subformula φ of an emptied siphon. A deadlock prevents a process from doing any move. Thus, the evaluation of $\mathbf{AG}(\neg \varphi)$ to *false* should be followed by application for every process (server and agent) the formula $\mathbf{AG}(\varphi \Rightarrow \mathbf{EF} \neg \varphi$/restricted to a process). The formula reads: always φ is inevitably followed by not φ. Of course, this procedure may be applied to a system of any shape, cycling or not. For the example of emptyable siphon in Fig. 3 (we can pretend that we do not know that the system is purely cyclic) the verification should be performed as follows:

- φ=((*S*[1],*neutral*)∨(*S*[1],*cons*)) ∧ (*buf,elem0*) ∧ (*S*[2],*cons*) ∧ ((*A*[1],*S*[1],*ok_put*)∨(*A*[1],*buf,get*)) ∧ ((*A*[2],*S*[2],*doSth*)∨(*A*[2],*S*[2],*ok_put*)∨(*A*[2],*buf,get*)),
- check $\mathbf{AG}(\varphi \Rightarrow \mathbf{EF}$ (*S*[1],*prod*)) for server *S*[1],

- check **AG**($\varphi \Rightarrow$ **EF** (*buf*,*elem*[1])) for server *buf*,
- check **AG**($\varphi \Rightarrow$ **EF** ((*S*[2],*prod*)$\vee$(*S*[2],*neutral*))) for server *S*[2],
- check **AG**($\varphi \Rightarrow$ **EF** ((*A*[1],*S*[1],*doSth*)$\vee$ (*A*[1],*S*[2],*ok_get*)$\vee$(*A*[1],*buf*,*put*))) for agent *A*[1],
- check **AG**($\varphi \Rightarrow$ **EF** ((*A*[2],*S*[2],*ok_get*)$\vee$(*A*[2],*buf*,*put*))) for agent *A*[2].

The result depends on a value of a formula for every process:

- *true* for every involved server – no communication deadlock,
- *false* for all involved servers, and all servers are involved – total communication deadlock,
- *false* for some involved servers, or for all servers involved but not all servers are involved – partial communication deadlock,
- *true* for every involved agent – no resource deadlock,
- *false* for all involved agents, and all agents are involved – total resource deadlock,
- *false* for some involved agents, or for all agents involved but not all agents are involved – partial resource deadlock.

In the last three cases concerning agents, only non-terminated agents are taken under consideration, i.e. the agents which messages are present in the configuration corresponding to the emptied siphon.

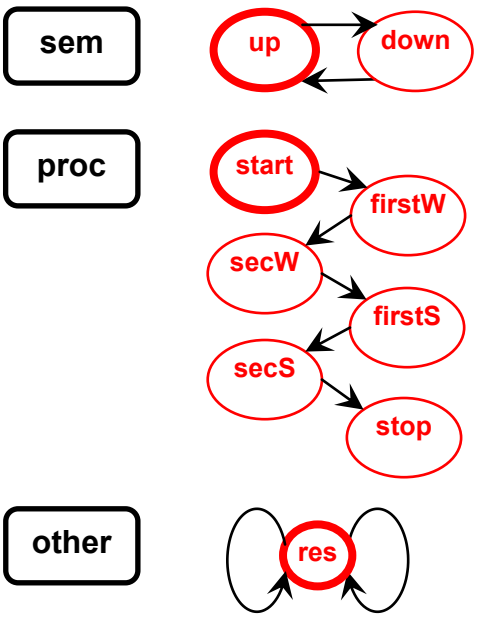


Fig. 6 Automata-like model of servers in *two-semaphores* system, agents not shown

The verification procedure uses several well-known and widely used algorithms. Computational aspects of finding siphons are discussed in [34]: elementary siphons may be found in linear time for a large class of Petri nets (and needs some preprocessings in general case). Also, parallel solutions exist [35].

The complexity of CTL model checking is P-Complete [36], which means that the time of temporal formula evaluation is $|\text{LTS}|\times|\psi|$, where $|\text{LTS}|$ denotes the number of nodes in a Kripke structure (it is the LTS of a verified system) and $|\psi|$ is the length of a formula ψ. Every formula **AG**($\varphi_1 \Rightarrow$ **EF** φ_2) contains two temporal operators, so this evaluation cost is fixed. The verification should be repeated for every siphon, and according to each siphon for every server and every agent, i.e., the complexity is a number of elementary siphons $\times$ (n+k) $\times$ $|\text{LTS}|$, where n is a number of servers and k is a number of agents.

VII. Example System with Leaders

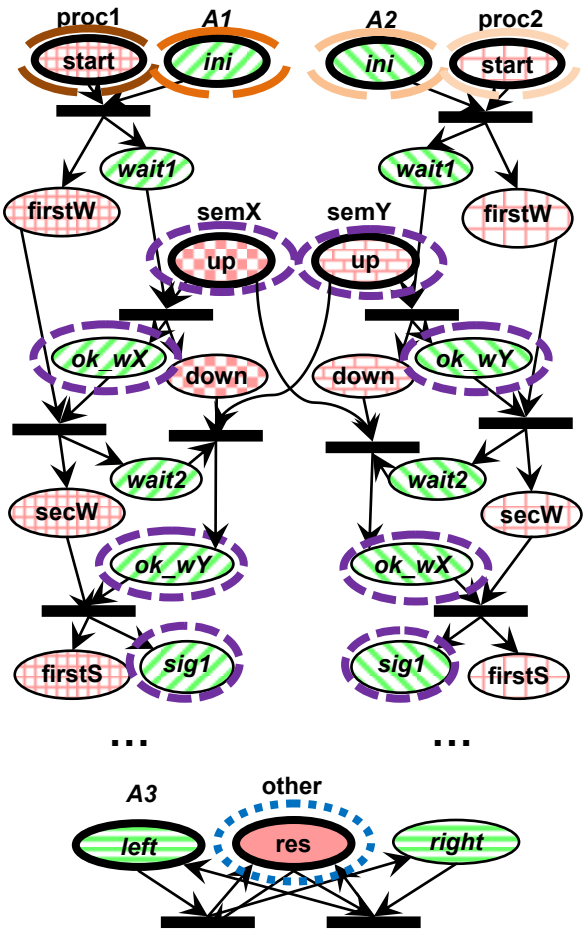


Fig. 7 Petri net representation of the *two-semaphores* system

As an example of system with leaders, we present a *two-semaphores* system consisting of two agents *A1* and *A2*, each one running on its own server (*proc1* and *proc2*). The agents use two semaphores *semX* and *semY*. They use the semaphores "crosswise", i.e., *A1* issues operation *wait* to *semX* than to *semY*, while *A2* does it in opposite order. To show a partial deadlock (not concerning all the servers/agents), the third agent *A3* is added, running on its own server *other* and

performing some looping calculations. The automata-like view of the system is presented in Fig. 6 (only servers' states and actions are shown, input and output messages in actions are omitted).

The system converted to a Petri net if shown in Fig. 7 (a part after second *wait* in the agents is suppressed). The general shape of the system consists in sequences of actions in agents *A1* and *A2*, leading from their start to their termination, and a separated Ending Strongly Connected Subgraph of agent *A3* (depicted in Fig. 8).

A verification in Charlie shows a siphon of an obvious deadlock, shown as places surrounded with denser dashed ovals (violet). This siphon is emptyable and the temporal formulas identifying processes involved show:

- Both agents *A1* and *A2* fall into resource deadlock.
- Both servers *semX* and *semY* fall into communication deadlock.

There is also a siphon containing the place *res* (dotted oval, dark blue), but it is not emptyable - this does not denote a deadlock. There are four such siphons in the system (three of them are not indicated in the figure).

Four siphons formed by places *proc1.start, proc2.start, A1.proc1.ini* and *A2.proc2.ini* are emtyable (they are depicted as sparsely dashed ovals on the top of Fig. 7), but the temporal formulas show that these siphons do not denote deadlocks. They are typical leader siphons.

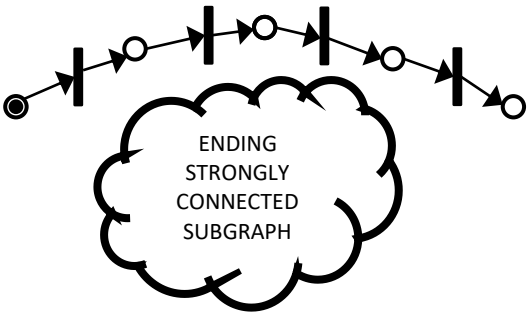


Fig. 8 General shape of the *two-semaphores* system. The chain - servers *proc1, proc2, semX, semY* and agents *A1, A2*. The cycle – server *other* and agent *A3*.

Summing up:

- A partial communication deadlock of processes *semX* and *semY* is identified, in which processes *proc1, proc2* and *other* are not involved.
- A partial resource deadlock of processes *A1* and *A2* is identified, in which process *A3* is not involved.
- Four not emptyable siphons are found (no-deadlock siphons).
- Four emptyable leader siphons are found (no-deadlock siphons).

VIII. Example Application to Automatic Vehicle Guidance System

We chose an Automatic Vehicle Guidance System (AVGS) verification to illustrate an application of our method. The AVGS system consists of a set of road segments (identifiers are taken from cardinal directions) with their controllers modeled as servers, for example on a crossing depicted in Fig. 9. The controllers are very simple: they allow or deny a vehicle to take up a road segment, depending on its freeness or occupation. Vehicles are modeled as agents. When more than one vehicle approaches the crossing, routes for all the vehicles are prepared, for instance using a genetic algorithm. A route connects an entrance segment (*A...*) and a target segment (*T...*) of vehicle's travel. Obviously, in the model every route terminates. The routes are tested for deadlock freeness using siphon detection and temporal verification, described in this paper. This can be performed automatically, because in our methodology deadlock detection formulas are independent on the structure of a verified system. Deadlock-free route sets are executed while routes exposed to deadlocks are rejected. If a new vehicle appears, the whole procedure is repeated from current positions of all the vehicles on the crossing and approaching ones.

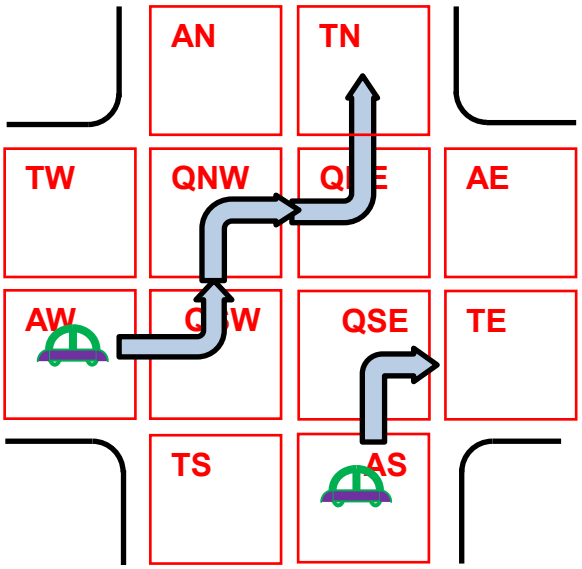


Fig. 9 Automatic Vehicle Guidance System

The described procedure may find solutions unusual in ordinary vehicle traffic, for example one of the routes shown in Fig. 9 causes a vehicle to apply left side traffic for a while. Also, safe routes may be fond even if one road segment is blocked, for example by a broken vehicle. In verification, AVGS is similar to the *two-semaphores* system, without server *other* and agent *A3*.

Other examples of systems that are based on terminating processes, which may be modeled using our approach, is taxi

service [37], business processes [38] or multi-agent systems based on Belief-Desire-Intention paradigm [39]

IX. Conclusions

An approach to deadlock detection is presented which is based on coupling IMDS formalism with Petri net structural analysis and model checking. The methodology allows to find total and partial deadlocks in two perspectives: servers communicating by messages and agents communicating by states of servers. As a result, communication deadlocks and resource deadlocks are identified, which highlights communication duality in distributed systems. Also, the specification in IMDS clearly identifies processes running in a system, which is sometimes difficult in ordinary Petri nets.

The methodology finds multiple deadlocks in a distributed system, preserving locality of decisions, autonomy of servers, and asynchrony of behavior and communication. In some rare cases, in which siphons can be emptied in various ways, not all deadlocks are identified in a single run, but the procedure may be repeated after correction of found deadlocks. However, we have not found any real-life example of a system with such feature, typically all deadlocks are found in a single run.

The presented deadlock detection procedure may be applied for system of arbitrary shape: cycling like a class of FMS systems, terminating like WF-nets, lasso-shaped, or IoT systems compound of multiple terminating and looping autonomous nodes.

A collection of programs is used for the described procedure: Dedan for specification, Charlie for siphon identification, internal Dedan model checker and Uppaal for verification. In the future, the whole procedure will be integrated in Dedan. This will allow to check for deadlocks and to perform other types of structural analysis in a uniform environment.

In the future, a timed version of the prosed procedure is planned, with application of UPPAAL timed automata [40].

Acknowledgment

Extensive discussions with Wlodek Zuberek helped to significantly improve the article, especially in the new formulation of IMDS.

References

[1] S. Chrobot and W. B. Daszczuk, “Communication Dualism in Distributed Systems with Petri Net Interpretation,” *Theor. Appl. Informatics*, vol. 18, no. 4, pp. 261–278, 2006. arXiv: 1710.07907

[2] W. B. Daszczuk, “Communication and Resource Deadlock Analysis using IMDS Formalism and Model Checking,” *Comput. J.*, vol. 60, no. 5, pp. 729–750, 2017. doi: 10.1093/comjnl/bxw099

[3] C. Baier and J.-P. Katoen, *Principles of Model Checking*. Cambridge, MA: MIT Press, 2008. ISBN: 9780262026499

[4] Dedan, http://staff.ii.pw.edu.pl/dedan/files/DedAn.zip

[5] W. Reisig, *Petri Nets - An Introduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985. doi: 10.1007/978-3-642-69968-9

[6] D. C. Craig and W. M. Zuberek, “Two-stage siphon-based deadlock detection in Petri nets,” in *Current Advances in Computing, Engineering and Information Technology*, P. Petratos and P. Dandapami, Eds. Palermo, Italy: Int. Society for Advanced Research, 2008, pp. 317–330.

[7] F. Chu and X.-L. Xie, “Deadlock analysis of Petri nets using siphons and mathematical programming,” *IEEE Trans. Robot. Autom.*, vol. 13, no. 6, pp. 793–804, 1997. doi: 10.1109/70.650158

[8] M. Uzam, “An Optimal Deadlock Prevention Policy for Flexible Manufacturing Systems Using Petri Net Models with Resources and the Theory of Regions,” *Int. J. Adv. Manuf. Technol.*, vol. 19, no. 3, pp. 192–208, Feb. 2002. doi: 10.1007/s001700200014

[9] J. Ezpeleta, J. M. Colom, and J. Martinez, “A Petri net based deadlock prevention policy for flexible manufacturing systems,” *IEEE Trans. Robot. Autom.*, vol. 11, no. 2, pp. 173–184, Apr. 1995. doi: 10.1109/70.370500

[10] W. B. Daszczuk and W. M. Zuberek, “Deadlock Detection in Distributed Systems Using the IMDS Formalism and Petri Nets,” in *12th International Conference on Dependability and Complex Systems, DepCoS-RELCOMEX 2017*, Brunów, Poland, 2-6 July 2017. AISC vol 582, W. Zamojski et al., Eds, Cham, Switzerland: Springer International Publishing, 2018, pp. 118–130. doi: 10.1007/978-3-319-59415-6_12

[11] R. Agarwal and S. D. Stoller, “Run-Time Detection of Potential Deadlocks for Programs with Locks, Semaphores, and Condition Variables,” in *Proc. of the Workshop on Parallel and Distributed Systems: Testing and Debugging (PADTAD-IV), ISSTA, 2006,* Portland, ME, 17-20 July 2006, 2006, pp. 51–59. doi: 10.1145/1147403.1147413

[12] N. Kaveh, “Using Model Checking to Detect Deadlocks in Distributed Object Systems,” in *2nd International Workshop on Distributed Objects*, Davis, CA, 2-3 November 2000, LNCS vol.1999, 2001, pp. 116–128. doi: 10.1007/3-540-45254-0_11

[13] J. Cho, J. Yoo, and S. Cha, “NuEditor – A Tool Suite for Specification and Verification of NuSCR,” in *Lecture Notes in Computer Science vol. 3647*, Berlin Heidelberg: Springer, 2006, pp. 19–28. doi: 10.1007/11668855_2

[14] O. Inverso, T. L. Nguyen, B. Fischer, S. La Torre, and G. Parlato, “Lazy-CSeq: A Context-Bounded Model Checking Tool for Multi-threaded C-Programs,” in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Lincoln, NE, 9-13 November 2015, 2015, pp. 807–812. doi: 10.1109/ASE.2015.108

[15] Y. Yang, X. Chen, and G. Gopalakrishnan, “Inspect: A Runtime Model Checker for Multithreaded C Programs", Report UUCS-08-004, University of Utah, Salt Lake City, UT, 2008, http://www.cs.utah.edu/docs/techreports/2008/pdf/UUCS-08-004.pdf

[16] P. C. Attie, “Synthesis of large dynamic concurrent programs from dynamic specifications,” *Form. Methods Syst. Des.*, vol. 47, no. 131, pp. 1–54, Jun. 2016. doi: 10.1007/s10703-016-0252-9

[17] D. Fahland, C. Favre, J. Koehler, N. Lohmann, H. Völzer, and K. Wolf, “Analysis on demand: Instantaneous soundness checking of industrial business process models,” *Data Knowl. Eng.*, vol. 70, no. 5, pp. 448–466, May 2011. doi: 10.1016/j.datak.2011.01.004

[18] S. J. C. Joosten, F. V. Julien, and J. Schmaltz, “WickedXmas: Designing and Verifying on-chip Communication Fabrics,” in *3rd International Workshop on Design and Implementation of Formal Tools and Systems, DIFTS’14*, Lausanne, Switzerland, October 20, 2014, 2014, pp. 1–8. https://pure.tue.nl/ws/files/3916267/889737443709527.pdf

[19] S. Duri, U. Buy, R. Devarapalli, and S. M. Shatz, “Application and experimental evaluation of state space reduction methods for deadlock analysis in Ada,” *ACM Trans. Softw. Eng. Methodol.*, vol. 3, no. 4, pp. 340–380, Oct. 1994. doi: 10.1145/201024.201038

[20] X. Guan, Y. Li, J. Xu, C. Wang, and S. Wang, “A Literature Review of Deadlock Prevention Policy Based on Petri Nets for Automated Manufacturing Systems,” *Int. J. Digit. Content Technol. its Appl.*, vol. 6, no. 21, pp. 426–433, Nov. 2012. doi: 10.4156/jdcta.vol6.issue21.48

[21] M. A. Reniers and T. A. C. Willemse, “Folk Theorems on the Correspondence between State-Based and Event-Based Systems,” in *37th Conference on Current Trends in Theory and Practice of Computer Scienc*e, Nový Smokovec, Slovakia, January 22-28, 2011, 2011, pp. 494–505. doi: 10.1007/978-3-642-18381-2_41

[22] W. Penczek, M. Szreter, R. Gerth, and R. Kuiper, "Improving Partial Order Reductions for Universal Branching Time Properties," *Fundam. Informaticae*, vol. 43, no. 1–4, pp. 245–267, 2000. doi: 10.3233/FI-2000-43123413

[23] W. Jia and W. Zhou, *Distributed Network Systems. From Concepts to Implementations*. New York: Springer, 2005. doi: 10.1007/b102545

[24] B. Czejdo, S. Bhattacharya, M. Baszun, and W. B. Daszczuk, "Improving Resilience of Autonomous Moving Platforms by real-time analysis of their Cooperation," *Autobusy-TEST*, vol. 17, no. 6, pp. 1294–1301, 2016. arXiv: 1705.04263

[25] W. B. Daszczuk, "Asynchronous Specification of Production Cell Benchmark in Integrated Model of Distributed Systems," in *23rd International Symposium on Methodologies for Intelligent Systems, ISMIS 2017*, Warsaw, Poland, 26-29 June 2017, Studies in Big Data, vol. 40, Bembenik, R. et al., Eds, Cham, Switzerland: Springer International Publishing, 2019. pp. 115-129. doi: 10.1007/978-3-319-77604-0_9

[26] M. Schwarick, M. Heiner, and C. Rohr, "MARCIE - Model Checking and Reachability Analysis Done EffiCIEntly," in *2011 Eighth International Conference on Quantitative Evaluation of SysTems*, Aachen, Germany, 5-8 Sept. 2011, 2011, pp. 91–100. doi: 10.1109/QEST.2011.19

[27] Charlie, http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software/Charlie

[28] M. Heiner, M. Schwarick, and J.-T. Wegener, "Charlie – An Extensible Petri Net Analysis Tool," in *36th International Conference, PETRI NETS 2015*, Brussels, Belgium, 21-26 June 2015, 2015, pp. 200–211. doi: 10.1007/978-3-319-19488-2_10

[29] Z. Li and M. Zhou, "Elementary Siphons of Petri Nets and Their Application to Deadlock Prevention in Flexible Manufacturing Systems," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 34, no. 1, pp. 38–51, Jan. 2004. doi: 10.1109/TSMCA.2003.820576

[30] M. H. Abdul-Hussin, "Elementary Siphons of Petri Nets and Deadlock Control in FMS", *J. of Comput. Commun.,* vol.3, No.7, pp. 1–12, Jul 2015. doi: 10.4236/jcc.2015.37001

[31] G. Behrmann, A. David, K. G. Larsen, P. Pettersson, and W. Yi, "Developing UPPAAL over 15 years," *Softw. Pract. Exp.*, vol. 41, no. 2, pp. 133–142, Feb. 2011. doi: 10.1002/spe.1006

[32] T. Latvala, A. Biere, K. Heljanko, and T. Junttila, "Simple Bounded LTL Model Checking," in *International Conference on Formal Methods in Computer-Aided Design*, Austin, TX, 15-17 Nov 2004, LNCS 3312, 2004, pp. 186–200. doi: 10.1007/978-3-540-30494-4_14

[33] W. M. P. van der Aalst, "Workflow Verification: Finding Control-Flow Errors Using Petri-Net-Based Techniques," in *Business Process Management, LNCS vol.1806*, W. van der Aalst, J. Desel, and A. Oberweis, Eds. Berlin Heidelberg: Springer, 2000, pp. 161–183. doi: 10.1007/3-540-45594-9_11

[34] M. Yamauchi and T. Watanabe, "Time Complexity Analysis of the Minimal Siphon Extraction Problem of Petri Nets," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E82–A, no. 11, pp. 2558–2565, 1999.

[35] F. Tricas and J. Ezpeleta, "Computing minimal siphons in Petri net models of resource allocation systems: a parallel solution," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 36, no. 3, pp. 532–539, May 2006. doi: 10.1109/TSMCA.2005.855751

[36] P. Schnoebelen, "The complexity of temporal logic model checking," in *4th Conference Advances in Modal Logic (AiML'2002), Toulouse, France, 30 Sept - 2 Oct 2004, Advances in Modal Logic vol. 4*, 2003, pp. 437–459. http://www.aiml.net/volumes/volume4/Schnoebelen.ps

[37] R. Klimek and P. Szwed, "Verification of ArchiMate proc*ess specifications based on deductive temporal reasoning," in FEDCSIS 2013 - Federated Conference on Computer Science and Information Systems*, Kraków, Poland, 8-11 Sept 2013, 2013, pp. 1109–1116. https://ieeexplore.ieee.org/document/6644153/

[38] P. Szwed, "Efficiency of formal verification of ArchiMate business processes with NuSMV model checker," in *FEDCSIS 2015 - Federated Conference on Computer Science and Information Systems*, Łódz, Poland, 13-16 Sept 2015, 2015, pp. 1427–1436. doi: 10.15439/2015F44

[39] A. T. E. Dib and Z. Sahnoun, "Model Checking of Multi Agent System Architectures Using BigMC," in *FEDCSIS 2015 - Federated Conference on Computer Science and Information Systems*, Łódź, Poland, 13-16 Sept 2015, 2015, pp. 1717–1722. doi: 10.15439/2015F300

[40] F. Cicirelli, A. Furfaro, L. Nigro, and F. Pupo, "Modelling Java Concurrency: An Approach and a UPPAAL Library," in *FEDCSIS 2013 - Federated Conference on Computer Science and Information Systems*, Kraków, Poland, 8-11 Sept 2013, 2013, pp. 1373–1380, https://ieeexplore.ieee.org/document/6644196

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 437–440
DOI: 10.15439/2018F83
ISSN 2300-5963 ACSIS, Vol. 15

Job-shop scheduling with machine breakdown prediction under completion time constraint

Łukasz Sobaszek
Lublin University of Technology,
Nadbystrzycka 38 D,
20-618 Lublin, Poland
Email: l.sobaszek@pollub.pl

Arkadiusz Gola
Lublin University of Technology,
Nadbystrzycka 38 D,
20-618 Lublin, Poland
Email: a.gola@pollub.pl

Edward Kozłowski
Lublin University of Technology,
Nadbystrzycka 38 D,
20-618 Lublin, Poland
Email: e.kozlovski@pollub.pl

***Abstract* — This paper discusses the problem of time-constrained job-shop scheduling with technological machine breakdown prediction. The first section gives short characteristics of the field of research, and describes the effects of machine failure on job completion times. Secondly, the discussed problem is represented by means of mathematical equations and solved with original algorithms for machine failure prediction and implementation of redundant service times, and finally, the proposed solutions are verified by means of simulation. In the computational experiment stage a typical production case with taking into account 3 machines failure was considered. The last part of this paper draws conclusions from the study and presents directions of future research work.**

I. Introduction

Job scheduling has been receiving considerable attention of researchers [1]–[2]. Scheduling problems are approached from the perspective of production systems [3]–[4], dynamic character and randomness [5], time-dependence [6]–[7], and its relevance to industrial conditions [7]. In each of its aspects scheduling is governed by a variety of constraints that must be accounted for in scheduling.

We may distinguish two broad categories of constraints considered in scheduling: Resource-Constrained Scheduling and Time-Constrained Scheduling [8]–[9]. Although this division concerns mainly the field of project scheduling [10]–[11], it may be nonetheless found in scheduling of production as well [12]–[14].

Numerous constraints are discussed in the job-shop environment, yet they tend to be included as limiting constraints, for the sake of simplification of given scheduling problems [15]. Therefore, Robust Scheduling has been attracting an increasing amount of focus as it addresses the presence and the negative effect of various conditions and factors influencing the execution of production jobs [16].

II. Completion Time and Machine Failure Effect

Scheduling production jobs is most frequently considered under the objective function C_{max} – makespan/completion time of all jobs. This parameter is implemented in both analysis of test problems (aimed at evaluating the effectiveness of solutions) [17], as well as in determining job completion times [2], [6]. In practical industrial applications scheduling must strictly follow the C_{max} as exceeding these times and missing order delivery times may enforce contractual fines and lead to losing customers [2]. It is therefore a typical time constraint.

Apart from time constraints, there is a wide variety of factors that may only be addressed and resolved by implementing suitable methods and solutions. Scheduling Under Machine Failure is one of the most frequently considered factors, which is of considerable importance to observing order deadlines and therefore to time-constrained scheduling, as downtime of one technological machine may negatively affect the schedule through delaying jobs (Fig. 1).

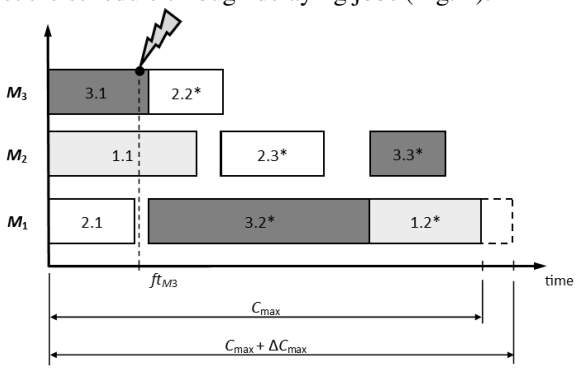


Fig. 1 M_3 machine failure and its impact on makespan (* – delayed jobs)

III. Problem Formulation

The job-shop scheduling problem consists in assigning jobs from the set of jobs $J = \{J_1, J_2,..., J_n\}$ to the set of available machines $M = \{M_1, M_2,..., M_m\}$ so that the schedule is optimised according to the objective function. Processing job J_j on machine M_i is described as operation. Simultaneously, we must consider the technology of processes, which is described by the matrix of machine orders $MO = [o_{ij}]$. Times of particular operations are described by the matrix of processing times $PT = [pt_{ij}]$. The size of matrices MO and PT is $m \times n$ [15].

This work was not supported by any organization

In order to implement the machine failure constraint, each machine must be described by means of the following sets:

- $FT_{Mi} = \{ft_{Mi1}, ft_{Mi2},..., ft_{Min}\}$, describing potential failure times of machines (expressed in hours),
- $P_{Mi} = \{p_{Mi1}, p_{Mi2},..., p_{Min}\}$ describing the probability of machine failure,
- $TB_{Mi} = \{tb_{Mi1}, tb_{Mi2},..., tb_{Min}\}$ defining time buffers (length of potential service times in minutes).

Moreover, implementing job completion time constraints requires that the schedule based on the set of feasible solutions accounts for the machine failure and adheres to the objective $C_{max} = C_{max} + \min(\Delta C_{max})$.

IV. Suggested problem solution

The solution we propose to the scheduling problem under time and machine failure constraints employs algorithms based on actual historical data, which are recorded in the sets of failure times $T_{Mi} = \{t_1, t_2,..., t_n\}$ and repair times $RT_{Mi} = \{rt_1, rt_2,..., rt_n\}$. Each set is specified individually for each machine.

A. Failure prediction algorithm

The first algorithm defines the key times in the schedule where failure may occur. Executed algorithm produces information regarding failure times (FT_{Mi}) and failure occurrence probability (P_{Mi}). The subsequent steps of the algorithm are as follows:

1. Define machine M_i and load data from set T_{Mi}.
2. Determine sequence and sort observations in an increasing order:
$$\{(t_i,d_i)\}_{1\le k\le n} \quad (1)$$
where: t_i – failure times,
d_i – number of occurrences.
3. Filter data – delete outliers.
4. Estimate the survival function based on:
$$\hat{S}(t) = \begin{cases} 1, & dla \quad t < t_1 \\ \prod_{t_i \le t} \frac{r_i - d_i}{r_i}, & dla \quad t_1 < t \end{cases} \quad (2)$$
where: r_i – the total number of failures expressed by:
$$r_i = \sum_{j=i}^{k} d_j \quad (3)$$
5. From the survival function determine failure times and failure probability (Fig. 2).

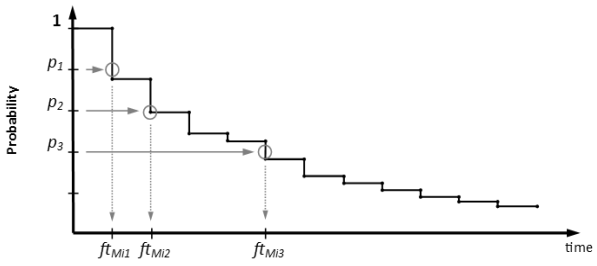


Fig. 2 Determining failure times from survival probability

6. Determine probability of failure:
$$p_{Mij} = 1 - p_i \quad (4)$$
7. Save data in sets FT_{Mi} and P_{Mi}.

Execution of this algorithm enables to pinpoint the times in the schedule where redundant time buffers should be implemented.

B. Algorithm for estimation and implementation of redundant time buffers

The second proposed algorithm determines the span of the service times and their implementation in order to make the schedule robust to failures. The algorithm is composed of the following steps:

1. Load observations from RT_{Mi} and sort in an increasing order.
2. Filter data – delete outliers.
3. Divide into two subsets according to:
$$RT_{Mi} = \{RT_{Mi1}, RT_{Mi2},\ldots, RT_{Mi8}\} \quad (5)$$
where: RT_{Mij} – subset of repair times, and:
$RT_{Mi1} = \{rt_i \in (0; 60>\}$
$RT_{Mi2} = \{rt_i \in (60; 120>\}$
…
$RT_{Mi8} = \{rt_i \in (420; 480>\}$
4. Define service buffer times – determine subset of RT_{Mi} of maximum weight – determine auxiliary sets RT'_{Mi} and TB'_{Mi}:
$$RT'_{Mi} = \left\{\overline{\overline{RT_{Mi1}}}, \overline{\overline{RT_{Mi2}}},\ldots,\overline{\overline{RT_{Mi8}}}\right\} \quad (6)$$
$$\underset{\max(RT'_{Mi})}{\vee} TB'_{Mi} = RT_{Mij} \quad (7)$$
$$TB'_{Mi} = \{rt_i \in \langle \min(RT_{Mij}); \max(RT_{Mij}) \rangle\} \quad (8)$$
where: RT_{Mij} – maximum weighted subset.
5. Estimate service time buffers:
$$TB_{Mi} = \{tb_{Mi1}, tb_{Mi2},\ldots, tb_{Min}\} \quad (9)$$
$$tb_{Mij} = \frac{\max(TB'_{Mi})}{n_p} \cdot j \quad 10)$$
where: j – number of element of set TB_{Mi},
n_p – number of considered levels of probability (elements of set P_{Mi}).
6. Take (p_{Mij}, ft_{Mij}) from sets FT_{Mi} and P_{Mi}.
7. Determine control probability:
$$p_{kMil} = \frac{l}{n_p} \quad (11)$$
where: $l = \langle 1,2,\ldots,n_p \rangle$.
8. Select buffer according to:

```
IF pMij ≤ pkMi1
   select buffer tbMi1 (minimal)
ELSE IF pMij > pkM1 and pMij ≤ pkMi2
   select buffer tbMi2 (where tbMi2 > tbMi1)
   ...
```

```
ELSE IF p_Mij > p_kMi(n-1) and p_mi ≤ p_kMin
   select buffer tb_Min (where tb_Min > tb_Mi(n-1)
END IF
```

9. Implement selected buffers of set TB_{Mi} in points defined by elements of set FT_{Mi} – in case buffers double – select higher.

The presented algorithm determines time buffers which make the schedule robust to failure of technological machines.

C. Schedule optimisation for makespan

As a result of execution of proposed algorithms the obtained schedule becomes robust. However, in order to meet the specified deadlines for orders, the most suitable schedule should be selected. Hence, the schedule selected from the group of solutions will be the one that offers the optimum solution to $\min(\Delta C_{\max})$ constraint. In the following case then, the best scheduling methods will be the exact ones and methods based on expert knowledge.

V. Simulation tests

The proposed solutions were verified by executing the scheduling process in job-shop conditions. The scheduling concerned 8 production orders processed on a stock of 5 machine tools. The processing times, the job order, and the number of jobs assigned to machines were randomly generated with available software, on the basis of the following assumptions:

- processing times cannot exceed one shift (maximum processing time is 7.5 h),
- machine loading is specified at 75%,
- machine routing is predetermined and not subject to change.

Moreover, the schedule allowed for the failure of 3 technological machines. The machine failure data was obtained from actual historical data of a production company. The source of information about machines failure times were electronic forms collected by maintenance department. In the data-treatment process authors used RStudio software with selected libraries. 3 failure probability levels were considered: $p_{Mi1} = 0.25$, $p_{Mi2} = 0.5$ and $p_{Mi3} = 0.75$. The results of the executed prediction algorithm are shown in Table 1.

TABLE I.
RESULTS OF MACHINE FAILURE PREDICTION ALGORITHM EXECUTION

Probability level	Predicted failure times [h]		
	M_1	M_2	M_3
$p_{Mi1} = 0.25$	$ft_{M11} = 8$	$ft_{M21} = 8$	$ft_{M31} = 8$
$p_{Mi2} = 0.50$	$ft_{M12} = 16$	$ft_{M22} = 16$	$ft_{M32} = 24$
$p_{Mi3} = 0.75$	$ft_{M13} = 40$	$ft_{M23} = 32$	$ft_{M33} = 48$

Execution of time buffer algorithm determined redundant service buffers and their implementation times (Table 2).

TABLE II.
DETERMINED SERVICE BUFFERS

Implementation time	Buffer length [min.]		
	M_1	M_2	M_3
after 8 h	$tb_{M11} = 20$	$tb_{M21} = 20$	$tb_{M31} = 20$
after 16 h	$tb_{M12} = 40$	$tb_{M22} = 40$	$tb_{M31} = 20$
after 24 h	$tb_{M11} = 20$	$tb_{M21} = 20$	$tb_{M32} = 40$
after 32 h	$tb_{M12} = 40$	$tb_{M23} = 60$	$tb_{M31} = 20$
after 40 h	$tb_{M13} = 60$	$tb_{M21} = 20$	$tb_{M31} = 20$
after 48 h	$tb_{M12} = 20$	$tb_{M22} = 40$	$tb_{M33} = 60$

The effectiveness of the proposed solutions was assessed by means of the following criteria:

- criterion $C_{\max}$,
- criterion y_J – the number of critical operations in reference to technology (operations of one job),
- criterion y_M – the number of critical operations in reference to machines.

The conducted verification tests implemented the following popular dispatching rules: FCFS, EDD, LPT and SPT. In the computational experiment stage LiSA software (*Library of Scheduling Algorithms*) was used. The tests were carried out on a typical PC-class computer. Subsequently, the obtained robust schedule was optimised under the specified objective function. The results of the calculations are shown in Tables 3–4.

TABLE III.
VALUES OF CRITERION C_{MAX}

Dispatching rules	C_{max} value [h]		
	Nominal schedule	Robust schedule	Optimized schedule
LPT	48	67	47.83
SPT	46	54.50	
FCFS	43.5	56.50	
EDD	49	63.17	

Implementation of service time buffers leads to delaying the makespan (Table 3). The mean delay is 14.77 h, which amounts to approx. 2 shifts. The delay in completion time of all jobs, however, increases the stability of production and indicates the feasible production completion time, simultaneously accounting for technological machine failure.

Determination of the actual production completion time allows preventing potential contractual fines. In the case of strictly specified order delivery deadline, implementation of the proposed algorithms and optimisation with the use of expert knowledge produces a robust algorithm under the specified completion time constraint. In the presented problem the schedule was optimised according to the defined objective function, in which case the makespan generally became slightly delayed, and in 2 cases was shortened; therefore the obtained schedules did meet the constraint $\min(\Delta C_{\max})$.

TABLE IV. VALUES OF CRITICAL OPERATIONS

Dispatching rules	Number of critical operations [–]					
	Nominal schedule		Robust schedule		Optimized schedule	
	y_J	y_M	y_J	y_M	y_J	y_M
LPT	17	20	3	12	4	12
SPT	11	17	6	9	5	10
FCFS	8	20	4	11	4	10
EDD	9	18	4	10	4	10

The robust schedule was characterised by the decreased number of critical operations – both in terms of processed jobs and operations on particular machines (Table 4). Both robust and optimised schedules showed a nearly 50% drop in the number of such operations, compared to the nominal schedule. The abovementioned leads to increasing the stability of production and reducing uncertainty and nervousness, as even if the failure of the machine should occur, the allocated service time absorbs the potential negative impact of failure on processing subsequent operations.

VI. CONCLUSION

It must be remarked that scheduling production jobs in industrial applications is inherently connected with a wide range of potentially disruptive factors, which ought to be treated as constraints. This paper presented scheduling under constraint of completion time of all jobs and machine failure. The proposed algorithms were verified in simulation, which proved their effectiveness and applicability. Future research works should be continued and could incorporate other factors occurring in job scheduling, such as alternating processing times, transport between workstations, *etc*). The proposed algorithms and methods could be implemented in the real production environments – at the managerial and production planning departments for instance.

REFERENCES

[1] M. T. Jensen, "Robust and Flexible Scheduling with Evolutionary Computation," Aarhus, 2001.

[2] Ł. Sobaszek, A. Gola, A. Świć, "Predictive scheduling as a part of intelligent job scheduling system," in Intelligent Systems in Production Engineering and Maintenance – ISPEM 2017: proceedings of the First International Conference on Intelligent Systems in Production Engineering and Maintenance ISPEM 2017, D. Mazurkiewicz, A. Burduk, Ed. Switzerland, 2018, pp. 358–367.

[3] J. Louis, Zhijie Xu, "Genetic Algorithms for Open Shop Scheduling and Re-Scheduling," Departament of Computer Science, University of Nevada, 1999.

[4] In-Chan Choi, Dae-Sik Choi, "A Local Search Algorithm for Jobshop Scheduling Problems with Alternative Operations and Sequence-Dependent Setups," Computers & Industrial Engineering, 42 (2002), pp. 43–58.

[5] P. Sharma, A. Jain, "Performance analysis of dispatching rules in a stochastic dynamic job shop manufacturing system with sequence-dependent setup times: Simulation approach," CIRP Journal of Manufacturing Science and Technology, Vol. 10, 2015, pp. 110–119.

[6] M. Pawlak, "Algorytmy ewolucyjne jako narzędzie harmonogramowania produkcji," Wydawnictwo Naukowe PWN, Warszawa, 1999.

[7] Yu. N. Sotskov, N. Yu. Sotskova, Lai T.-C., F. Werner, "Scheduling under Uncertainty – Theory and Algorithms," Belorusskaya nauka, Minsk, 2010.

[8] J. S. Norwood, "An evaluation of the time constrained and resource constrained scheduling features of commercially available project management software," Naval Postgraduate School, Monterey, California, 1996.

[9] M. Klimek, "Priority algorithms for the problem of financial optimisation of a multi stage project," Applied Computer Science, vol. 13, no. 4, 2017, pp. 20–34.

[10] P. Pekczynski, "Scheduling Constraints," Department of Computer Science, Saarland University, Germany, 2005.

[11] S. Van de Vonder, E. Demeulemeester, W. Herroelen, "Proactive Heuristic Procedures for Robust Project Scheduling: An Experimental Analysis," European Journal of Operational Research, 189 (2008), pp. 723–733.

[12] Hung-Kai Wang, Chen-Fu Chien, Che-Wei Chou, "An empirical study of bio manufacturing for the scheduling of hepatitis in vitro diagnostic device with constrained process time window," Computers & Industrial Engineering, Volume 114, 2017, pp. 31–44.

[13] A. M. Aguirre, L. G. Papageorgiou, "Resource-constrained formulation for production scheduling and maintenance," in Computer Aided Chemical Engineering, A. Espuña, M. Graells, L. Puigjaner, Ed. Elsevier, vol. 40, 2017, pp. 1375–1380.

[14] E. Gebennini, L. Zeppetella, A. Grassi, B. Rimini, "Production scheduling to optimize the product assortment in case of constrained capacity and customer-driven substitution," IFAC – PapersOnLine, vol. 48(3), 2015, pp. 1954–1959.

[15] Ł. Sobaszek, A. Gola, E. Kozłowski, "Application of survival function in robust scheduling of production jobs," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 575–578.

[16] Jian Xiong, Li-ning Xing, Ying-wu Chen, "Robust Scheduling for Multi-Objective Flexible Job-Shop Problems with Random Machine Breakdowns," International Journal of Production Economics, vol. 141(1), 2013, pp. 112–126.

[17] M. T. Jensen, T. K. Hansen, "Robust solutions to Job Shop problems," The 1999 Congress on Evolutionary Computation, July 1999, pp. 1138–1144.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 441–444
DOI: 10.15439/2018F143
ISSN 2300-5963 ACSIS, Vol. 15

Lecturers' competences configuration model for the timetabling problem

Jarosław Wikarek
Kielce University of Technology
Al. 1000-lecia PP 7, 25-314 Kielce, Poland,
Institute of Management and Control Systems
e-mail:j.wikarek@tu.kielce.pl

***Abstract*—The article presents the problem of academic teachers' competences configuration in the context of the university course timetabling problem (UCTP). Usually when solving UCTP, the set of available academic teachers and the competences they have is defined. The sets of lecture rooms, subjects (courses), student groups, time-slots, etc. are also known. Problems can emerge when it is not possible to find a satisfactory UTCP solution due to the missing sufficient number of specific academic teachers' (lecturers') competences, which is reflected in the possibility to teach specific classes (courses). In order to detect early such a situation and effectively manage the available and required lecturers' competences, a mathematical model of lecturers' competences configuration has been formulated in the form of a MILP (Mixed Integer Linear Programming) problem. Its solution has a direct impact on UTCP. The article also presents the implementation of the model in the LINGO solver environment and computational experiments.**

I. Introduction

University course timetabling problem (UCTP) is a problem in operation research, which raises interest in the communities both in the area of operational research (OR) and artificial intelligence (AI). In the most general manner, UCTP can be defined as the allocation of students and academic teachers (lecturers) to classes (courses) with consideration given to the available resources such as: lecture halls, laboratories and research workrooms. Additionally, the allocation must take place in appropriate periods (time-slots) that most often include a semester of teaching classes divided into single units (these are most often weeks). It is assumed that before starting work on an UCTP solution the number of all the resources and their availability are known; also, the characteristic features of the resources are known, such as: number of seats in the halls, number of students in particular student groups, the set of competences held by particular lecturers, whether a given room has a multimedia overhead projector, etc. With regard to the lecturers, each of them may have a certain set of competences, which is usually directly reflected in the list of courses which he or she may offer. Apart from the sets of resources with specific properties and time-slots, there are also numerous constraints for UCTP. The most important of them include the allocation of the lecturers and student groups to the halls, limited teaching load (teaching hours) per lecturer in a given semester, limited capacity of the lecture halls, laboratories and research workrooms, etc. Very often, there are also additional time restrictions related to preferences and availability of the lecturers as well as requirements for different courses. At the beginning of every semester, the dean, or the head of the department, faces the problem of an optimal solution to the UCTP problem.

Let's ask the question, what is going to happen, if it turns out, when solving UTCP, that it is impossible to find any acceptable solution due to a certain resource? The answer seems to be evident: the accessibility and/or the number of the given resource should be increased. Unfortunately, it is not always easy. It is particularly difficult with regard to the resource of lecturers' competences which are available when solving UCTP, as this requires them to complete additional training, studies, internships, and/or employment of new lecturers. The article proposes a model for the problem of managing and configuration such competences. On its basis an answer can be given to two key questions:

a. Do we have the appropriate resource of lecturers' competences, which will allow an UCTP solution to be found according to the rules (not exceeding the teaching load excessively)?
b. What and how many competences are missing to find an UCTP solution? What is the minimum number of the missing competences that guarantees an UCTP solution?

Fig. 1 presents the location of the problem of competences configuration in the context of UCTP. The main contribution of the presented research is a unique model for the problem of lecturers' competences configuration in the form of a MILP (Mixed Integer Linear Programming) problem [1]. The examined problem can be classified as a problem of resource distribution (teaching loads) with constraints, among others concerning the teaching load. Additionally, the implementation of the model in the MP solver environment and numerous computational experiments have been presented.

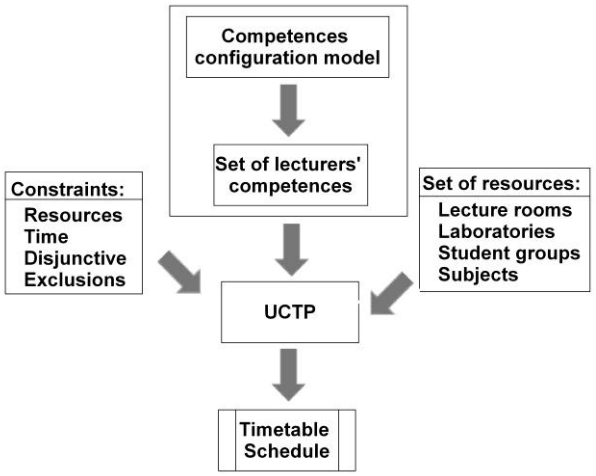


Fig. 1 Place of the competences configuration problem in the context of UCTP

II. LITERATURE REVIEW

For the first time the timetabling problem appears in [2] as a problem consisting of three sets of (i) teachers, (ii) classes and (iii) timeslots. Since then, the timetabling problem has been the subject of interest of many scientists and practitioners [3-6]. Over the years, many approaches have been developed to modeling and solving the principal variety of the problem, the so-called UCTP. The most important of them include: (a) operational research (OR) methods based on Integer/Linear programming (IP/ILP), Graph Coloring (GC), (b) methods and techniques of constraint programming (CP) and constraint logic programming (CLP), (c) metaheuristic methods, such as Case Base Reasoning method (CBR), Genetic Algorithms (GAs), Ant Colony Optimization (ACO), Partial Swarm Optimization (PSO), Variable Neighborhood Search (VNS), Tabu Search Algorithm (TS), etc., (d) hybrid methods and (e) multi-agent methods [7-10]. Practically, the competences configuration problem has not been considered in any of the presented approaches; it has been assumed that the set of lecturers with specific competences is given prior to finding an UCTP solution.

III. PROBLEM DECRIPTION

The competences configuration problem is discussed for the selected organizational unit of the university, which can be: chair, department or faculty. In the given organizational unit, lecturers are employed $E=\{e_1, \ldots, e_k, \ldots, e_{ZE}\}$ where ZE – number of lecturers employed in the unit. Each of the lecturers k has a certain teaching load allocated s_k i.e. the minimum number of hours to be realized in the given period (semester, academic year etc.). For instance, according to the valid law at Polish universities the teaching load is most often: 150, 210, 240, or 360 hours per academic year. In practice many lecturers teach courses in the number of hours exceeding their teaching load. For this reason, z_k coefficient has been introduced. If $z_k=1$, this means that lecturer k agrees to teach courses in the number of hours exceeding the teaching load (otherwise $zk=0$). *Wsp* coefficient has also been introduced, which determines by which percent a lecturer's teaching load can be exceeded without the need to obtain his or her consent (currently it is 15%). Certain types of courses are allocated to the given organizational unit (different forms for the given subject: lectures, projects, laboratory classes etc.) $P=\{p_1,\ldots, p^i, \ldots, p_{ZP}\}$ where ZP – number of types of courses in particular subjects assigned to the organizational unit. Each type of courses i has a specified number of hours l_i in which it is realized. In addition, for all classes the number of student groups h_i is defined. Lecturers of a given unit have certain qualifications (competences) to teach certain types of courses (coefficient $g_{i,k}=1$ means that lecturer k, without any further training, courses or postgraduate studies, etc. may offer subject i, otherwise $g_{i,k}=0$).

A. Illustrative example

In the example chair at a technical university, 18 different teaching courses are provided $P=\{p_1,\ldots, p_i, \ldots, p_{18}\}$. Each course has two characteristic parameters i.e.: the number of teaching hours in a semester and the number of student groups for which it is provided. The respective numerical data are presented in Table 1.

TABLE I.
DATA DESCRIBING THE SUBJECTS FOR THE ILLUSTRATIVE EXAMPLE.

courses	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Number of hours taught under the course	15	15	30	15	30	15	30	15	15
Number of student groups assigned to the course	1	2	1	2	1	2	1	2	1
courses	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}
Number of hours taught under the subject	15	30	15	30	15	30	30	30	15
Number of student groups assigned to the course	2	1	3	1	3	1	3	1	3

In the considered organizational unit, four lecturers are employed. Each of the lecturers has competences authorizing him or her to teach specific groups of courses, which have been specified in Table 2 (where "1" means having qualifications certifying the given competence, and "0" means that it is missing). The respective numerical data are presented in Table II.

TABLE II.
DATA DESCRIBING THE LECTURERS FOR THE ILLUSTRATIVE EXAMPLE.

courses / lecturers	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}	P_{18}
E_1	1	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
E_2	0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
E_3	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	1
E_4	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1

Each of the lecturers has an allocated teaching load in the number of 150 hours in a given semester and each of them

expresses his or her consent to teach excess hours (exceeding the teaching load).

For the so described problem, two primary questions come up:

Q1. Is there any allocation of the courses that will guarantee that all the lecturers realize the teaching load without the need to change competences?

Q2. If it is not possible to allocate the courses without changing the competences, how to minimize the number of new/changed competences (additionally, it can be required that change of the competences of any of the lecturers should not be greater than *B* times)?

For the so allocation problem, a mathematical model has been formulated (Chapter B), its implementation has been made in the LINGO solver language (Appendix A) and computational experiments have been conducted (Chapter C).

B. Competences management problem model

The mathematical model has been formulated in the form of a MILP problem. Table 3 presents the description of parameters, indexes and decision variables of the model. Table 4 presents the description of constraints. The function minimizing the number of changes in the competences of the lecturers has been adopted as the objective function (1). The above model allows you to find the answer to question Q2. In order to answer the Q1 question instead of (1) we introduce an additional restriction (10).

Table III.
DESCRIPTION OF PARAMETERS, INDEXES AND DECISION VARIABLES OF THE MODEL.

Symbol	Description
	Sets
P	Set of lecturers
E	Set of courses
Indexes	
k	Index $k \in P$
i	Index $i \in E$
	Parameters
l_i	Number of hours taught under the course *i*, $i \in E$
h_i	Number of student groups assigned to the course *i*, $i \in E$
s_k	Number of hours of the teaching load of lecturer $k \in P$
z_k	If the lecturer *k* has given their consent to overtime hours $z_k=1$, otherwise $z_k=0$
$g_{i,k}$	If lecturer *k*, without any further training, may teach courses *i* $g_{i,k}=1$, otherwise $g_{i,k}=0$, $i \in E$, $k \in P$
wsp	Percent exceeded teaching load without lecturer's necessary consent
A	Arbitrarily high constant
B	Maximum number of changes in competences for lecturer *k*
	Decision variables
$X_{i,k}$	Number of student groups for course *i* which will be taught by lecturer k $i \in E$, $k \in P$
$Y_{i,k}$	If the competences of lecturer *k* to teach course *i* have been changed $Y_{i,k}=1$ otherwise $Y_{i,k}=0$, $i \in E$, $k \in P$

$$\min \sum_{i \in E} \sum_{k \in P} Y_{i,k} \quad (1)$$

$$X_{i,k} \le (G_{i,k} + Y_{i,k}) \cdot A \forall i \in E, k \in P \quad (2)$$

$$\sum_{k \in P} X_{i,k} = h_i \forall i \in E \quad (3)$$

$$\sum_{i \in E} l_i \cdot X_{i,k} \ge s_k \forall k \in P \quad (4)$$

$$\sum_{i \in E} l_i \cdot X_{i,k} \le wsp \cdot s_k \forall k \in P : z_k = 0 \quad (5)$$

$$\sum_{i \in E} \sum_{k \in P} Y_{i,k} \le B \quad (6)$$

$$Y_{i,k} = 0 \forall i \in E, k \in P : g_{i,k} = 0 \quad (7)$$

$$X_{i,k} \in C^+ \forall i \in E, k \in P \quad (8)$$

$$Y_{i,k} \in \{0,1\} \forall i \in E, k \in P \quad (9)$$

$$Y_{i,k} = 0 \forall i \in E, k \in P \quad (10)$$

Table IV.
DESCRIPTION OF CONSTRAINTS AND OBJECTIVE FUNCTION

Constraints	Description
1	Minimum number of changed competences (Q2)
2	Only the employees having appropriate competences can teach the course which requires them
3	Each course is provided for the designated number of student groups.
4	The constraint ensures that the teaching load is realized
5	The constraint does not allow the teaching load to be exceeded by more than *wsp*, if there is no lecturer's consent
6	Change/extension of the competences may happen only a specified number of times;
7,8,9	Binarity and discreteness
10	No possibility to change competences (Q1).

C. Computational experiments

For the proposed model (Chapter 3.B) using the LINGO environment [11] and the implementation model (Appendix A), computational experiments have been conducted. The experiments have been concerned with the illustrative example (P1) and other organizational units such as: chair, department and faculty (P2..P6). The obtained results are presented in Table V and Table VI.

Table V.
RESULTS OF CALCULATIONS

Nr	Unit	ZP	ZE	V	C	T	Q1	Q2
P1	Chair 1	18	4	132	121	1	YES	0
P2	Chair 2	16	6	177	145	1	NO	1
P3	Depart. 1	40	15	1159	717	2	NO	3
P4	Depart. 2	60	20	2287	1267	3	NO	1
P5	Faculty 1	120	50	6170	6352	10	NO	4
P6	Faculty 2	160	80	13040	13290	25	NO	6

ZP Number of curses ZE Number of lecturers
V Number of decision variables C Number of constraints
T Calculation time

Detailed results for the illustrative example (P1) are presented in Table VI.

Table VI
DETAILED RESULTS OF CALCULATIONS FOR P1

k	i	$X_{i,k}$	$Y_{i,k}$	k	i	$X_{i,k}$	$Y_{i,k}$
1	1	1	0	3	5	1	0
1	4	1	0	3	6	2	0
1	11	1	0	3	7	1	0
1	12	3	0	3	15	1	0
1	16	2	1	3	16	1	0
2	2	2	0	4	8	2	0
2	3	1	0	4	9	1	0
2	4	1	0	4	10	2	0
2	13	1	0	4	17	1	0
2	14	3	0	4	18	3	0

IV. CONCLUSION

The work presents the problem concerning lecturers' competences configuration. A MILP model has been proposed for this problem, which has been implemented in the LINGO package environment. Before starting work on an UCTP solution, it is necessary to designate the set of the missing competences and supplement them. As a result of solving the problem of lecturers' competences configuration (Table V), we obtain information how many and what competences are missing, which lecturers should supplement them, and we obtain the allocation of the lecturers to the classes and the student groups, namely we partly/initially solve UCTP. The received allocation meets the constraints related to the teaching load for different lecturers. As part of further works, research is planned on modeling and solving UCTP integrated with the configuration model (Chapter B) and additional logic constraints are to be introduced [12,13] related to lecturers' preferences as to the times and forms of the curses, halls etc. It is also planned to apply the method of hybrid modeling and solving to the above-mentioned problem [14-16].

APPENDIX A

```
Model:
Sets:
 lecturers   /1..4/:s,z;
 subjects   /1..18/:l,h;
 pom_1 (subjects,lecturers):g,X,Y;
EndSets
Data:
 s= 150 150 150 150; z= 1 1 1 1;
 l= 15 15 30 15 30 15 30 15 15 15 30 15 30 15 30
 30 30 15;
 h= 1 2 1 2 1 2 1 2 1 2 1 3 1 3 1 3 1 3;
 g= 1 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 1
 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0
 0 1 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 1 1;
 wsp=1.25;  A=100; B=3;
EndData
Min=@sum(pom_1(i,k):Y(i,k));
 @for(pom_1(i,k): X(i,k)<=(G(i,k)+Y(i,k))*A);
 @for(subjects(i):
  @sum(lecturers(k):X(i,k))=h(i));
 @for(lecturers(k):
  @sum(subjects(i):l(i)*X(i,k))>=s(k));
 @for(lecturers(k)|z(k)#EQ#0:
   @sum(subjects(i):l(i)*X(i,k))<=s(k)*wsp);
 @sum(pom_1(i,k):Y(i,k))<=B;
 @for(pom_1(i,k):@GIN(X(i,k)));
 @for(pom_1(i,k):@BIN(Y(i,k)));
 @for(pom_1(i,k)|G(i,k)#EQ#1:Y(i,k)=0);
end
```

References

[1] A. Schrijver, “Theory of Linear and Integer Programming”. John Wiley & sons, ISBN:0- 471-98232-6, 1998.

[2] C.C. Gotlib, “The construction of class-teacher timetables”, *Proceedings of IFIP Congress*, 62, 1963, 73–77.

[3] R.J. Willemen, “School timetable construction: algorithms and complexity” *Printed by Universiteitsdrukkerij Technische Universiteit Eindhoven*, 2002, doi:10.6100/IR553569.

[4] J.H. Kingston, “Timetable Construction: The Algorithms and Complexity Perspective”, Ann Oper Res, 2014, 218-249, https://doi.org/10.1007/s10479-012-1160-z

[5] M. Ilic, P. Spalevic, S. Ilic, M. Veinovic, Z. Milivojevic, B. Prlincevic: “Data mining techniques for student timetable optimization”, *INFOTEH-JAHORINA* Vol. 14, March 2015, 578-583.

[6] H. Babaei, J. Karimpour, A. Hadidi, “A survey of approaches for university course timetabling problem”, *Computers & Industrial Engineering 86*, 2015, 43–59.

[7] T.A. Redl, “A study of university timetabling that blends graph coloring with the satisfaction of various essential and preferential conditions” Ph.D. Thesis, Rice University, Houston, 2004, Texas.

[8] H. Asmui, E.K. Burke, J.M Garibaldi, “Fuzzy multiple heuristic ordering for course timetabling”, *The proceedings of the 5th United Kingdom workshop on computational intelligence (UKCI05)*, London, UK, 2005, 302–309.

[9] S. Abdullah, E.K. Burke, B. McColloum, “Using a randomised iterative improvement algorithm with composite neighborhood structures for university course timetabling. metaheuristic – Program in complex systems”, *Optimization*, 2007, 153–172.

[10] M. Mühlenthaler, “Fairness in Academic Course Timetabling”, *Lecture Notes in Economics and Mathematical Systems 678*, Springer International Publishing, Switzerland, 2015, doi:10.1007/978-3-319-12799-6_2.

[11] Lindo, http://www.lindo.com/, Accessed May 04 2018.

[12] P. Sitek, I.E Nielsen J. Wikarek, “A Hybrid Multi-agent Approach to the Solving Supply Chain Problems”, Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, Volume 35, 2014, 1557-1566,.doi:https://doi.org/10.1016/j.procs.2014.08.239

[13] P. Sitek, “A Hybrid Approach to the Two-Echelon Capacitated Vehicle Routing Problem (2E-CVRP)”, *Automation, Robotics and Measuring Techniques. Advances in Intelligent Systems and Computing*, 267. Springer, Cham, 2014, 251-263, doi:https://doi.org/10.1007/978-3-319-05353-0_25.

[14] P. Sitek, J. Wikarek, “A multi-level approach to ubiquitous modeling and solving constraints in combinatorial optimization problems in production and distribution”, *Applied Intelligence 48*, 2018, 1344–1367, doi:https://doi.org/10.1007/s10489-017-1107-9

[15] P. Sitek, J. Wikarek, P. Nielsen, “A constraint-driven approach to food supply chain management”, *Industrial Management & Data Systems 117(9)*, 2017, 2115-2138, doi: 10.1108/IMDS-10-2016-0465/

[16] P. Sitek, "A hybrid multi-objective programming framework for modeling and optimization of supply chain problems", *2015 Federated Conference on Computer Science and Information Systems (FedCSIS),* 2015, 1631-1640, doi: https://doi.org/10.15439/2015F83

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 445–453
DOI: 10.15439/2018F131
ISSN 2300-5963 ACSIS, Vol. 15

Generation of Synthetic Business Process Traces using Constraint Programming

Piotr Wiśniewski, Krzysztof Kluza, Antoni Ligęza
AGH University of Science and Technology
al. A. Mickiewicza 30, 30-059 Krakow, Poland
E-mail: {wpiotr,kluza,ligeza}@agh.edu.pl

Anna Suchenia
Cracow University of Technology
ul. Warszawska 24, 31-155 Kraków, Poland
Email: asuchenia@pk.edu.pl

***Abstract*—Juxtapositioning manually created business process models with diagrams generated using process discovery algorithms exposes high complexity of the latter. As a consequence, their formal verification requires significant computational resources due to a large state space. Nevertheless, an analysis of the generated model is needed to assure its correctness and the ability to represent source data. As a solution to this problem, we present an approach for constraint-based generation of a complete workflow log for a given BPMN model. In this paper, we propose a method to extract directed subgraphs representing token flows in the process together with a set of predefined constraints. Likewise, in the case of process simulation, these constraints ensure the correctness of the generated traces. Ultimately, the obtained results can be compared to the original workflow log used for process discovery in order to verify the obtained model.**

***Index Terms*—business process management, process verification, workflow logs, constraint programming**

I. Introduction

BUSINESS process models aim to represent knowledge about workflows which take place in an organization. Such chains of different activities are represented by graphical diagrams that hold information about their dependencies, execution conditions, alternative flows and other properties included in the used modeling notation. Creation of a process model may be performed manually by a process designer who builds the diagram in a graphical editor or prepares other representations such as UML activity diagram [1], structured text [2], natural text description [3] or a spreadsheet representation [4]. Another way to design a model is to discover it from event logs generated by existing IT systems [5], [6].

Although process mining appears to be a convenient technique which does not require much effort in the phase of process modeling, the discovered workflow models can suffer from various defects. Such flaws can be caused either by the imperfection of the selected algorithm [7] or by corrupted logs [8]. Raw data recorded from real system may be characterized by missing events, imprecise or incorrect data, as well as irrelevant artifacts.

Several methods were developed to evaluate discovered models. They include simulations performed on a generated Petri Net which represents the workflow [9], application of evaluation metrics [10], as well as validating models against temporal logic formulae [11]. In order to improve data analyzed by a process miner, the technique of real-time log monitoring can be applied [12]. Another related approach consists in generating synthetic traces based on a structured declarative model [13].

The existing methods can be used to improve quality of process mining from various perspectives, however they tend to operate on different workflow representations, such as raw event data or Petri Nets, without considering the output model. The purpose of the proposed method is to evaluate the created business process model and verify its behavior by generating its admissible execution traces. The synthetic log can be then compared to the set of real execution sequences obtained from a computer system in order to indicate areas of imperfection and take corrective actions, such as changing the mining algorithm or refining log data. A schematic illustration of our approach is presented in Figure 1.

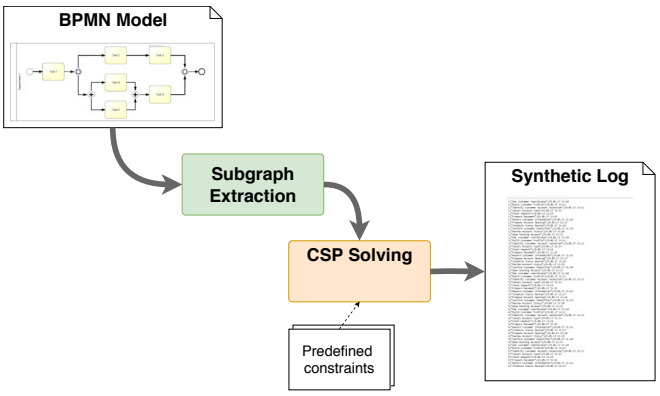


Figure 1. Overview of the proposed approach.

This paper is organized as follows: Section II provides an overview of Business Process Management which includes its definition and phases. Section III describes the concept of a BPM lifecycle and its aspects. A definition of workflow log was presented in Section IV and followed by a brief description of process mining, where logs are used to extract process knowledge, which was provided in Section V. Our main contribution is included in Sections VI and VII where we proposed methods to determine the number of traces in a process log, as well as presented a constraint-based model to generate a synthetic set of traces. Concluding remarks and plans for future works were summarized in Section IX.

II. Business Process Management

Business Process Management (BPM) [14] is a modern approach to improving organization's workflow, which focuses on reengineering of processes to obtain optimization of procedures, increase efficiency and effectiveness by constant process improvement.

The key aspect of BPM is a Business Process (BP). Although there is no single definition of a Business Process, the existing definitions have many things in common [15], [16], [17], [18]. A BP is usually described as a collection of related activities which transform different kinds of clearly specified inputs to produce a customer value, mainly considered as products or services and organizational goals, as output.

Different definitions emphasize various aspects of such defined processes. Davenport described a process highlighting the importance of producing an output for a customer – how work is done [15]. Definition of Eriksson and Penker emphasizes how work is performed rather than describing products or services, results of a process [19]. Jacobson, in turn, underlined that a process should be customer-oriented and meet an individual customer's needs [20]. A wider conceptualization of process was presented by Melao and Pidd [21]. They gave four perspectives on business process to understand BPs more fully. In their approach, BPs can be seen as either deterministic machines, complex dynamic systems, interacting feedback loops or social constructs.

Thus, Business Process Management requires a specification of many aspects, such as goals, inputs, outputs, used resources, activities and their order, impact to other organizational units, customers and owners for each of managed processes to enable real benefits. It unifies the previously distinct disciplines such as Process Modeling, Simulation, Workflow, Enterprise Application Integration (EAI), and Business-to-Business (B2B) integration into a single standard [22].

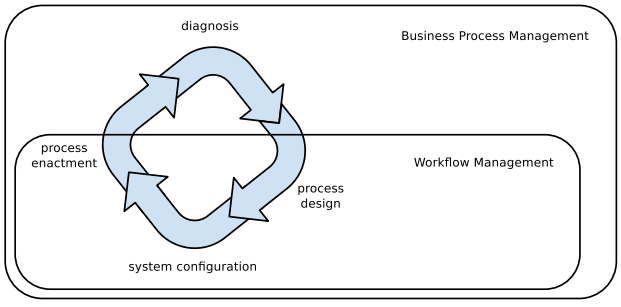


Figure 2. Comparison of Workflow Management and Business Process Management (based on [23]).

Therefore, BPM is often considered as either a legacy or the next step after workflows. Workflow Management Coalition (WfMC) [24] defines a workflow [25] in terms of automation of a business process during which documents, information or tasks are passed from one participant to another for action, according to a defined set of procedural rules. According to van der Aalst et al. [23], BPM is a broader term than Workflow Management (WFM). BPM supports business processes using methods, techniques, and software in designing, enacting, controlling, and analyzing processes involving humans, organizations, applications, documents and other information sources. It is restricted to operational processes, thus it excludes processes that cannot be made explicit.

A simple approach to process management distinguishes four phases of supporting processes [23]:

1) *process design*, in which the process is designed or redesigned,
2) *system configuration*, in which the design is implemented by configuring process management system,
3) *process enactment*, in which the operational business process is executed using the configured system,
4) *diagnosis*, in which the process is analyzed or verified to identify things that can be improved.

The relationship between WFM and BPM is presented in Figure 2. As one can observe, BPM extends the traditional WFM approach. In the case of the WFM systems, they do not support diagnosis phase, and such features as simulation, verification or validation of process designs.

III. BPM Lifecycle

Although many aspects of BPM have been debated in literature, one of the fundamental BPM issues is a repeating sequence of steps, the so-called Business Process Management Lifecycle (see Figure 3). The main idea behind the BPM lifecycle is to manage and improve BPs over business changes. Due to the use of clearly defined phases, BPM enables the continuous maintenance and the evolution of processes. During iterations, such parameters of business processes as cost, time, quality of output or customer satisfaction can be improved causing an improvement of the whole process.

Thus, BPM is in fact the application of the management cycle to organization's business processes [26]. The BPM lifecycle starts with specification of organizational and process goals as well as an assessment of environmental factors having an effect on the organization BPs. In the next process design phase, the organization processes are to be identified or redesigned. In this phase, the particular process details should be specified, and the proper variables that will influence the process design should be identified as well. During the next phase the previously specified process models are implemented in the environment, usually manually via procedure handbooks or using BPM or workflow software. Finally, the implemented process can be instantiated and executed. During execution, the performance is monitored in order to control and improve the process. Data produced during the process enactment and monitoring phases, aggregated from multiple process instances, can be used in the evaluation phase, whose purpose is to formulate the results suitable for process improvement.

Our area of research focuses on analyzing process models discovered from event logs and verifying them in terms of admissible execution sequences. Therefore it covers the evaluation phase and its side activities such as auditing event logs as well as providing measures of improvement for the whole process.

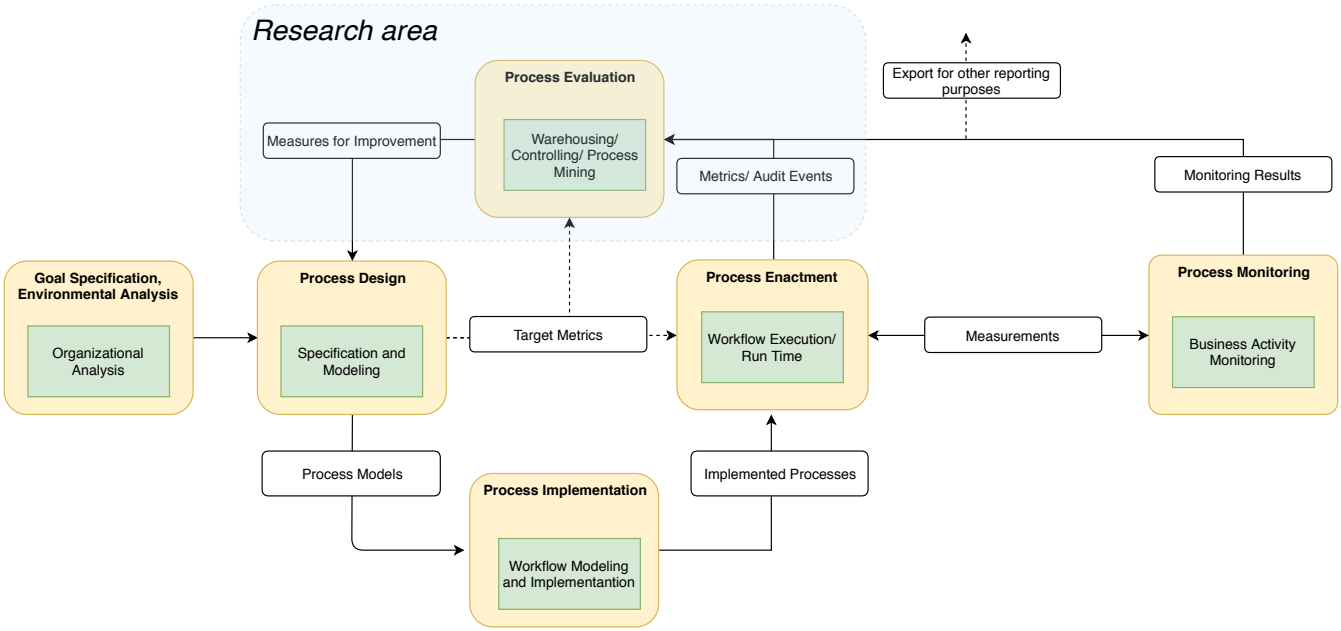


Figure 3. Business Process Management Lifecycle with the indication of our area of research (based on [26]).

IV. Workflow Log

In every business process, regardless whether it is executed manually or by an IT system, completion of each activity should be recorded with a proper timestamp. Such a record is often referred to as a log event [27] and may include information about a person or unit performing the task, as well as its cost and used resources. A set of log events ordered by their completion timestamps is called a workflow trace:

$$\sigma = \{\alpha^{(1)}, \alpha^{(2)}, ...\alpha^{(K)}\}, \tag{1}$$

where α is a log event and K represents the length of the trace.

One of the most important features of an event trace is that all recorded activities are ordered chronologically. In other words, even if two different activities are executed in parallel, their accomplishment time differs and it is always possible to distinguish the one which was completed first. However, as business processes are repeated many times, the order of their completion may differ depending on the instance of the process. In addition, processes may contain alternative gateways that, based on a logical condition, determine which task should be executed and which should remain unused. Therefore, in order to gather the information about the whole process, one should record a workflow trace for a number of times to ensure, that all or nearly all the possible execution sequences were collected. A set of workflow traces is called a workflow log:

$$W = \{\sigma_1, \sigma_2, ...\sigma_L\}, \tag{2}$$

where σ_i are separate workflow traces, also referred to as cases or workflow instances and L is the number of recorded traces.

A workflow log can be considered complete if it covers all the possible execution sequences of the process. In the case when activities are executed in loops, the number of possible traces may be infinite. Therefore, we weaken this requirement to a notion of sufficient completeness explained in Definition 1. It limits the required number of traces to those where the number of repetitions for each log event is equal to the number of cycles which include the corresponding activity.

Definition 1. (Sufficiently complete workflow log) Let G_P be a business process graph [28] representing the analyzed process and S_C be a set of all simple cycles in G_P. Function $C_C(\tau)$ determines the number of occurrences of the vertex representing activity τ in S_C. Workflow log W is sufficiently complete if it contains all the possible execution sequences where the number of occurrences for each activity τ is lower than or equal to $C_C(\tau) + 1$.

V. Process Mining

Process mining is an area of research which focuses on extracting knowledge from event logs [29] which were described in details in Section IV. One of the challenging tasks within process mining is process discovery [30] which includes algorithms able to generate process models in a flexible way, and in some cases without the need of any human actions. Although process discovery methods can produce syntactically correct workflow nets, the result is a general process model which is not directly applicable for execution in a runtime BPMN environment. BPMN diagrams can be obtained directly from event logs [31]. However, they require significant enhancements to be suitable for execution. Such modifications can be based on decision mining [32] which extends the process model by providing conditions for alternative or exclusive gateways.

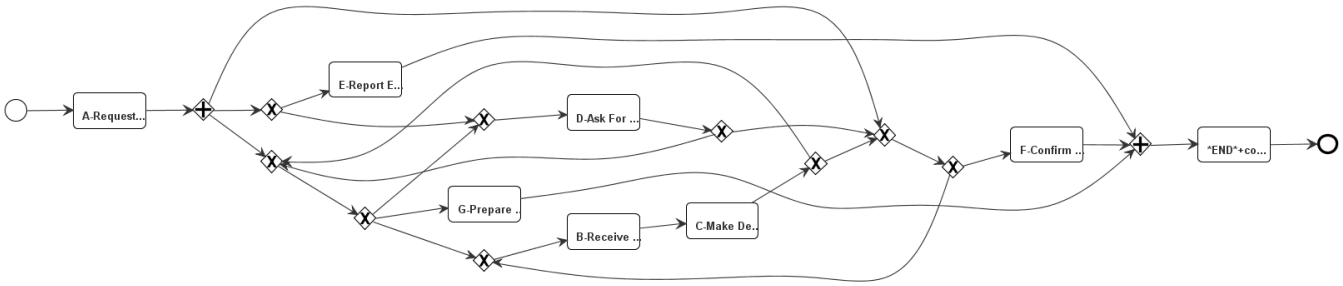


Figure 4. Result of applying a process mining algorithm to a workflow log of 36 traces.

Regarding process execution, the generated model still needs to be validated in terms of structural anomalies which may result in wrong dynamic behavior of a process [33].

Figure 4 presents a BPMN model discovered using ILP Miner [34] based on 36 different execution traces. In this example, the number of parallel and exclusive gateways (11) is higher than the number of activities (8). This implies a large number of routings and may possibly result in various exceptions during execution. Several complexity metrics exist which evaluate the complication level of business process models [35]. From a runtime point of view, the control-flow perspective should be considered. Cardoso et al. [36] propose a set of such metrics of which two can be applied in this case:

- Control Flow Complexity (CFC) which can be calculated as a sum of states induced by all the split gateways. Given n_{out} as the number of outputs of a gateway, each exclusive (XOR) split induces n_{out} states while a parallel split corresponds only to one state, as all the output branches are always used.
- Coefficient of Network Complexity (CNC) which is a quotient of the number of arcs (sequence flows) and the number of all activities, joins and splits in the process.

In the example presented in Figure 4, the Control Flow Complexity is equal to 12 and the Coefficient of Network Complexity has a value of 1.47. It is worth noting that values of these metrics for a simple workflow without any branching elements are equal to 0 and $(n_a + 1)/n_a$, where n_a is the number of activities, respectively.

VI. Determining the Number of Traces in a Process

The first step towards the estimation of the number of distinct traces is based on a business process model. A sequential workflow consisting of one start event, one end event and no gateways produces only one trace. A trivial example of such a process model is shown in Figure 5. However, models representing a simple workflow are rarely used in practice. According to the survey [37], in 90% of BPMN models the number of gateways varies between 5 and 15.

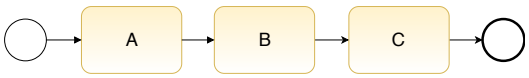


Figure 5. Simple sequential workflow.

In order to analyze business process model from an execution perspective, its token flow must be considered. It is assumed that the start event of a process generates a token which runs through the whole workflow to be consumed by an end event [38]. Although business processes can contain multiple start events, it is not regarded as a good practice as it remains unclear in the BPMN specification whether all the start events should occur before the process execution [39]. When determining a sufficiently complete log we refer to best practices of process modeling [40] following the statement that a well designed process model should contain only one start event which creates exactly one token.

On the other hand, multiple end events are a common practice in business process models, as they may represent different final states (e.g. goal and error states). However, only one of the end events is triggered in a single process instance. It may consume one or more tokens, depending on the number of incoming sequence flows.

Token flow in a business process is managed by logical gateways which determine branching flows. A single token created at the beginning of each process instance can be processed differently depending on the type of a gateway. The following actions are possible for a split gateway with n output branches:

- An exclusive (XOR) gateway places the token in one of the output sequence flows where the corresponding condition is met. As a result n different actions are possible.
- An inclusive (OR) gateway multiplies the incoming token by the number of conditions met. This action is followed by placing the created tokens on the corresponding sequence flows. As at least one output branch must be active, 2^{n-1} actions are possible.
- A parallel (AND) gateway multiplies the incoming token by the number of output branches. This result in only one possible action.

Since a single sequence flow should not transfer more than one token at a time, join gateways are responsible for merging multiple sequence flows into one output branch. An exclusive merge gateway only receives a token from one of its incoming branches and passes it to its output. A more complex situation occurs in case of a synchronization of multiple sequence flows which may be done in the following way:

- an inclusive merge consumes all the tokens created by its corresponding split gateways,
- a parallel merge consumes the tokens for all its input branches.

Each of the synchronizations result in creation of a token and passing it to the outgoing sequence flow except the situation when a parallel merge is declared implicitly by multiple sequence flows leading to an end event.

Since according to the best practices OR gateways should be avoided in BPMN modeling [40] and regarding the fact that in most cases they can be replaced by a sequence of exclusive and parallel gateways [41], this type of routing object was not taken into further analysis. As a consequence, the only flow objects in a well modeled process where token creation or consumption occurs are parallel gateways.

Figure 6 represents a simple BPMN model with one parallel and one exclusive gateway. As stated before in this section, an XOR split gateway generates as many possible states as is the number of its outgoing sequence flows. Thus, in this case there will be two possible execution sequences of the SESE (Single Entry Single Exit) block determined by two exclusive gateways: one sub-trace consisting of task D and one empty sequence. Although the presence of a single AND gateway induces one state, the generated tokens represent separate sequences of activities. As a result they can be executed independently until they reach a synchronization element represented by a parallel merge. Therefore, the first SESE block which follows the start event may also be executed in two ways, namely $\{A, B\}$ and $\{B, A\}$. Besides these two process fragments the remaining activity, denoted as C is executed in every process instance. This analysis leads to a conclusion that the number of execution traces in such a process is a multiplication of the corresponding values calculated for its SESE blocks.

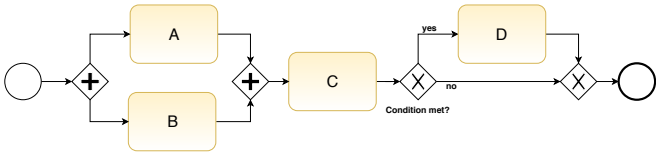


Figure 6. Example process with two basic types of gateways.

As stated before in this section, sequence flows followed by parallel gateways are executed independently. Therefore, if each branch within a SESE block delimited by AND gateways contains exactly one activity, these activities can be executed in any order. The example presented in Figure 6 illustrates a general rule for determining the number of admissible traces in a BPMN model which is expressed in Theorem 1.

Theorem 1. Let $\mathbb{P}$ be a BPMN process model containing a set of exclusive split gateways $\mathbb{G}_{XOR} = \{g_{X1}, g_{X2}, ...g_{Xk}\}$, a set of parallel split gateways $\mathbb{G}_{AND} = \{g_{A1}, g_{A2}, ...g_{Al}\}$ and two corresponding sets of merge gateways, namely $\mathbb{M}_{XOR}$ and $\mathbb{M}_{AND}$. If $\mathbb{P}$ consists of k SESE blocks determined by XOR gateways and l SESE blocks determined by AND gateways where each of l sequence flows contains exactly one activity, then the number of sequence flows in sufficiently complete workflow log W_{SC} can be expressed by Formula 3.

$$|W_{SC}| = \prod_{i=1}^{k} n(g_{Xi}) \cdot \prod_{i=1}^{l} n(g_{Ai})! \tag{3}$$

where $n(g)$ determines the number of outgoing sequence flows of a split gateway.

Proof. Let us assume that each of k exclusive gateways in $\mathbb{P}$ has exactly n outgoing flows. Knowing that every SESE element in a process model can be reduced to a subprocess which is a single BPMN activity [28] and that every XOR gateway allows for n actions, there are n possible states of every k subprocess in $\mathbb{P}$. This implies that the number of all admissible states is equal to n^k. Since the number of outgoing flows may vary for different gateways, $n(g_{Xi})$ has to be multiplied k times. □

VII. Constraint-Based Log Generation

The method presented in Section VI refers to these BPMN diagrams where SESE gateway structures are easily distinguishable. However, in automatically generated process models, as shown in Figure 4, gateways can be nested and the number of merge gateways does not have to match the number of splits. As a result, the number of traces is hardly calculable using analytical methods. Figure 7 shows a simple process model with nested gateways where Formula 3 cannot be applied.

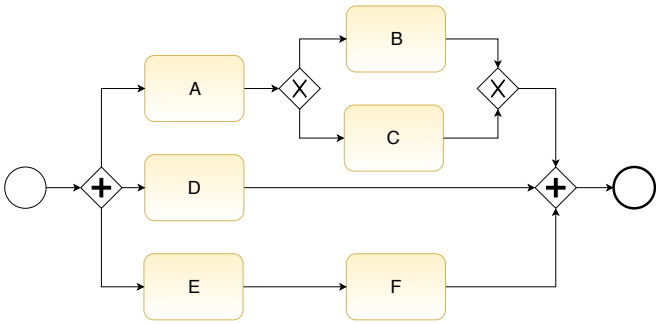


Figure 7. Example process with nested gateway structures.

The BPMN model presented in Figure 7 consist of six tasks formed in a parallel block with three branches. In business process models without loops, the upper limit of the number of traces is equal to the factorial of the number of activities. This would occur if all the tasks in the process were executed independently. In this case, however, there are following constraints which limit the set of possible sequences:

1) A must occur before B or C.
2) B is executed if and only if C is not.
3) E must occur before F.

In the analyzed example such constraints are easily identifiable and they can be expressed in a temporal logic [42]. However, dynamic generation of constraints for a complex

process model requires checking of all the dependencies between each pair of tasks which is an exponential problem. As a solution to this issue, we propose to extract process subgraphs from the model in a such manner that each of the graphs will represent the flow of a single token. It was stated in Section VI that each process instance ends with a single end event. As a consequence, all the generated tokens have to be consumed either by one of the parallel join gateways or by an event. Thus, the method can be applied to those SESE blocks in the process which start with a parallel split and whose last flow object is either a parallel join or an end event with multiple incoming sequence flows.

Let us denote the set of subgraphs as $S_G = \{s_1, s_2, ...s_q\}$. In the model presented in Figure 7, there are three tokens generated at a parallel split gateway and all of them are synchronized by a parallel merge before the end event. As a result, the extraction will provide three subgraphs, each representing tasks on a single branch (see Figure 8).

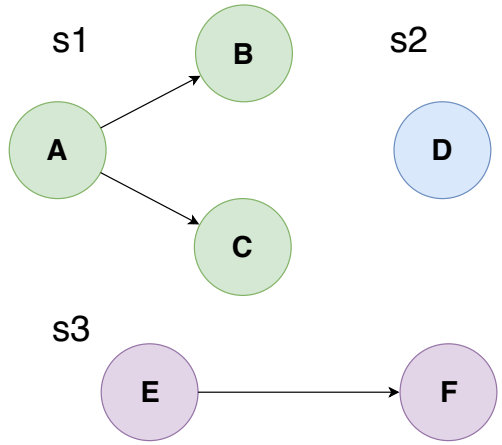


Figure 8. Token flow subgraphs extracted from a business process model.

In order to determine admissible workflow traces, a path should be calculated in every subgraph. The beginning of such a path is a vertex with no incoming edges. A path should end if a vertex with an outdegree equal to zero is reached. In the next step, all these paths are merged to vector β of length $p \cdot q$ where p is the overall number of tasks in the analyzed SESE block and q is the number of extracted subgraphs.

In Constraint Satisfaction Problems (CSP), a state which represents a set of elements must satisfy a collection of finite constraints over variables [43], [44]. CSPs can be solved using the constraint programming technique whose applications include design and modeling [45], as well as planning and scheduling [46]. If the problem is not over-constrained, a CSP algorithm always finds all the admissible solutions for finite domains [44]. Solving a Constraint Satisfaction Problem consists in general of four following steps:

1) Ordering the decision variables according to preference criteria.
2) Assigning values to each of the variables with respect to their domains.
3) Verifying if any of the constraints is violated at any step of the solving process.
4) If a complete or partial constraint-violating assignment is found, backtracking is enforced, and a succeeding set of values is assigned.

In this case, to calculate a process trace, we developed a constraint-based model in MiniZinc environment which consists of the following core elements:

1) Input data:
   - a list of tasks in the analyzed block,
   - vector e_x of maximum numbers of executions for each task (1 by default),
   - adjacency matrices $A_1, ...A_q$ for each subgraph.
2) Decision variables:
   - a subgraph trace matrix S_t of size $q \times p$,
   - a vector of merged traces β,
   - a vector γ of size p representing a workflow trace.

In order to improve the understandability of the code, we define two custom predicates which are further used in the predefined constraints:

- connected – returns true if two tasks are connected in the subgraph represented by its adjacency matrix,

```
predicate connected(src, dest, adj)
    = (adj[src, dest] == 1);
```

- same_block – returns true if a pair of indices of vector β is in the same subgraph trace.

```
predicate same_block(idx1, idx2)
    = (idx1 > 0 /\ idx2 > 0
        /\ idx1 div p == idx2 div p
    );
```

Constraints included in the model can be divided into three main groups, depending on a decision variable to which they are related. Let us briefly present these constraints along with their simplified representations in the MiniZinc language:

1) Subgraph traces:
   - the count of occurrences for each task should be lower then or equal to the input value,

```
forall(i in 1..p, j in 1..q)(
    count_geq(row(S_t,j), i, e_x[i])
);
```

   - the value 0 in the event log represents an idle task,

```
idle_task = 0;
```

   - the last log event should not be preceded by an idle task,

```
forall(i in 1..q)(
    count_neq(row(S_t,i), idle_task,
      last_process_index+1)
);
```

- all tasks after the last log event should be idle,

```
forall(i in 1..p, j in 1..q)(
    if i > last_task_index[j]
    then S_t[j,i] == idle_task
    else S_t[j,i] != idle_task
    endif
);
```

- the first element of a trace should be represented by a subgraph vertex without any incoming edges,

```
forall(i in 1..p, j in 1..q)(
  if S_t[j,1] == i
  then count(row(A_j, i),-1,0)
  endif
);
```

- the last log event should be represented as a vertex without outgoing edges,

```
forall(i in 1..p, j in 1..q)(
  if s_t[j,1] == i
  then count(row(A_j, i),1,0)
  endif
);
```

- if one task directly follows another in a trace, then it is connected by a directed edge in the corresponding subgraph.

```
forall(i in 1..q, j in
  1..last_process_index) (
  connected(S_t[i,j],S_t[i,j+1], A_i)
    \/ (S_t[i,j] == 0
        /\ S_t[i,j+1] == 0)
    \/ (j == last_task_index[1]
        /\ S_t[i,j+1] == 0)
);
```

2) Merged vector β:
   - the vector β is a concatenation of trace matrix rows.

```
forall(i in 1..last_process_index, j
  in 1..q)(
  if i < last_task_index[j]
  then beta[(j-1)*last_process_index
    + i] = S_t[j,i]
  else beta[(j-1)*last_process_index
    + i] = 0
  endif
);
```

3) Final trace vector:
   - the vector γ holds non-zero indices of β,

```
forall(i in gamma_indices)(
  if gamma[i] > 0 then
    beta[gamma[i]] != 0
  else beta[gamma[i]] == 0
  endif
);
```

- all the elements of γ are different except zero,

```
alldifferent_except_0(gamma);
```

- all the elements following the last non-zero element of γ are equal to zero,

```
forall(i in gamma_indices)(
  if i > last_gamma_index then
    gamma[i] == 0
  else gamma[i] != 0
  endif
);
```

- for each pair of tasks in γ if the pair is in the same row of matrix S_t then these task should be ordered in the same way as in S_t.

```
forall(i in gamma_indices, j in
  gamma_indices)(
  if i != j /\ gamma[i] != 0 /\
    gamma[j] != 0 /\
    same_block(gamma[i], gamma[j])
    then
    if i > j then
      gamma[i] > gamma[j]
    else
      gamma[i] < gamma[j]
    endif
  endif
);
```

In order to run the MiniZinc solver two files are needed:

- the model file *trace_id.mzn* which contains definitions of decision variables, predicates and constraints,
- the data file *subgraphs.dzn* where activity names, their maximum number of executions and subgraph adjacency matrices are defined.

For the workflow trace generation the search goal should be set for constraint satisfaction by using the statement `solve satisfy`.

The analyzed example model contains 6 tasks and 3 subgraphs, then $p = 6$ and $q = 3$. Formula 4 presents an example subgraph trace matrix S_t for subgraphs shown in Figure 8. Creation of vector β consists in merging all the subgraph traces (see Formula 5). Indices of β are used to generate trace vector γ (see Formula 6).

$$S_t = \begin{bmatrix} A & B & 0 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 0 \\ E & F & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4}$$

$$\beta = [A\ B\ 0\ 0\ 0\ 0\ D\ 0\ 0\ 0\ 0\ 0\ E\ F\ 0\ 0\ 0\ 0] \tag{5}$$

$$\gamma = [13\ 7\ 1\ 2\ 14\ 0] \tag{6}$$

The resulting trace is an ordered set (see Formula 1) determined by non-zero elements of β whose indices are ordered by values of γ. Its value for the running example represented by the BPMN model in Figure 7 and trace matrix S_t defined by Formula 4 is shown in Formula 7.

$$\sigma = \{E, D, A, B, F\} \tag{7}$$

In order to generate all the admissible task sequences the solver should be set to print all solutions. Execution of the model results in a sufficiently complete log of a SESE process block. If a process contains multiple token split blocks then all of them should be handled separately.

VIII. Log-Based Verification of Process Models

Several metrics were proposed to analyze results of process discovery algorithms, namely: replay fitness, simplicity, precision, and generalization [47]. Since we tend to compare two sufficiently complete workflow logs without analyzing the graphical layout of the generated model, the following quality measures have been considered and adopted for the purpose of log comparison:

- model fitness – the percentage of traces from the original log which were generated based on the discovered model,
- execution precision – the percentage of generated workflow traces that are allowed in the original log.

It is worth noting that values for both metrics should be calculated in order to validate the resulting model. To illustrate this problem, let us analyze the example shown in Figure 6. Its original complete log can be easily determined analytically:

$$\begin{aligned} W_C = \{&\{A, B, C\}, \{B, A, C\}, \\ &\{A, B, C, D\}, \{B, A, C, D\}\}. \end{aligned} \tag{8}$$

Now let us assume that this log was used to discover a BPMN model whose traces were then generated using the constraint-based approach. The synthetic log W_S was determined as follows:

$$W_S = \{\{A, B, C, D\}, \{B, A, C, D\}, \{A, C, B, D\}\}. \tag{9}$$

Traces where activity D does not occur were not present in W_S. Thus, only half of the original traces are reproduced by the model which results in a model fitness equal to 50%. On the other hand, trace $\{A, C, B, D\}$ is not an element of W_C. In this case the execution precision will be equal to 66,67%, as only two synthetic traces out of three can be found in the original workflow log.

The application of the proposed method to the example process model presented in Figure 4 resulted in generation of 11820 distinct workflow traces. Table I presents results of a log-based verification performed for the example BPMN model.

Table I
Evaluation of the example BPMN model.

Parameter	Value
Number of original traces	36
Number of synthetic traces	11820
Traces not generated	0
Model fitness	100%
Synthetic traces not included in the log	11784
Execution precision	0.03%

The results show that the discovered process model is characterized by low execution precision (0.03%). It means that the selected process mining algorithm has a tendency to generalize, i.e. it allows for much more behavior than included in the original workflow log. Therefore, the next step of the verification process should be to check if the synthetic traces not included in the original log can be allowed in the real process. If not, then the choice of the process mining algorithm should be reconsidered in order to provide more accurate process representations.

IX. Conclusions

In the paper, we presented a novel constraint-based algorithm which results in generation of a sufficiently complete workflow log for a given business process model. The proposed approach may serve as an additional tool to verify BPMN diagrams generated using process mining techniques. Comparison of the real execution log with a synthetic one helps to choose the most suitable discovery algorithm for the analyzed process or gives clues to the user how the model can be enhanced manually.

As future works, we plan to develop an automated tool for a comparison of two workflow logs which will be able to identify flaws occurring as a result of process discovery. Such a solution could be also used as a decision support system that, based on a re-created log, provides advice to process designers which mining algorithm to use.

References

[1] J. R. Nawrocki, T. Nedza, M. Ochodek, and L. Olek, "Describing business processes with use cases," in *BIS*, 2006, pp. 13–27.

[2] K. Kluza and K. Honkisz, "From SBVR to BPMN and DMN models. proposal of translation from rules to process and decision models," in *Artificial Intelligence and Soft Computing: 15th International Conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016, Proceedings, Part II*, ser. Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds. Springer International Publishing, 2016, vol. 9693, pp. 453–462.

[3] F. Friedrich, J. Mendling, and F. Puhlmann, "Process model generation from natural language text," in *Advanced Information Systems Engineering*. Springer, 2011, pp. 482–496.

[4] K. Kluza and P. Wiśniewski, "Spreadsheet-based business process modeling," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1355–1358.

[5] A. A. Kalenkova, W. M. van der Aalst, I. A. Lomazova, and V. A. Rubin, "Process mining using BPMN: relating event logs and process models," *Software & Systems Modeling*, vol. 16, no. 4, pp. 1019–1048, 2017.

[6] W. M. Van Der Aalst, "A general divide and conquer approach for process mining," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE, 2013, pp. 1–10.

[7] A. Rozinat, A. K. A. de Medeiros, C. W. Günther, A. Weijters, and W. M. van der Aalst, "The need for a process mining evaluation framework in research and practice," in *International Conference on Business Process Management*. Springer, 2007, pp. 84–89.

[8] S. Suriadi, R. Andrews, A. H. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Information Systems*, vol. 64, pp. 132–150, 2017.

[9] A. Rozinat and R. Mans, "Mining cpn models: discovering process models with data from event logs," in *In Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN*. Citeseer, 2006.

[10] A. Rozinat, A. A. De Medeiros, C. W. Günther, A. Weijters, and W. M. Van der Aalst, "Towards an evaluation framework for process mining algorithms," *BPM Center Report BPM-07-06, BPMcenter. org*, vol. 123, p. 142, 2007.

[11] W. M. van der Aalst, H. De Beer, and B. F. van Dongen, "Process mining and verification of properties: An approach based on temporal logic," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2005, pp. 130–147.

[12] D. Loreti, F. Chesani, A. Ciampolini, and P. Mello, "A distributed approach to compliance monitoring of business process event streams," *Future Generation Computer Systems*, 2018.

[13] F. Chesani, A. Ciampolini, D. Loreti, and P. Mello, "Abduction for generating synthetic traces," in *International Conference on Business Process Management*. Springer, 2017, pp. 151–159.

[14] M. Weske, *Business Process Management: Concepts, Languages, Architectures 2nd Edition*. Springer, 2012.

[15] T. H. Davenport, *Process Innovation: Reengineering Work Through Information Technology*. Boston, MA, USA: Harvard Business School Press, 1993.

[16] M. Hammer and J. Champy, *Reengineering the Corporation: A Manifesto for Business Revolution*. New York, NY, USA: Harper Business, 1993.

[17] S. A. White and D. Miers, *BPMN Modeling and Reference Guide: Understanding and Using BPMN*. Lighthouse Point, Florida, USA: Future Strategies Inc., 2008.

[18] A. Lindsay, A. Dawns, and K. Lunn, "Business processes - attempts to find a definition," *Information and Software Technology*, vol. 45, no. 15, pp. 1015–1019, December 2003, elsevier.

[19] H.-E. Eriksson and M. Penker, *Business Modeling with UML: Business Patterns at Work*. Wiley, 2000.

[20] I. Jacobson, M. Ericsson, and A. Jacobson, *The object advantage: business process reengineering with object technology*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1994.

[21] N. Melao and M. Pidd, "A conceptual framework for understanding business processes and business process modelling," *Information Systems Journal*, vol. 10, no. 2, pp. 105–129, 2000.

[22] M. Owen and J. Raj, "BPMN and Business Process Management. Introduction to the new business process modeling standard." OMG, Tech. Rep., 2006, www.bpmn.org.

[23] W. van der Aalst, "Business process management: a personal view," *Business Process Management Journal*, vol. 10, no. 2, 2004.

[24] WfMC, "Workfow Management Coalition," http://www.wfmc.org/.

[25] P. Lawrence, Ed., *Workflow Handbook*. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[26] M. zur Muehlen and D. T.-Y. Ho, "Risk management in the BPM lifecycle," in *Business Process Management Workshops*, 2005, pp. 454–466.

[27] W. Van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[28] P. Wiśniewski, "Decomposition of business process models into reusable sub-diagrams," in *ITM Web of Conferences*, vol. 15. EDP Sciences, 2017, p. 01002.

[29] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, J. Buijs *et al.*, "Process mining manifesto," in *International Conference on Business Process Management*. Springer, 2011, pp. 169–194.

[30] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[31] A. A. Kalenkova, M. de Leoni, and W. M. van der Aalst, "Discovering, analyzing and enhancing BPMN models using ProM?" in *Business Process Management-12th International Conference, BPM*, 2014, pp. 7–11.

[32] A. Rozinat and W. M. van der Aalst, "Decision mining in prom," in *Business Process Management*, ser. Lecture Notes in Computer Science. Springer, 2006, vol. 4102, pp. 420–425.

[33] A. Suchenia (Mroczek), P. Wiśniewski, and A. Ligęza, "Overview of verification tools for business process models," in *Communication Papers of the 2017 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 13. PTI, 2017, pp. 295–302. [Online]. Available: http://dx.doi.org/10.15439/2017F308

[34] J. M. E. Van der Werf, B. F. van Dongen, C. A. Hurkens, and A. Serebrenik, "Process discovery using integer linear programming," in *International conference on applications and theory of petri nets*. Springer, 2008, pp. 368–387.

[35] K. Kluza and G. J. Nalepa, "Proposal of square metrics for measuring business process model complexity," in *Proceedings of the Federated Conference on Computer Science and Information Systems – FedCSIS 2012, Wroclaw, Poland, 9-12 September 2012*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., 2012, pp. 919–922. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6354395

[36] J. Cardoso, J. Mendling, G. Neumann, and H. A. Reijers, "A discourse on complexity of process models," in *Proceedings of the 2006 international conference on Business Process Management Workshops, Vienna, Austria*, ser. BPM'06, S. D. e. a. J. Eder, Ed. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 117–128.

[37] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012.

[38] V. S. W. Lam, "A precise execution semantics for BPMN," *IAENG International Journal of Computer Science (IJCS)*, vol. 39, no. 1, pp. 20–33, 2012.

[39] R. M. Dijkman, M. Dumas, and C. Ouyang, "Semantics and analysis of business process models in BPMN," *Information and Software technology*, vol. 50, no. 12, pp. 1281–1294, 2008.

[40] J. Mendling, H. A. Reijers, and W. M. van der Aalst, "Seven process modeling guidelines (7pmg)," *Information and Software Technology*, vol. 52, no. 2, pp. 127–136, 2010.

[41] C. Favre and H. Völzer, "The difficulty of replacing an inclusive or-join," in *International Conference on Business Process Management*. Springer, 2012, pp. 156–171.

[42] R. Klimek, L. Faber, and M. Kisiel-Dorohinicki, "Verifying data integration agents with deduction-based models," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE, 2013, pp. 1029–1035.

[43] P. Sitek and J. Wikarek, "A hybrid method for modeling and solving constrained search problems," in *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE, 2013, pp. 385–392.

[44] A. Ligęza, "Models and tools for improving efficiency in constraint logic programming," *Decision Making in Manufacturing and Services*, vol. 5, no. 1, pp. 69–78, 2011. [Online]. Available: https://journals.agh.edu.pl/dmms/article/view/537

[45] P. Wiśniewski, K. Kluza, M. Ślażyński, and A. Ligęza, "Constraint-based composition of business process models," in *International Conference on Business Process Management*. Springer, 2017, pp. 133–141.

[46] P. van Beek and X. Chen, "Cplan: A constraint programming approach to planning," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, ser. AAAI '99/IAAI '99. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1999, pp. 585–590. [Online]. Available: http://dl.acm.org/citation.cfm?id=315149.315406

[47] J. C. Buijs, B. F. Van Dongen, and W. M. van Der Aalst, "On the role of fitness, precision, generalization and simplicity in process discovery," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2012, pp. 305–322.

3rd International Workshop on Language Technologies and Applications

DEVELOPMENT of new technologies and various intelligent systems creates new possibilities for information processing. Natural Language Processing (NLP) addresses problems of automated understanding, processing, evaluation and generation of natural human languages. LTA workshop provides a venue for discussion and presenting innovative research in NLP domain, but not restricted, to: computational and mathematical modeling, analysis and processing of any forms (spoken, handwritten or text) of human language, interactions via Virtual Reality and Augmented Reality, Computational Intelligence models and applications but also other various applications in decision support systems. We welcome papers covering innovative applications and practical usage of theoretical aspects. The LTA workshop will provide an opportunity for researchers and professionals to discuss present and future challenges as well as potential collaboration for future progress in the field.

TOPICS

The submitted papers shall cover research and developments in all NLP aspects, such as (however this list is not exhaustive):

- Computational Intelligence methods applied to language & text processing
- text analysis
- language networks
- text classification
- language networks, resources and corpora
- document clustering
- various forms of text recognition
- machine translation
- intelligent text-to-speech (TTS) and speech-to-text (STT) methods
- authorship identification and verification
- author profiling
- plagiarism detection
- sentiment analysis
- NLP applications in education
- knowledge extraction and retrieval from text and natural language structures
- multi-modal and natural language interfaces
- innovative language-oriented applications and tools
- interactions models and applications via Virtual Reality and Augmented Reality
- NLP for text analysis in forensic linguistics and cybersecurity

EVENT CHAIRS

- **Damasevicius, Robertas,** Kaunas University of Technology, Lithuania
- **Martinčić – Ipšić, Sanda,** University of Rijeka, Croatia
- **Napoli, Christian,** Department of Mathematics and Informatics, University of Catania, Italy
- **Woźniak, Marcin,** Institute of Mathematics, Silesian University of Technology, Poland

PROGRAM COMMITTEE

- **Artiemjew, Piotr,** University of Warmia and Mazury, Poland
- **Burdescu, Dumitru Dan,** University of Craiova, Romania
- **Calixto, Iacer,** University of Amsterdam, The Netherlands
- **Čukić, Bojan,** UNC Charlotte, United States
- **Cuzzocrea, Alfredo,** University of Trieste, Italy
- **Dobrišek, Simon,** University of Ljubljana, Slovenia
- **Ganchev, Ivan,** University of Limerick, Ireland / University of Plovdiv "Paisii Hilendarski", Bulgaria, Ireland
- **Gelbukh, Alexander,** Instituto Politécnico Nacional, Mexico
- **Grigonytė, Gintarė,** University of Stockholm, Sweden
- **Harbusch, Karin,** Universität Koblenz-Landau, Germany
- **Kapočiūtė-Dzikienė, Jurgita,** Vytautas Magnus University, Lithuania
- **Krilavičius, Tomas,** Vytautas Magnus University, Lithuania
- **Kurasova, Olga,** Vilnius University, Institute of Mathematics and Informatics, Lithuania
- **Lopata, Audrius,** Vilnius University, Lithuania
- **Marszałek, Zbigniew,** Silesian University of Technology, Poland
- **Maskeliūnas, Rytis,** Kaunas University of Technology, Lithuania
- **Matson, Eric T.,** Purdue University, United States
- **Meštrović, Ana,** University of Rijeka, Croatia
- **Mikelić-Preradović, Nives,** University of Zagreb, Croatia
- **Nowicki, Robert,** Czestochowa University of Technology, Poland
- **Połap, Dawid,** Institute of Mathematics, Silesian University of Technology, Poland
- **Pulvirenti, Alfredo,** University of Catania, Italy
- **Rosen, Alexandr,** Charles University, Czech Republic
- **Sanada, Haruko,** Rissho University, Japan

- **Skadina, Inguna,** University of Liepaja, Latvia
- **Śluzek, Andrzej,** Khalifa University, United Arab Emirates
- **Šnajder, Jan,** University of Zagreb, Croatia
- **Stanković, Ranka,** University of Belgrade, Serbia
- **Starczewski, Janusz,** Czestochowa University of Technology, Poland
- **Steinberger, Josef,** University of West Bohemia, Czech Republic
- **Szymański, Julian,** Gdansk University of Technology, Poland
- **Tahmasebi, Nina,** University of Gothenburg, Sweden
- **Tambouratzis, George,** Institute for Language and Speech Processing, Athena Research Centre, Greece
- **Tramontana, Emiliano,** University of Catania, Italy
- **Wang, Lipo,** Nanyang Technological University, Singapore
- **Wei, Wei,** School of Computer Science and Engineering, Xi'an University of Technology, China
- **Yanushkevich, Svetlana,** University of Calgary, Canada
- **Žabokrtský, Zdeněk,** Charles University

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 457–464
DOI: 10.15439/2018F304
ISSN 2300-5963 ACSIS, Vol. 15

Do Actions Speak Louder Than Words? Predicting Influence in Twitter using Language and Action Features

Fatima Al-Raisi, Shadab Alam, Bruno Vavala, Mao Sheng Liu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
Email: {fraisi,shadaba,bvavala,maoshenl}@andrew.cmu.edu

***Abstract*—This work explores the connection between language, personality, and influence in a social media network. It clusters users based on two types of features: account activity features and stream content (word) features and compares the usefulness of these different types of features in categorizing users according to their influence and leadership potential in the network. Results of clustering using different sets of features are examined to answer questions about distribution of Twitter users from the influence perspective. These results are compared against distributions of personality traits obtained from previous research on personality types and established assessment tools that measure leadership aptitude and style. Experiments with different clustering algorithms are described and their performance and cluster outputs are reported.**

I. Introduction

This work pursues the research question of how language, personality, and influence are connected. We use Twitter data to analyze users from the perspective of influence. For the purpose of this analysis, we use both user account data and content data to cluster users in different categories according to their leadership or influence abilities. We refer to features extracted from each type of data above as "activity" (or account) features and "word" (or language) features, respectively. We also compare the different types of features in their usefulness for modeling and predicting influence. The underlying reference model is DISC [1], [2]: a behavioral assessment tool that measures different personality traits including leadership aptitude and style. It was first proposed by psychologist William Marston in 1928 and later developed by industrial psychologist Walter Clarke. DISC is a widely used personality assessment tool that is based on studies in various fields and involves surveying a large number of people from different backgrounds, professions, and personalities [3], [4]. A significant portion of this assessment tool is based on the use of language; hence, motivating its use for extracting features to identify influence in tweets. We use this specific tool as a reference for identifying different personality traits associated with DISC categories and focus on the categories of leaders and influential people.

This kind of profiling has many important applications. It can be used to identify points of influence in a large social network. This is of especial interest in places where social media is used as an alternative means to exercise influence or express opinions otherwise not represented in main stream media. For commercial purposes, companies may need to identify influential users to provide them a product or a service so they may recommend it to their followers. This is essentially linked to the "activity shaping" problem in social networks. This kind of analysis can also be used to answer important questions about social media networks such as the similarity in behavioral distribution to patterns/distributions found in larger populations. In addition to answering questions about social media, this kind of analysis can help understand and better model the dynamics of influence, trust, and information propagation. Other applications include targeted advertisement and personalized interface design.

This project is on one hand exploratory work aimed at examining the nature of Twitter data in terms of whether typical DISC distribution patterns can be found in Twitter and whether content/word features are as useful in predicting influence as account/activity features. On the other hand, this work can be viewed as a first step towards automating the task of DISC profiling in social media networks.

II. Background and Related Work

In recent years, a line of research connecting natural language processing and social media analysis has emerged. Several related studies focused on various aspects of personality and interaction including prediction of social relationships and tie strength [5], [6], prediction of Big Five personality traits [7], and prediction of anti-social traits [8]. *Influence* which is an important aspect of personality has been studied using language features or account activity features but has not been explored, to our knowledge, in social media analysis using and contrasting both types of features.

While previous work focused on personality models based mostly on the Big Five personality traits [7], [5], we use DISC model in this work to explore the relationship between language and influence. DISC is an assessment tool that has been developed for different personality analysis purposes including testing leadership aptitude and

style [1], [2]. It basically distributes the population in a space of two dimensions that roughly correspond to 1) people-oriented vs. task-oriented and 2) outgoing/active/fast-paced vs. reserved/reflective/moderately-paced as shown in Figure 1 on page 3. Different variants of the test focus on different aspects of the classification depending on which dimension is more relevant to the problem and its domain. According to DISC studies focusing on leadership aptitude, only 4% of the population falls in the two extremes of leadership aptitude and at most 2% are natural leaders who have strong leadership qualities regardless of training and environment. The majority of the population is found in between not deviating much from the mean in a distribution that resembles a bell curve. The DISC model has been used in different studies for different purposes such as improving team performance through behavioral assessment profiling [9], identifying behavioral factors of individuals in high managerial ranks [3], and even studying the influence of personality style on performance of students in educational settings [4]. In this work we focus on using DISC to answer questions about categorization of Twitter users according to leadership aptitude and style and compare empirical findings to expected distribution based on domain knowledge. We also explore whether language features are as useful as action features (e.g.,#followers, #following,#retweets, etc.) in modeling and predicting influence. We first discover clusters based on user account features and word features separately and then examine whether we obtain similar or different results. We compare the usefulness of each type of features for coming up with clusters that resemble DISC categories and decide whether "actions speak louder than words," "words speak louder than actions," or whether they convey the same information in this context. This also allows us to see differences between the influence aspect of personality and other aspects that are accurately predictable using linguistic content as shown in previous work [7], [8], [6], [5].

III. Data and Feature Engineering

Next we describe the dataset, the different features computed, the motivation and method for feature selection and numerical scaling of features.

A. Data

In this work, we use a subset of Twitter obtained and published in previous work on social user profiling for inferring home locations [10]. The dataset contains network data for 3 million users (profile/account data) and 147 thousand tweet streams. There are about 78 thousand users for which we have both account data and tweet streams.

B. Feature Engineering

We consider two types of features: account (action) features and language content (word) features. Account features include the number of followers, number of friends (following), ratio of the previous two numbers, number of tweets, retweets, and favorite (liked) tweets. The last two are used as an indication of tweet popularity/impact as tweets that tend to be retweeted and liked frequently have more influence and propagate further. In addition to these features that are almost available with the data and required little computation, we also compute the page rank of a user as another account feature. The page rank of a user A is given by the formula:

$$PageRank(A) = 1 - d + d \sum_{i=1}^{n} \frac{PageRank(i)}{L(i)}$$

where n is the number of A followers, $L(i)$ is the number of i's followers and d is the damping factor[1]. Table 1 lists all account (action) features. Note that these features are not all independent. We experiment with different functions and subsets of features for clustering users.

The other type of features is language/content features. These features are based on a bag of words language model in which words are either grouped or treated as individual features. Several linguistic content categorization systems exist including Linguistic Inquiry Word Count (LIWC) system that is commonly used for personality analysis [8], [12], [5], DISC categories which are based on grouping words according to DISC categories: Dominance, Influence/Inducement, Submission, and Compliance, and finally broad categorization of most frequent words into linguistic categories such as function words, common verbs, and pronouns and semantic categories such as social processes, emotions, and work-related words. Both normalized frequencies and tf-idf were used in different experiments to explore the effect of relative weighting in this clustering task.

1) Extracting Content Features: Since this work is exploratory in nature, different sets of word features were used. In one set, words are grouped into categories and one frequency counter is maintained for each category, another set was formed by splitting words into separate features (frequency counter for each word), a third set was formed by including words describing different categories in the DISC assessment, and another variant of the word features was based not on counts but tf-idf scores. The set in which words were split as individual features resulted in very sparse representation of some features so we used the grouped version of the word features (as done in linguistic analysis of most related work). The tweet stream was preprocessed before computing the features. The text was converted to lower case, irrelevant punctuation and other markers were removed, and constant keywords were ignored. However, we did not perform stemming on the tweet stream; weighing the computational cost and information gain we decided that counting variants of the surface word sharing the same stem was not computationally expensive and often important to differentiate.

2) Extracting and Scaling Account/Activity Features: Most action features were readily available in the data. User profiles include the number of followers, friends, retweets, and favorite (liked) tweets. However, in addition to normalizing these

[1]We use the commonly assumed value of 0.85 [11].

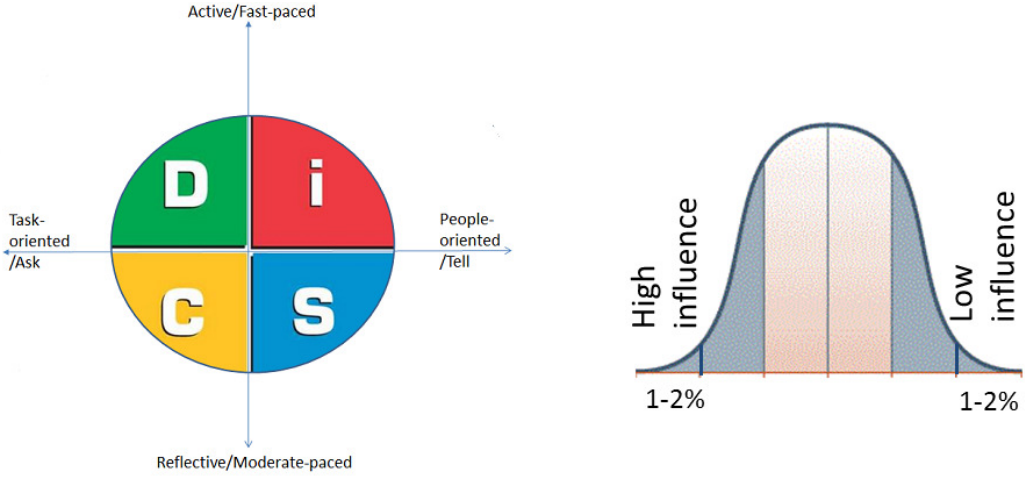


Fig. 1. DISC categories and distribution

activity features	word features
# followers A	LIWC
# following B	DISC
ratio A/B	LIWC + DISC
# tweets	tf-idf of LIWC
# retweets	tf-idf of DISC
# hashtags	tf-idf of LIWC + DISC
# likes	individual words variant of above
pageRank	grouped words variant of above

TABLE I
ACCOUNT ACTIVITY FEATURES AND CONTENT/WORD FEATURES

counts and creating features based on simple functions of these counts (ratios for example), we computed a "pageRank" feature which captures deeper influence in the network than simply the number of followers or the ratio between followers and friends. Interestingly, results show that the pageRank follows a power-law distribution; i.e., there are very few users with high pageRank and there is a long tail of users with small pageRank, while the number of followers have a smoother distribution.

Analyzing user account data, we noticed a wide range of values with a long tail for several key features. This called for scaling of features before feeding the values into clustering algorithms, otherwise the algorithms may yield unexpected results or convergence behavior due to the skewed distribution. The following account statistics are obtained from three million user accounts (3123283 Twitter user profiles). Table 2 shows that the data varies in a wide range with a very long tail across all features. This long tail phenomenon has to be addressed before clustering. Therefore, all features are scaled by the median of the feature set except the *likes* count feature (because its median is 0) which is scaled by 10 times the mean of that count.

3) Feature Selection: Since the number of combined features is very large, Principle Component Analysis (PCA) was done to produce a lower number of linearly uncorrelated features that are most informative. The total number of features exceeds 500, since not all features may be informative and a large number of features may slow the convergence of clustering, we used PCA to represent these hundreds of features in a few eigen components not exceeding 15. Indeed, dimensionality is reduced with PCA and clustering was performed on the projected feature space produced by PCA.

IV. METHOD

Section 2 detailed the different feature types and different linguistic categorizations for word features (LIWC, DISC, individual, grouped). This section presents the clustering algorithms and different experimental settings created from various combinations of feature types and clustering algorithms.

A. Empirical Support for Hypothesis

One of the main questions in this work is whether language and influence are related. We describe an experiment conducted to test the hypothesis that language and influence are related before running clustering algorithms. In this experiment, users were ranked according to each action feature (#followers, #following, etc.), resulting in n different rankings/lists (where n = #action features), then the top 5% of users in each list were extracted, the pair-wise intersection of user lists (i.e., intersection of top 5% users according to each pair of features) was obtained, and finally the union of resulting sets was taken.

Feature	min	max	mean	median	sum
# FRIENDS	0	695509.0	719.4	218	2246906359.0
# FOLLOWERS	0	11060753.0	1348.48	136.0	4211702993
# TWEETS	0	982934.0	2319.71	272.0	7245121544.0
# LIKES	0	3200.0	27.41	0.0	85628538.0

TABLE II
ACCOUNT DATA STATISTICS

The final set contained the top 5% users according to action features. Initially, the experiment was designed to simply take the intersection of all n lists and regard that intersection as the top 5% influential users according to action features but that intersection was almost empty[2] which suggests that these features were not redundant and that each feature targets a different "action" and therefore possibly different kinds of users. The other option was to simply union all top 5% users obtained from rankings by different features but that set may contain users that are not as influential as those who are ranked highly by more than one feature. Taking the pair-wise intersection and then taking the union of all resulting sets is a balanced option in between. The 5% lowest rank users were sampled following a similar procedure.
We then examined the language use for these two sets of users that are on different extremes according to action features. A clearly different use of language across all linguistic categories is observed. The variation is measured in differences in (normalized) frequencies of words across categories as shown in Figure 2. One observation is that the top 5% users tend to use more words (i.e., express themselves more) and that the ratio between the two sets of users varies across categories from double to more by a third or less. This supports the intuition that language and influence are related. The following section describes the clustering experiments conducted to further examine this hypothesis and answer other questions about influence distribution in Twitter.

B. Clustering

Clustering was done using three different clustering algorithms: k-means, EM, and spectral clustering. In each experiment users were clustered according to word features and account features, separately. The resulting sets of clusters are examined for similarity and overlap. The idea is that if we cluster using action features and cluster using word features separately and then find that the resulting clusters are similar and overlap then we can infer that these different sets of features (actions and words) model the same phenomenon: influence, and that although they are different in nature they are strongly related as they can make similar predictions about the same phenomenon. Algorithm 1 on page 4 is high-level description of the clustering and analysis steps.

k-means: We experimented with different values of k. Based on the problem domain, however, we selected 4 as it reflects the number of main categories in DISC. Although k-means is suitable in settings where the data is expected to be separable, we noticed that clusters we obtained are dense around the mean with far fewer points spread further. So even with k-means we were able to obtain clusters that were clearly distinct.

[2]The intersection of all top 5% lists included only 6 users from the original list of 78 thousand users.

Algorithm 1 Cluster and Analyze

```
for each feature set S do
  for each clustering algorithm A do
    cluster users according to S using A
  end for
end for
for each clustering algorithm A do
  examine overlap between action-based clusters and word-based clusters
  examine similarity of clusters obtained using different linguistic content categorization
  examine relative sizes of clusters in each clustering
end for
return overlapping clusters, cluster size distribution, and corresponding algorithm A
```

EM: Since EM is suitable for soft clustering where clusters are expected to overlap, we clustered the users using EM on mixture-of-Gaussian models. This was motivated by the large number of data points ($>$ 78K users) calling for the applicability of the central limit theorem as a reasonable assumption and more significantly that the overall distribution of influence resembles a normal distribution as depicted in II on page 3.

Spectral: We also applied spectral clustering to see if it confirms results of other clustering algorithms or behaves differently. Spectral clustering is further motivated by its applicability in settings where clusters may overlap. We noticed that spectral clustering does not scale with a large dataset. To successfully apply spectral clustering, k-means was first run to reduce the dimensionality of the data and then spectral clustering was run on the dimension-reduced dataset. Different values of k ranging from 700 to 50 were tried, spectral clustering scaled only to the smallest dataset; i.e., the maximally reduced set with 50 means.
In the following section, we present results and compare the performance of clustering algorithms.

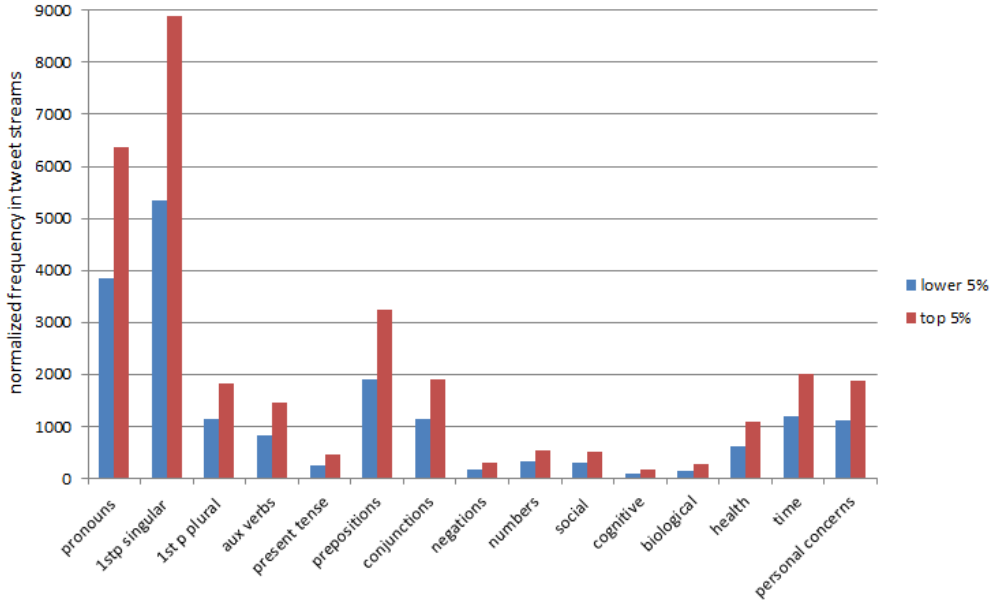


Fig. 2. language use in different groups of users

V. Results

A. Comparison of Different Clustering Algorithms

Table 3 summarizes differences in performance and output of clustering algorithms. The runtime reported is based on running the algorithms on the largest data set of profiles (> 2.5 million users). The runtime for spectral clustering includes the runtime of running k-means with k=50 first and then running spectral clustering on the result.

B. Cluster Overlap Results

Clusters based on action features and those based on word features are examined for overlap. Overlap between two clusters is measured in the number of users they have in common. The idea is that if action-based clustering and word-based clustering result in respective clusters that overlap; i.e., have many users in common, we conclude that action features and word features can mirror each other in making similar predictions and therefore are related. Figure V-B shows overlap patterns for clustering using different sets of features. A diagonal means that overlap is observed in all four clusters. A partial diagonal means that overlap is observed only in some subset of clusters. We present the plots that represent common patterns across algorithms and discuss these findings in the analysis section.

VI. Analysis and Discussion of Results

The following are observations from the experimental results along with analysis for each result.

1) Clustering based on account features and on word features show that the sizes of clusters produced closely match the typical size of population categories according to DISC with two small clusters representing the extremes on influence abilities and two large (possibly overlapping) clusters representing most of the population. In fact, all clustering combinations almost always produce a skewed distribution of cluster sizes which matches the typical distribution of individuals according to the theory of influence and leadership. It is also observed that the distribution is more skewed when word features are used to cluster. We've attempted further analysis on the word content of these clusters in order to draw reliable conclusions about the nature of these clusters. This is discussed in cluster identification section below.
2) Similarity was found between clusters obtained using word features from different linguistic content categorization systems. More specifically, adding DISC word categories to LIWC does not significantly change clusters obtained from LIWC word categories. This shows that the latter is a comprehensive categorization that subsumes DISC categories. However, clustering based on DISC categorization of linguistic content results in better alignment with clusters obtained from action features. This agrees with the claim that DISC is designed to specifically target the influence aspect of personality.
3) Similar clusters are obtained from account features in original spaces and PCA projection space. This result is not surprising due to the small number of action features that even when projected using PCA result in principle components that are very similar to the original representation.
4) More importantly, a clear overlap is observed between clusters based on action features and clusters based on word features.

	run-time	scalability with		cluster sizes
		# features	# data points	
4-means	1m 1s	✓	✓	12,19,27,42(%)
EM	5m 3s	✓	✓	6,13,36,45(%)
Spectral	42m 44s	✗	✗	6,18,35,41(%)

TABLE III
COMPARISON OF DIFFERENT CLUSTERING ALGORITHMS

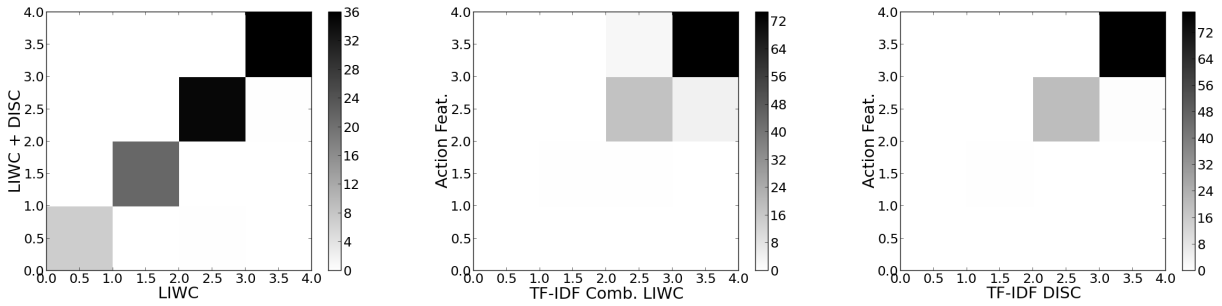


Fig. 3. Overlap between clusters obtained using different feature sets

5) Different clustering algorithms produce different clusters for both action and word features yet EM and spectral clustering results are closer to each other. This shows that each algorithm handles the data differently leading to different clusters. However, the overall relative sizes of clusters and their densities are very similar. We suspect that each of these algorithms is best at modeling a specific aspect of the intricate phenomenon of influence for example: degree of influence vs. style of influence.

VII. CLUSTER IDENTIFICATION

The objective of cluster identification is to label instances in clusters according to influence categories. In the initial experiments, our analysis was not limited to the sizes of clusters. We examined the word content for clusters based on user account (activity) features. The analysis was to see if a mapping exists between clusters obtained from account activity features and those obtained from word features. Although the use of language shows commonalities among clusters, there are variations in the word patterns from cluster to cluster suggesting some differences in the use of language across clusters obtained from account/action features only. There are at least two clusters that seem more similar and two that seem to exhibit different use of words. This agrees with the distribution of cluster sizes with two small clusters (different use of language) and two overlapping clusters (similar use of language).

We also checked membership of the top 5% users sampled according to action features to see if they belong to the same word-based cluster which we can use to identify the cluster of most influential users. These users were distributed across all clusters.

We sampled tweet streams of users from each cluster (closest to the centroids in k-means) and examined the content of

Fig. 4. visualization of language content of sample streams: small cluster vs. large cluster

the small vs. large clusters that we expect are the influential and ordinary clusters, respectively. Indeed, the sample from the small cluster streams shows that the discussions and topics were about work, strategy, marketing, and topics that are in general very different from those found in the other sample which are mostly related to personal concerns and daily activities as depicted in Figure 4 on page 6. This shows that the small clusters, as expected from background knowledge on typical influence distribution, correspond to influence extremes. The limitations of these samples should be taken into consideration, however, as these samples are small and streams are noisy.

A. Challenges and Remarks

Data Representation and Sparsity: Coming up with useful informative features require revisiting domain specific algorithms and implementing them. For example, pageRank features. Four different versions of word features and different grouping combinations and linguistic models were used. We tried running the clustering algorithms using all combinations

but some resulted in sparse representation of the data so PCA and clustering algorithms could not be run. Although this was useful for feature selection (eliminate features leading to sparsity), it was counter-intuitive as some of these features were carefully selected based on domain knowledge (i.e., relevance to DISC).

Dimensionality: With the large number of user account and stream content features, both feature extraction and data clustering require significant computation power and memory. Most of the experiments were run on a dedicated server yet some algorithms had excessively long run times or crashed due to memory errors. We attempted different clustering algorithms including DBSCAN (for density-based clustering) and spectral clustering which was useful in understanding scalability limitations of different clustering algorithms and possible approaches to address these limitations like preprocessing using k-means prior to spectral clustering. Moreover, we attempted to cluster in different spaces: original feature space and feature space from PCA. Running in original feature space required considerably more computational time and power so clustering was mostly done on the transformed feature space.

Unsupervised learning of latent parameters: With no labels and no explicit correspondence between "influence" and language in the noisy data present in social media, the problem of identifying users with leadership potential is a challenging one. Using a latent variable model such as LDA or a combination of semi-supervised and active learning methods may be useful in moving from clustering to classification and realizing the goal of automated leadership profiling in online communities.

VIII. Conclusion

The overlap between clusters obtained from word features and those obtained from action features suggests the usefulness of both types of features in predicting influence. This answers the main research question posed in this work and shows, based on empirical evidence, that language and influence are connected. It is important to contrast results obtained from these two different sources of information especially with the more recent phenomena of bought followers and likes that can render account (non-content) features less reliable. Our experiments show that using different systems for analyzing the language content of Twitter can lead to variations in results and that DISC seems to be the most appropriate tool for studying influence. The influence distribution found in this sample of Twitter resembles that obtained from large population statistics. It is highly skewed and shows two minority categories that are clearly separated and two larger overlapping categories. Manual labeling of sampled instances from different clusters shows different use of language and agrees with theoretical and empirical results suggesting that influential users are a minority.

Appendix: Supplemental Material

k-means is typically used in cases where clusters are expected to be separable. However, the following figures are provided to illustrate that even when applying k-means to cluster

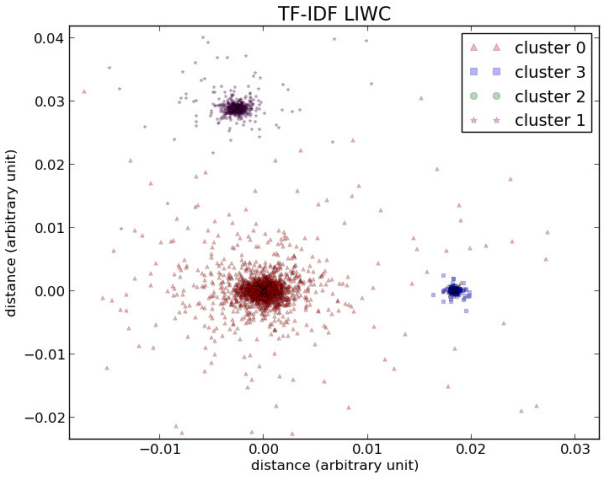


Fig. 5. Visualization of of k-means clusters based on tf-idf LIWC word features

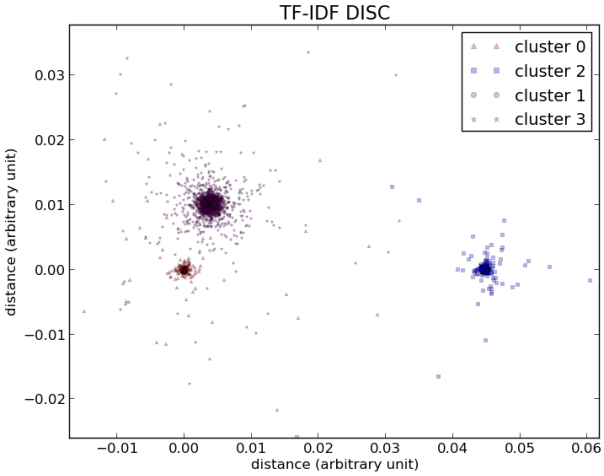


Fig. 6. Visualization of k-means clusters based on tf-idf DISC word features

users according to language features that are expected to result in some overlaps, we get clusters that are dense around the mean and clearly distinct. Since the data is multidimensional, a special distance function was designed to compute the distance between clusters and present their relative sizes and distances from each other in 2D. The two selected visualizations are for clusters obtained from tf-idf features on LIWC and DISC linguistic categorizations. The graphs show only three clusters because in one case the fourth cluster is too small to be seen relative to the other three clusters and in the other the fourth cluster is at a large distance from the other three clusters making it not possible to fit all four clusters in one reasonably scaled image.

References

[1] "DISC Instrument Validation Manual," http://www.coachannette.com/pdfs/DISCValidityManual.pdf, 2004.

[2] "DISC Classic®Validation Report," https://www.inscape-exchange.com/downloads/marketing_support/researchreports/DiSCClassicValidationResearchReport.pdf.
[3] G. Beamish, "How chief executives learn and what behaviour factors distinguish them from other people," in *Industrial and Commercial Training*. Emerald Group Publishing Limited, 2005, vol. 37, pp. 138–144.
[4] P. Blignaut and A. Naude, "The influence of temperament style on a student's choice of and performance in a computer programming course," *Comput. Hum. Behav.*, vol. 24, no. 3, pp. 1010–1020, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2007.03.005
[5] S. Adali, F. Sisenda, and M. M. Ismail, "Actions speak as loud as words: predicting relationships from social behavior data," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, New York, NY, USA, 2012, pp. 689–698. [Online]. Available: http://dx.doi.org/10.1145/2187836.2187930
[6] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 211–220. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518736
[7] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 149–156. [Online]. Available: http://dx.doi.org/10.1109/passat/socialcom.2011.33
[8] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets." in *ICMLA (2)*. IEEE, 2012, pp. 386–393. [Online]. Available: http://dblp.uni-trier.de/db/conf/icmla/icmla2012-2.html#SumnerBBP12
[9] J. Duck, "Making the connection: Improving virtual team performance through behavioral assessment profiling and behavioral cues," in *Developments in Business Simulation and Experiential Learning*, vol. 33, 2006, pp. 358–359.
[10] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations." in *KDD*, Q. Yang, D. Agarwal, and J. Pei, Eds. ACM, 2012, pp. 1023–1031. [Online]. Available: http://dblp.uni-trier.de/db/conf/kdd/kdd2012.html#LiWDWC12
[11] A. Altman and M. Tennenholtz, "Ranking systems: The pagerank axioms," in *Proceedings of the 6th ACM Conference on Electronic Commerce*, ser. EC '05. New York, NY, USA: ACM, 2005, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/1064009.1064010
[12] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of Language and Social Psychology*, 2010. [Online]. Available: http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 465–469
DOI: 10.15439/2018F155
ISSN 2300-5963 ACSIS, Vol. 15

Towards semantic search for mathematical notation

Agnieszka Bier
Institute of Mathematics
Silesian University of Technology
23 Kaszubska Str.
44-100 Gliwice, Poland
Email: agnieszka.bier@polsl.pl

Zdzisław Sroczyński
Institute of Mathematics
Silesian University of Technology
23 Kaszubska Str.
44-100 Gliwice, Poland
Email: zdzislaw.sroczynski@polsl.pl

***Abstract*—The paper concerns the design and implementation of a search engine for mathematical expressions given by the user in a convenient form of natural language or visual queries. Proper presentation and transcription of the mathematical notation is substantial for further processing and the adequate choice of the word distance measure for string comparison is an important issue as well. Within this project a complete solution for acquiring and processing the mathematical query and a searching algorithm is elaborated. We present results of exemplary search queries obtained for different types of input data format with application of two different word distance measures and discuss briefly the observed properties.**

I. Introduction

THE mathematical notation is an example of a visual language, which uses images for communication. A visual language consists of a set of planar diagrams representing expressions which are defined to be correct, and in which the syntax and semantics are determined by the mutual geometric relations between particular symbols.

The mathematical notation intended for machine processing may be transcribed in many different formats. The most popular ones are LaTeX, a dedicated document preparation system widely used for publication of scientific content, and MathML dedicated to Web applications. Both of these formats have their origins in text description languages, TeX and HTML/XML, respectively, and are designed to be open. Other solutions are used mainly in Desktop Publishing (DTP) and usually allow exporting to and importing from the above mentioned universal formats. We also mention a special format for mathematical Braille dialect, used by visually impaired people.

Independently of the chosen document description language, automation of mathematical notation processing requires transcription of the common planar visual language notation into a sequential notation used in programming languages. Hence all structural information on the mathematical formula, encoded in the geometric mutual relations between symbols, must be transcribed to the notation, in which the semantics of the given expression is incorporated solely in the sequencing of symbols possibly grouped by brackets (the significant dissimilarity of the Reverse Polish Notation to the natural mathematical notation excludes its application in most of the areas, especially for educational purposes). Such approach enables preparation of documents containing arbitrary complex mathematical expressions, while on the other hand it provides poor capabilities for searching or indexing of mathematical content.

The goal of the presented work was to develop a novel tool for searching for mathematical expressions based on queries containing parts of original or similar formulae, which are preferably given in a natural language.

II. Processing mathematical notation for semantic search

The problem of expressing the mathematical symbolic notation in the natural language has drawn interest of many researchers in different aspects within the scope of didactics of mathematics. A particular interest in this issue is observed in terms of assistance for the disabled people. A thorough survey on the English transcription of mathematical content can be found in [1], while some considerations on the ambiguity of certain notation rules are elaborated in [2], [3], [4], [5]. Various examples of algorithms implementing verbalization of mathematical content can be found in [6], [7], [8], [9], [10] and an algorithm designed to meet the needs of visually impaired is presented in [11]. Other results concern the problem of searching the mathematical expressions (see [12], [13], [14], [15], [16]) and some indexing algorithms for efficient search are elaborated in [17], [18], [19], [20], [21]. The automatization of content categorization is discussed in [22], methods for correctness validation are presented in [23], and certain search algorithms by means of natural language queries are elaborated in [24], [25], [26] and [27].

The verbalization of mathematics provided by math-to-speech engine used in our experiments allows the use of different dialects and national versions. There are English, Polish simplified/natural and full transcriptions available at the moment. The main difference between these versions is the level of details and abbreviations introduced to shorten the output and make it sound more or less similar to the actual verbalization done by human readers.

To build the query in an artificial description language with the use of complex notation and the set of non-familiar commands the user must have extensive technical experience. This is very unlikely for the majority of students at different levels of education. Moreover, translating the query into the special language distracts the user's attention from the actual search object.

There is also the additional challenge – to construct the query with the use of some characteristics of the searched mathematical expression, for example the presence of the particular key symbol as root or integral sign. This is important because the user very seldom remembers exactly the whole equation, rather reminds only some fragments, the general layout or is able to prompt the field of mathematics.

Summarizing, the important issue while searching for the mathematical expression in the database containing the collection of equations is the construction of the query. Is can be done with the use of common document description languages, but it requires complex lexical knowledge and complicates queries based on the fragmentary information and simplifications due to imperfections of human memory.

The verbalization of the mathematical notation stored in the database can help to solve problems mentioned above. There is a possibility to compare the spoken version of the equation with the query given in natural language with the use of the methods, tools and measures for approximate string matching. This way the accuracy of the semantic search for mathematical notation should significantly increase.

Given a query expression, the search algorithm follows the steps:

1) Converting the query expression entered in the Formula Editor into one of three formats: LaTeX, natural English, natural Polish
2) Comparison of the transcribed query with the existing database records by means of one of the text similarity measures: normalized Levenstein distance, and (weighted) cosine similarity measure.
3) Visual presentation of the list of 20 best matches (according to the chosen comparison measure results)

The Levenshtein distance $D_{Le}(a,b)$ is the measure of difference between text strings a and b of possibly different length, calculated as the minimum number of single-char edit operations, required to change one string (a) into another (b) [28], where operation categories are: insertions, deletions and replacements weighed equally 1. The normalized Levenshtein distance $D_{LeN}(a,b)$ additionally places the resulting values in the range 0 – 100%

$$D_{LeN}(a,b) = 100 \times \frac{D_{Le}}{\max\left(|a|,|b|\right)}$$

where $|s|$ is number of characters in string s.

Another useful text comparison measure is the (weighted) cosine similarity measure. Two strings are first transformed into numerical vectors $v = [v_1, ..., v_n]$ and $w = [w_1, ..., w_n]$ with coordinates being the counts of instances of a given key word in the string, and then the cosine of the angle between v and w is calculated by means of the standard dot product of the two vectors:

$$\cos\angle(v,w) = \frac{v \circ w}{|v| \cdot |w|} = \frac{\sum_{i=1}^{n} v_i \cdot w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \cdot \sqrt{\sum_{i=1}^{n} w_i^2}}.$$

Values close to 1 are interpreted for high similarity of the vectors, while values close to 0 indicate big differences in v and w. Additionally, by introducing weights in the cosine similarity measure, one can choose symbols that should have stronger or weaker impact on the comparison. For instance, since brackets are typically used only for ordering the expression and do not introduce their own meaning, one should pay less attention to them while comparing formulae. In our examples the following weights have been assigned to certain basic types of mathematical expressions: roots 30, fractions: 30; integral: 30; sums: 30; components: 15; products: 30; factors: 15; limits: 30; parentheses and brackets: 5; powers, including squares and cubes: 20. These weights have been chosen experimentally as ones appearing the most suitable for effective comparison.

In the case of comparison based on the verbalization (word transcription) of the formulae, an initial step of conversion to the spoken language is performed and the transcripts are stored in a database. In the conversion process the macrodefinitions formulated in an interpretable script language Lua are used. A model is loaded from the database to the inner engine, which stores the tree of objects corresponding to each element of the processed formula. To increase the efficiency of the comparison all database records are pre-processed and saved in an additional data structure of type in-memory-table. As the processing time of a formula consisting of tens of symbols is ca 50–100ms for typical PC unit, the initial pre-processing step significantly increases the performance of the presented algorithm.

III. Experimental results

The proposed algorithm has been implemented and tested for the quality of search, interpreted as the similarity of the original formula to the obtained results. The performance measure of the search algorithm was defined as the position of the original formula on the list of search results. The experimental queries were formulated in all accepted data formats:

1) the inner data format of the Equation wizard ("Edytor wzorów") application, based on LaTeX language,
2) the natural English transcription/verbalization,
3) the natural Polish transcription/verbalization.

With respect to the details of the content and possible application areas the following three categories of test queries were distinguished:

(C1) the query and original expressions differ in single symbols (correction),
(C2) the query was a part of the original expression (autocompletion),
(C3) the query expression contains few symbols, specific to the original expression (search).

Each test query was introduced to the search engine in two ways: by means of a visual editor and by entering the respective natural language key words. The performance of our search algorithm was then compared for the different

TABLE I
TOTAL SUCCESS RATE OF SEARCH WITH RESPECT TO FORMATS OF COMPARED DATA, QUERY INPUT (VISUAL / KEY WORDS), AND THE MEASURE USED FOR STRING COMPARISON.

Data format	Measure	Visual query	Key words query
Inner	Levenstein	66.6%	0%
	Cosine	66.6%	17%
English	Levenstein	83.3%	100%
	Cosine	100%	50%
Polish	Levenstein	100%	83.3%
	Cosine	100%	33.3%

data formats, within each category of queries and the input modes. The test consisted of two parts. Firstly, the queries were executed on a general database of mathematical expressions, containing various formulae from different branches of mathematics. Some of the database records were totally irrelevant to the queries. We present the obtained results in Table II, column GD.

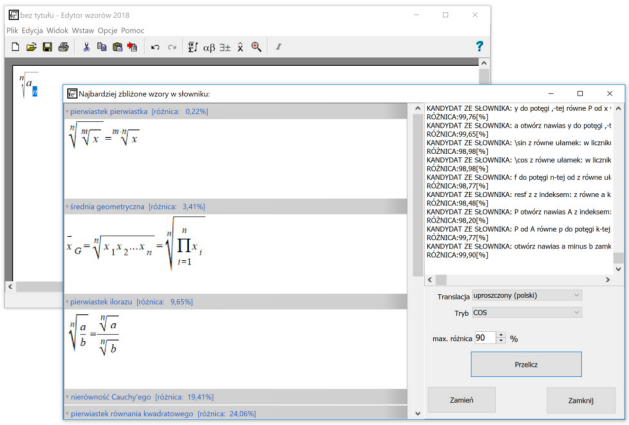


Fig. 1. The user interface of the Equation wizard („Edytor wzorów") test application with the results of the visual query.

For the second part of the test a specialized database was used in order to reveal possible weaknesses of the search algorithm or text similarity measures used for comparison. The test database consisted of 20 different formulae all containing quadratic subexpressions and having elements identical with the original formula. The column SD in Table II presents the search results for three different queries referring to the original formula $(a+b)^2 = a^2+2ab+b^2$ in different settings.

In order to compare the performance of the search algorithm for different settings, the total success rate has been calculated. The total success rate is defined to be the percentage of correctly found expressions, where correctly found means that the desired expression was among the first five entries on the list of search results. A summary of these rates is presented in Table I. The main observations are the following. The queries in lingual data formats provide the required expressions at average rate of more than 81%, while the success rate of queries processed from inner format is twice lower (around 37%). A less significant difference is observed when comparing the similarity measures used by the search algorithm - the Levenstein measure success rate is around 72%, while for cosine measure it is around 61%. However, if one looks into positioning of the queried formula, the advantage of Levenstein measure becomes more evident. Finally, the observed success rate of visual queries is significantly higher than for key word queries – nearly a half (47%) of key word queries provided desired outcome, while 86% of the visual queries appeared successful.

A few further results, presenting the overall performance issues of the proposed algorithm on a general database are given in Table III.

IV. CONCLUSIONS

The results of the experiments with the developed interactive application give the basis for the research of an efficient mathematical search engine. Both the search with the description languages as LaTeX as well as usage of verbalized text fragments was possible and accurate, especially while processing relatively not much different equations. Therefore, the proposed method is suitable for automatic or semi-automatic, supervised correction of documents containing complex mathematical notation.

On the other hand, the search for the given equation on a partial information basis, i.e. the excerpt of the formula or unordered sequence of symbols or keywords, has been also successful in many cases. Thanks to developed solution such an approximate matching is convenient and does not require any extra technical knowledge from the user, since the system accepts queries in natural language. The introduction of the advanced translator for the verbalization of the mathematical notation, taking into account the complexity of the subexpressions, dialect and national languages helped to increase the quality of text comparisons and thus the efficiency of the search.

The performed tests revealed also some general tendencies in the results obtained for different kinds of queries when compared by means of different similarity measures. The Levenstein measure seems more suitable for queries that differ slightly from the origin and therefore it could be applied for corrections. On the other hand, the cosine measure appeared to provide better matching for shorter queries, being parts of the original formula, and hence its area of application would preferably be searching or autocompletion.

One of the main directions for future experiments in this area could be the introduction of more flexible queries, possibly with the notation for logical operators as AND, OR and NOT, which could increase the accuracy of the search.

REFERENCES

[1] L. A. Chang, *Handbook for Spoken Mathematics (Larry's Speakeasy)*. Lawrence Livermore Laboratory, The Regents of the University of California, 1983.
[2] R. Fateman, "Handwriting + speech for computer entry of mathematics," *Style, Benjamin L. Kovitz, Manning Publications Company*, 2004.

TABLE II
SEARCH RESULTS FOR DIFFERENT PARTS OF FORMULA IN A GENERAL DATABASE (GD) AND SPECIALIZED DATABASE (SD) WITH RESPECT TO FORMATS OF COMPARED DATA (INNER – LATEX, ENG – NATURAL ENGLISH, POL – NATURAL POLISH), QUERY INPUT (VISUAL / KEY WORDS), AND THE MEASURE USED FOR STRING COMPARISON (L - LEV, COS - COSINE). LAST COLUMN PRESENTS THE POSITION OF THE INTENDED ORIGINAL FORMULA $(a+b)^2 = a^2 + ab + b^2$ ON THE LIST OF RESULTS (* – FORMULA OUTSIDE THE LIST OF 20 BEST MATCHES).

Query	Data format	Input Mode	GD		SD	
			Pos. L	Pos. COS	Pos. L	Pos. COS
Category: C1 (correction)						
$(a+b)^2 = a^2 + ab + b^2$	inner	visual	1	1	1	1
"square of sum of a and b equals a squared plus ab plus b squared"	inner	key words	*	*	9	7
"kwadrat sumy a i b równa się a kwadrat plus ab plus b kwadrat"	inner	key words	*	*	12	7
$(a+b)^2 = a^2 + ab + b^2$	Eng	visual	1	1	1	1
"square of sum of a and b equals a squared plus ab plus b squared"	Eng	key words	1	1	1	5
$(a+b)^2 = a^2 + ab + b^2$	Pol	visual	1	1	1	1
"kwadrat sumy a i b równa się a kwadrat plus ab plus b kwadrat"	Pol	key words	1	9	1	4
Category: C2 (autocompletion)						
$(a+b)^2$	inner	visual	14	7	4	2
square of sum of a and b	inner	key words	*	*	12	9
kwadrat sumy a i b	inner	key words	*	*	12	9
$(a+b)^2$	Eng	visual	2	3	1	2
square of sum of a and b	Eng	key words	9	*	4	9
$(a+b)^2$	Pol	visual	2	3	1	2
kwadrat sumy a i b	Pol	key words	7	8	5	8
Category: C3 (search)						
$()^2$	inner	visual	*	9	5	5
squared	inner	key words	*	*	12	7
do kwadratu	inner	key words	*	3	12	7
$()^2$	Eng	visual	8	2	1	3
squared	Eng	key words	8	2	3	6
$()^2$	Pol	visual	5	2	1	3
do kwadratu	Pol	key words	5	2	3	6

[3] ——, "How can we speak math?" Computer Science Division, EECS Department, University of California at Berkeley, Tech. Rep., 2013.

[4] J. Cuartero-Olivera, G. Hunter, and A. Pérez-Navarro, "Reading and writing mathematical notation in e-learning environments," *eLearn Center Research Paper Series*, no. 4, pp. 11–20, 2012.

[5] Z. Sroczyński, "Priority levels and heuristic rules in the structural recognition of mathematical formulae," *Theoretical and Applied Informatics*, vol. 22, no. 4, p. 273, 2010.

[6] A. Bier and Z. Sroczyński, "Adaptive math-to-speech interface," in *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*, ser. MIDI '15. New York, NY, USA: ACM, 2015. doi: 10.1145/2814464.2814471. ISBN 978-1-4503-3601-7 pp. 7:1–7:9. [Online]. Available: http://doi.acm.org/10.1145/2814464.2814471

[7] D. Attanayake, J. Denholm-Price, G. Hunter, E. Pfluegel, and A. Wigmore, "Intelligent assistive interfaces for editing mathematics." in *Intelligent Environments (Workshops)*, 2012, pp. 286–297.

[8] D. Attanayake, G. Hunter, J. Denholm-Price, and E. Pfluegel, "Novel multi-modal tools to enhance disabled and distance learners' experience of mathematics," *ICTer*, vol. 6, no. 1, 2013.

[9] T. Sancho-Vinuesa, C. Córcoles, M. Huertas, A. Pérez-Navarro, D. Marquès, R. Eixarch, and J. Villalonga, "Automatic verbalization of mathematical formulae for web-based learning resources in an on-line environment," *INTED2009 Proceedings*, pp. 4312–4321, 2009.

[10] M. Maćkowski, P. Brzoza, M. Żabka, and D. Spinczyk, "Multimedia platform for mathematics' interactive learning accessible to blind people," *Multimedia Tools and Applications*, pp. 1–18, 2017. doi: 10.1007/s11042-017-4526-z

[11] I. Kohanova, "The ways of teaching mathematics to visually impaired students," in *International Congress on Mathematical Education (ICME)*, 2008.

[12] S. Yang and Y. Ko, "Mathematical formula search using natural language queries," *Advances in Electrical and Computer Engineering*, vol. 14, no. 4, pp. 99–104, 2014. doi: 10.4316/AECE.2014.04015

[13] M. Líška, P. Sojka, and M. Ružicka, "Similarity search for mathematics: Masaryk university team at the ntcir-10 math task," in *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*. Citeseer, 2013, pp. 686–691.

[14] M. Líška, P. Sojka, and M. Růžička, "Math indexer and searcher web interface," in *International Conference on Intelligent Computer Mathematics*. Springer, 2014, pp. 444–448.

[15] J. Mišutka and L. Galamboš, "Extending full text search engine for mathematical content," *Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008*, pp. 55–67, 2008.

TABLE III
SEARCH RESULTS FOR VARIOUS VISUAL QUERIES IN A GENERAL DATABASE WITH RESPECT TO FORMATS OF COMPARED DATA (INNER – LATEX, ENG – NATURAL ENGLISH, POL – NATURAL POLISH), AND THE MEASURE USED FOR STRING COMPARISON (L - LEVENSTEIN, COS - COSINE). LAST 3 COLUMNS PRESENT THE BEST 3 MATCHES.

Query	Data format	Measure	Match #1	Match #2	Match #3
$\int_1^2 \cos x dx$	inner	Lev	$\int_a^b f(x)dx$	$Y-y=\frac{dy}{dx}(X-x)$	$(a+b)^2$
	inner	cos	$\iiint_V \left(\frac{\partial P}{\partial x}+\frac{\partial Q}{\partial y}+\frac{\partial R}{\partial z}\right)$	$\int_a^b f(x)dx$	$a_k=\frac{2}{T}\int_0^T f(x)\cos k\omega x dx$
	Eng	Lev	$\int_a^b f(x)dx$	$y'=P(x)y^2+Q(x)y+R(x)$	$a_n=a_1q^{n-1}$
	Eng	cos	$\int_a^b f(x)dx$	$a_k=\frac{2}{T}\int_0^T f(x)\cos k\omega x dx$	$\int_a^b ydx\approx h(y_0+y_1+...+y_{n-1})$
	Pol	Lev	$\int_a^b f(x)dx$	$(a+b)^2$	$y'=P(x)y^2+Q(x)y+R(x)$
	Pol	cos	$\int_a^b f(x)dx$	$\int_a^b ydx\approx h(y_0+y_1+...+y_{n-1})$	$\int_a^b ydx\approx\frac{1}{2}h(y_0+2y_1+...+2y_{n-1}+y_n)$
$\sqrt[n]{a_n}$	inner	Lev	$(a+b)^2$	$Y-y=\frac{dy}{dx}(X-x)$	$\int_a^b f(x)dx$
	inner	cos	$\sqrt[n]{\sqrt[m]{x}}=\sqrt[m\cdot n]{x}$	$\frac{a_1+a_2+...+a_n}{n}\geq\sqrt[n]{a_1a_2...a_n}$	$\sqrt[n]{x_1x_2...x_n}$
	Eng	Lev	$\int_a^b f(x)dx$	$a_n=a_1q^{n-1}$	$\sqrt[n]{\sqrt[m]{x}}=\sqrt[m\cdot n]{x}$
	Eng	cos	$\sqrt[n]{\sqrt[m]{x}}=\sqrt[m\cdot n]{x}$	$\sqrt[n]{\frac{a}{b}}=\frac{\sqrt[n]{a}}{\sqrt[n]{b}}$	$\sqrt[n]{x_1x_2...x_n}$
	Pol	Lev	$\sqrt[n]{\sqrt[m]{x}}=\sqrt[m\cdot n]{x}$	$\frac{a_1+a_2+...+a_n}{n}\geq\sqrt[n]{a_1a_2...a_n}$	$S_n=\frac{n(a_1+a_n)}{2}$
	Pol	cos	$\sqrt[n]{\sqrt[m]{x}}=\sqrt[m\cdot n]{x}$	$\sqrt[n]{x_1x_2...x_n}$	$\sqrt[n]{\frac{a}{b}}=\frac{\sqrt[n]{a}}{\sqrt[n]{b}}$
$\frac{x^2}{a^2}$	inner	Lev	$\sin z=\frac{e^{iz}-e^{-iz}}{2i}$	$\cos z=\frac{e^{iz}+e^{-iz}}{2}$	$\int_a^b f(x)dx$
	inner	cos	$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$	$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$	$(a-b)^2=a^2-2ab+b^2$
	Eng	Lev	$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$	$x=\frac{-b+\sqrt{b^2-4ac}}{2a}$	$\sqrt[n]{\frac{a}{b}}=\frac{\sqrt[n]{a}}{\sqrt[n]{b}}$
	Eng	cos	$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$	$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$	$(a-b)^2=a^2-2ab+b^2$
	Pol	Lev	$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$	$x=\frac{-b+\sqrt{b^2-4ac}}{2a}$	$\cos z=\frac{e^{iz}+e^{-iz}}{2}$
	Pol	cos	$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z$	$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$	$(a-b)^2=a^2-2ab+b^2$

[16] P. Sojka and M. Líška, "The art of mathematics retrieval," in *Proceedings of the 11th ACM symposium on Document engineering*. ACM, 2011, pp. 57–60.

[17] M. Adeel, M. Sher, and M. S. H. Khiyal, "Efficient cluster-based information retrieval from mathematical markup documents," *World Applied Sciences Journal*, vol. 17, no. 5, pp. 611–616, 2012.

[18] M. N. Quoc, K. Yokoi, Y. Matsubayashi, and A. Aizawa, "Mining coreference relations between formulas and text using wikipedia," in *23rd International Conference on Computational Linguistics*, 2010, p. 69.

[19] D. Formánek, M. Líška, M. Růžička, and P. Sojka, "Normalization of digital mathematics library content," in *Joint Proceedings of the 24 th Workshop on OpenMath and the 7 th Workshop on Mathematical User Interfaces (MathUI)*, 2012, p. 91.

[20] P.-Y. Chien and P.-J. Cheng, "Semantic tagging of mathematical expressions," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 195–204.

[21] M. E. Altamimi and A. S. Youssef, "Wildcards in math search, implementation issues." in *CAINE*. Citeseer, 2007, pp. 90–96.

[22] A. Niewiarowski and M. Stanuszek, "The mechanism of identification and classification of content (in Polish)," *Studia Informatica*, vol. 34, no. 2B, pp. 205–222, 2013.

[23] D. Połap, "Neural validation of grammatical correctness of sentences," *Ceur-ws*, 2016.

[24] S. Kozielski, M. Świderski, and M. Bach, "The use of natural language as an intuitive semantic integration system interface," in *Internet–Technical Development and Applications*. Springer, 2009, pp. 51–58.

[25] J. Jagielski and P. Wnęk, "Natural language in databases systems," *Studia Informatica*, vol. 31, no. 2B, pp. 281–290, 2010.

[26] F. Li and H. Jagadish, "Constructing an interactive natural language interface for relational databases," *Proceedings of the VLDB Endowment*, vol. 8, no. 1, pp. 73–84, 2014.

[27] R. Alexander, P. Rukshan, and S. Mahesan, "Natural language web interface for database (nlwidb)," *arXiv preprint arXiv:1308.3830*, 2013.

[28] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, Mar. 2001. doi: 10.1145/375360.375365. [Online]. Available: http://doi.acm.org/10.1145/375360.375365

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 471–478
DOI: 10.15439/2018F168
ISSN 2300-5963 ACSIS, Vol. 15

What was the Question? A Systematization of Information Retrieval and NLP Problems

Jens Dörpinghaus*, Johannes Darms† and Marc Jacobs‡
Fraunhofer Institute for Algorithms and Scientific Computing,
Schloss Birlinghoven, Sankt Augustin, Germany
Email: *jens.doerpinghaus@scai.fraunhofer.de, †johannes.darms@scai.fraunhofer.de, ‡marc.jacobs@scai.fraunhofer.de

***Abstract*—In this paper we suggest a novel systematization of Information Retrieval and Natural Language Processing problems. Using this rather general description of problems we are able to discuss and proof the equivalence of some problems. We provide reformulations of well-known problems like Named Entity Recognition using our novel description and discuss further research and the expected outcome. We will discuss the relation of two problems, cluster labeling and search query finding. With these results we are able to provide a novel optimization approach to both problems. This novel systematization approach provides a yet unknown view generating new classes of problems in NLP. It brings application and algorithmic approaches together and offers a better description with concepts of theoretical computer science.**

I. Introduction

A LOT of research in the last decades focused on the computational perspective of information retrieval and improving clusterings, partitions, search queries and document decomposing with and without feedback. Several authors like Manning et al. [1] or Clarc et al. [2] give an overview about the algorithmic part of computational linguistics and NLP. Applied researchers try to answer questions similar to "What was the question?", "What is the best description of this set of documents?", "How can we compare this and that clustering?". We realized that there are several names for the same or at least similar problems and approaches. For example Hagen et al. [3] tried to find search queries for a given set of documents and used it as a cluster labeling approach. This seems somehow obvious, as well as some other equivalent problems might also sound obvious. But nevertheless, we think a formal description and proof is necessary.

During our literature search and evaluation of several algorithms for query optimization and clustering, we realized that a formal Schema would ease the task. For finding and grouping This we propose such a schema within this work. We claim every NLP problem can be described using a five-tuple. To proof our theory a discussion on several problems and the proof of equivalent problems will follow. This novel systematic approach has a different perspective focusing on the computational view on this research area. We hope this early research will lead to a valuable discussion and more research on the theoretical and algorithmic fundamentals of natural language processing.

In table I we list some prominent NLP and IR probleme in our proposed five tuple with a corresponding description. The details of the tuple are introduced and discussed in the following chapters. However the connection between the problems can already be seen within this table.

TABLE I: Example formulations of information extraction problems as five-tuple. The first element describes the domain set, the second the domain subset of interest. The third element is a description function f. We either note this function or the image set of this function. The last entries are a feasible similarity or error measure and a reference standard. These problems are introduced in sections III and IV.

Problem Formulation	Problem Description
$\mathbb{D}\|R\|\mathbf{X}\|err\|R$	Generating of optimal Search Queries
$\mathbb{D}\|R\|\mathbf{X}\|err\|R$	Generating of optimal Cluster Labels
$\mathbb{S}\|\emptyset\|\boldsymbol{f}\|e\|\{\mathbb{S}\times[0,1]\}$	Named Entity Recognition
$\mathbb{D}\|R\|\mathbf{L}\|sim\|\emptyset$	Text summarization
$\mathbb{D}\|R\|\mathbf{K}\|sim\|\emptyset$	Keyword identification
$\mathbb{D}\|R\|\mathbf{C}\|sim\|\emptyset$	Document Clustering in $C=\{1,...,n\}$ cluster.
$\mathbb{D}\|R\|\mathbb{D}^{\mathbb{D}}\|sim\|\emptyset$	Relation Extraction
$\mathbb{D}\|R\|\mathbb{D}\|sim\|R$	Document Subset Finding Problem
$\mathbb{D}\|R\|\mathbf{G}\|sim\|\emptyset$	Parse tree

II. Notation

We want to introduce our problem description approach using a five-tuple. Therefore we define a domainset $\mathbb{D}$ and subset $R\subseteq\mathbb{D}$, a description set $\mathbb{X}$ and a description function $f:\mathbb{D}\rightarrow\mathbb{X}$, an evaluation function $e:\mathbb{E}\rightarrow[0,1]$ and a reference standard E. Hence NLP problems can be given as:

$$p=\mathbb{D}|R|f:\mathbb{D}\rightarrow\mathbb{X}|e:\mathbb{E}\rightarrow[0,1]|E \tag{1}$$

At first we have to introduce and describe the notation and sets. Some examples and applications will be provided within the next sections. For an illustration of sets and functions we refer to figure 1.

A. Domain Set

Let $\mathbb{D}$ be a finite *domain set* containing all instances of a Probleme, e.g documents, text data, speech or any other semantic content. A definition is $\mathbb{D}=\{d_1,...,d_n\}$ where d_i is a vector of documents or semantic data. We may see a textual

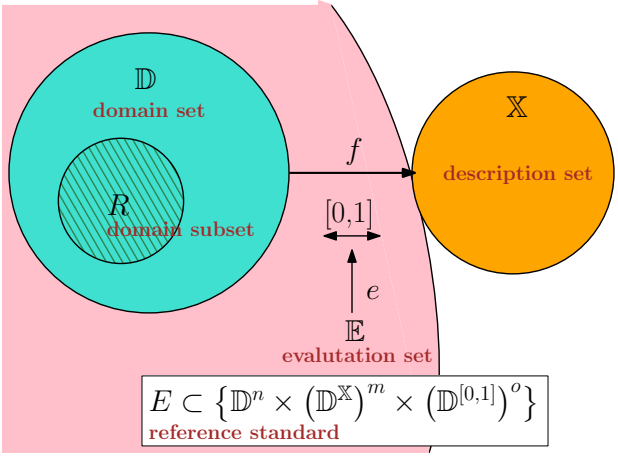


Fig. 1: Illustration of sets in equation 1. If not feasible, these sets or functions can be set to $\emptyset$ or id. The cut between $\mathbb{X}$ and the other sets is not necessarily empty.

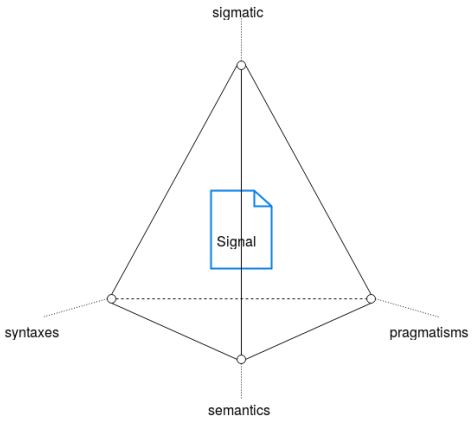


Fig. 2: Illustration of an SDA. All four properties, (a) Sigmatics (identification) (b) Pragmatics (what is being represented) (c) Semantics (what is being represented) and (d) Syntax (how is the signal constructed), are wrapped around the digital asset to provide more meta-data. All five elements for the semantic digital asset which. Multiple SDAs connect if they share at least one of these elements.

document d as a vector containing n meta data as well as full text etc., describing a document as follows

$$d_i = \begin{pmatrix} d_i^1 \\ d_i^2 \\ d_i^3 \\ \vdots \\ d_i^{n_i} \end{pmatrix} = \begin{pmatrix} \text{title} \\ \text{authors} \\ \text{fulltext} \\ \vdots \\ \text{NE} \end{pmatrix}$$

Here $\mathbb{D} = L^{\mathfrak{n}}$ with $\mathfrak{n} = \max_i n_i$ is the vector space of $\dim(\mathbb{D}) = \mathfrak{n}$ data fields of a natural language L. We may also store binary data within this vector. For a better generalization we can set $\mathbb{D} = \mathcal{P}(\mathbb{S})$ as the vector space of semantic digital assets. Here $\mathcal{P}$ denotes the power set. In this case a document d is a list $d = \{s_1, ..., s_n\}$ of semantic digital assets (SDA) $s_i \in \mathbb{S}$. These SDAs are an optimal carrier for meta-data or annotations.

A semantic digital asset can be defined as "an asset that exists only as a numeric encoding expressed in binary form" [4]. This definition includes text, images, sound files, tables, and so on. In a nutshell we can conclude that "digital assets include any electronically stored information" [5]. In addition some meta data is included. Thus we can describe a SDA as a variation of the information tetrahedron introduced in [6] where four semiotic properties are wrapped around each signal. These semiotic properties are (a) Sigmatics (b) Pragmatics (c) Semantics and (d) Syntax, see figure 2. As described by Hodapp et al. in [7] SDAs are highly flexible and can be easily connected. The hierarchy is connected into annotations. Applications can be found in [7] and [8].

Since it is of crucial importance, we will discuss the hierarchical connection in a nutshell, but refer to [7] for further information. Let a and b be defined with Sigmatics *sentence:S153:1322066041* and *sentence:S153:1322066041* – the first one with the Pragmatics *sentence* and the second one *Mus_musculus*. Here, the first SDA contains the sentence in natural language, for example "Spontaneous antepartal RhD alloimmunization" whereas the signal of the second one contains the named entity information *{"begin": 24,"end": 27,"attr": "original","ref": "MM137098:Rhd"}*. Thus it is not really necessary to store explicit relations between SDAs, since they are implicitly given in the structure of SDAs.

B. Domain Subset

To find a proper problem description, we can either focus on the complete domain set $\mathbb{D}$ or a subset $R \subseteq \mathbb{D}$ This can either be a manually created subset or created by semi-automatic tools. One can imagine the result of a search query.

In addition, we allow a ranking of elements in R in the intervall $[0, 1]$ of real numbers. Then $R = R' \times [0, 1]$ with $R' \subseteq \mathbb{D}$.

C. Description Function

If necessary, we may also add a *description function* f for documents or subsets of $\mathbb{D}$. Given a description set $\mathbb{X}$, this function can have several forms, in general denoted by $f : \mathbb{D} \to \mathbb{X}$. In our short notation we can either note the function f or the set $\mathbb{X}$ if we need to focus on this set. If both information are needed, we can write $f, \mathbb{X}$.

If we want to map elements in R to a meta data subset of a document $d \in \mathbb{D}$ like publishers, authors etc. f has the form $f : \mathbb{D} \to D$ with $\dim(D) \leq \dim(\mathbb{D})$ and $f(d_i) = f_i^j$. This may also be a combination of vector entries.

A description function may also return several discrete values, for example *true* or *false*. In this case f is given by $f : \mathbb{D} \to N \subset \mathbb{N}$. If we want to describe concepts from a terminology T, f is given by $f : \mathbb{D} \to T$. The function may also return words σ^* from a language L which leads to $f : \mathbb{D} \to \Sigma^*$. Here Σ^* denotes the set of all words (or strings)

over a given alphabet or language. We can even consider a subset from the language which leads to $f : \mathbb{D} \to L$. If we do not need a description function we can simply set $f = \text{id}$ to the identity function. As we can see the cut between $\mathbb{X}$ and the other sets is not necessarily empty.

In some cases it is useful to assume that $\mathbb{X} = f(\mathbb{D})$ and thus q as a surjective mapping. It follows that f has a right inverse q with $f \circ q = \text{id}_{\mathbb{X}}$. This function is necessary to model some problems. Usually we cannot assume that f is also an injective mapping: A description set $x \in \mathbb{X}$ may have more than one origin in $\mathbb{D}$.

Considering an element $\mu \in \mathbb{X}$ and $\mathbb{X}$ as the description set of a search engine, R can be explicitly set as $E = q(\mu)$ with the right inverse of f. Here q denotes the search function with $q : \mathbb{X} \to \mathbb{D}$.

D. Evaluation Function

For several problems we need an *evaluation function* $e : \mathbb{E} \to [0, 1]$ which is either a similarity measure sim, an error measure err or a weight $weight$. If it is not applicable, we may use the identity function id. The set $\mathbb{E}$ must be set according to our optimisation goal. If we optimise $\mathbb{D}$, for example by adding new documents or additional information, $\mathbb{E} = \mathbb{D} \times \mathbb{D}$. The same holds, if we want to find an optimal subset $R \subset \mathbb{D}$.

If we want to optimise our description function f, we must use the function space $\mathbb{D}^{\mathbb{X}} = \mathbb{E}$. An evaluation of the reference standard will be even more complex, see below. Then $\mathbb{E} = E$ applies.

E. Reference Standard

Usually the evaluation process cannot be done without an external criterion. In this cases we can add a *reference standard* or *evaluation set*

$$E \subset \left\{ \mathbb{D}^n \times \left(\mathbb{D}^{\mathbb{X}}\right)^m \times \left(\mathbb{D}^{[0,1]}\right)^o \right\}$$

to optimize our result. We either have one single subset of $\mathbb{D}$, or two subsets – a positive and a negative reference standard. We may also have a ranked list of subsets of $\mathbb{D}$. A description function in the function space $\mathbb{D}^{\mathbb{X}}$ or n of them could also be set as a reference standard, as well as one or o evaluation functions out of the function space $\mathbb{D}^{[0,1]}$. This is sometimes denoted as one or many *gold standards*. This can be very complex, but usually problems only need one of these sets.

If not feasible or unused, we may also set $E = \emptyset$.

F. Problem Description

Natural Language Processing problems can thus be described by a five-tuple. We can denote them by a combination p of

$$p = \mathbb{D}|R|f : \mathbb{D} \to \mathbb{X}|e : \mathbb{E} \to [0, 1]|E$$

with a domain set $\mathbb{D}$, a domain subset R, a description set $\mathbb{X}$ and a description function f as well as an evaluation function e evaluating on the set $\mathbb{E}$ according to the reference standard E. We usually have four parameters given and want to obtain an optimal solution for the fifth. The optima result with respect to the problem will be denoted in bold letters. We can add an additional index for ambiguous notations.

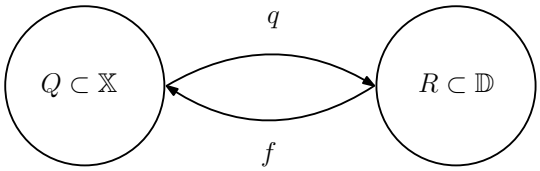
If we have an optimal algorithm we only need one computation step. If we have a heuristic returning approximate values, we may use the output of the first iteration as an input for the next iteration. We will discuss, that similar approaches usually only differ with respect to the chosen set $\mathbb{X}$.

III. Search Queries and Cluster Labels

A. Generating and optimisation of Search Queries

In this paper we use a very generic definition of search engines and search queries. A search engine is a function $q : \mathbb{X} \to \mathbb{D}$ which outputs a set of documents or any other content of the domain set if the input is a subset of a description set $\mathbb{X}$ which we cal search query.

The problem of generating search queries usually has a domain set $\mathbb{D}$ restricted by the database of the search engine. The return value of our problem is a search query $\mu \in \mathbb{X}$ so that $q(\mu) = R$. Thus R is the subset of documents for which we want to create a search query. We have a mapping from



one element in $\mathbb{X}$ to a subset of $\mathbb{D}$. q is thus the right inverse of the description function f and $f \circ q = \text{id}_{\mathbb{X}}$. Not only does f have to be surjective, but we also have to assume that even q is surjective. Every document in the target set $\mathbb{D}$ should be a target of some search query.

It is very easy to see that this is usually not given in reality: Assume q is a websearch, $\mathbb{X}$ the web search description and $\mathbb{D}$ the set of all web pages available. Some of them may not be indexed due to restrictions made to the robots crawling and indexing the web. We can sail around this by restricting $\mathbb{D}$ to $q(\mathbb{X})$. Then f should be the right inverse of q with $q \circ f = \text{id}_{\mathbb{D}}$.

We can also see, that $\forall d \in \mathbb{D}$ several $\mu_1, ..., \mu_n \in \mathbb{X}$ exists with $d \in q(\mu_i)$ – neither q nor f are injective mappings. If we want to find the optimal μ we need to define some sort of metric on elements in $\mathbb{X}$. This can be very complex. If we assume, that we have a terminology T and a simple algebra with $\vee$ and $\wedge$, we can simplify $\mathbb{X} = \mathcal{P}(T, \vee, \wedge)$ and take the length of $\mu \in \mathbb{X}$ as a metric. But if all documents have a unique index stored in $\mathbb{X}$ the shortest search query might consist of a concatenation of these indexes listing all documents in R.

Thus, the simplest evaluation function $e : \mathbb{D} \to [0, 1]$ is set by

$$err_1(d_i, d_j) = \begin{cases} 1 & i \neq j, f(d_i) \neq f(d_j), d_i, d_j \in R \\ 1 & i \neq j, f(d_i) = f(d_j), d_i \text{ or } d_j \in R \\ 0 & \text{else} \end{cases}$$

If f, the description function with the image set $\mathbb{X}$, does not map two documents in R to the same element, which is the

search query $\mu \in \mathbb{X}$, we count an error. Same happens, if another document not in R is mapped to R. Thus we want to find a description function f so that $f(R) = \mu \in \mathbb{X}$ with $q(\mu) = R$. It follows that the problem is given by

$$p = \mathbb{D}|R|\mathbf{X}|err_1|R$$

This is the simplest formulation of the stated problem. As discussed, it can be more complex. We have not defined a proper quality measure for search queries $\mu \in \mathbb{X}$. In addition, the space $\mathbb{X}$ may be very complex and it is not clear, if it is – like $\mathbb{D}$ – a discrete space with a proper metric. In addition, although f is a surjective mapping and q can be set to be surjective, it is left open, if one of these mapping might also be injective.

B. Generating and optimisation of Cluster Labels

A clustering is usually done on a domain set $\mathbb{D}$ and leads to several clusters $C_1, ..., C_n$, $n \in \mathbb{N}$. If $\mathbb{D} = \mathcal{P}(\mathbb{S})$, these clusters are explicitly coded in the set $\mathbb{D}$. Finding cluster labels is the task of assigning a subset of a description set $\mathbb{X}$ with the description function $f : \mathbb{D} \to \mathbb{X}$ to a cluster $R \in \{C_1, ..., C_n\}$. We might consider an evaluation function measuring the distance between the description between two documents in R, $|f(d_i) - f(d_j)|$. But we need to assume a proper metric on $\mathbb{X}$ to do so. This leads to very complex questions. For example: What is a proper metric on a space of boolean algebra? The easiest evaluation function is thus given by

$$err_2(d_i, d_j) = \begin{cases} 1 & i \neq j, f(d_i) \neq f(d_j), d_i, d_j \in R \\ 1 & i \neq j, f(d_i) = f(d_j), d_i \text{ or } d_j \in R \\ 0 & \text{else} \end{cases}$$

Here we define that every two documents in R must share the same cluster labels. This cluster label has to be unique to this cluster. The reference standard can also be set to R. Thus the problem of generating and optimisation of cluster labels is given by

$$p = \mathbb{D}|R|\mathbf{X}|err_2|R$$

where the resulting label set is the image $f(R) \subset X$. Depending on the choice of X this either leads to a set of metadata, terms, sentences or any subset of natural language. Again, this problem can be very complex.

C. Search Queries and Cluster Labels are closely connected

In our introduction we already discussed, that Hagen et al. found out that both problems are similar, see [3]. It is easy to proof that given the same domain set $\mathbb{D}$, image set $\mathbf{X}$ of the description function and the same evaluation function both problems are equivalent. Thus, they are closely connected.

Lemma III.1. *Let $\mathbb{X}$ be a description image set. For every solution f of $p_1 = \mathbb{D}|R|\mathbf{X}|err_1|R$ this is also an optimal solution of $p_2 = \mathbb{D}|R|\mathbf{X}|err_2|R$.*

Proof. This follows directly, since $err_1 = err_2$. □

Same follows directly for the inverse:

Lemma III.2. *Let X be a description image set. For every solution f of $p_2 = \mathbb{D}|R|\mathbf{X}|err_2|R$ this is also an optimal solution of $p_1 = \mathbb{D}|R|\mathbf{X}|err_1|R$.*

Thus both problems are equivalent if we consider the same domain set $\mathbb{D}$, image set X of the description function and the same evaluation function. We can conclude that we can use the same or similar heuristics for solving both problems. Usually a search query language is not used for representing cluster labels. But since query languages and natural languages are not only highly connected but merge more and more (see [9] or [10]) we follow that in future both problems will be even more connected. We will now discuss a small example.

D. Example

We will do a generation and optimization of cluster labels with a similar approach to Borkowski et al. [11], Kanavos et al. [12] and Demner et al. [13]. All of them use a taxonomy of categories like Medical Subhect Headings (MeSH, see https://www.nlm.nih.gov/mesh/) and process the documents using tfidf-method. We will use SCAIView, see [14] or https://www.scaiview.com), an information retrieval system for knowledge discovery for a similar approach. SCAIView was used in many recent research projects, for example regarding neurodegenerative diseases [15], brain imaging features [16] and other theoretic research like document clustering, see [17]. The advantage is, that SCAIView already provides us with Named Entities for MeSH but also other ontological representing biomedical entities. Thus we get a better coverage of text with named entities.

Our domain set $\mathbb{D}$ is MEDLINE data, and R a subsets of MEDLINE data. MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database maintained by the National Center for Biotechnology Information and covers a large number of scientific publications from medicine, psychology, and the health system. For the clustering use case, we study MEDLINE abstracts and associated metadata that are processed by ProMiner, a named entity recognition system, see [18], and indexed by the semantic information retrieval platform SCAIView.

Our goal is, to find a unique representation of R in $\mathbb{X}$. Let $f(d) = \mu$ for all $d \in R$. We have to define the description set $\mathbb{X}$.

Borkowski et al. [11] processed a ranked list of categories with their weights. We will follow Kanavos et al. [12] and use all ontologies available at SCAIView. To make our approach easier, we will limit our image set $\mathbb{X}$. Let $\mathbb{X}$ be the SCAIView search query set limited to NE.

R is the document set retrieved by a list of pubmed identifiers. In our example, we have R as the result of a list of 38 PMIDs. $\mathbb{D}$ is the set of all documents in SCAIView databse. Querying the Lucene backend we find a list of 654 NE $N = n_1, ..., n_{654}$ and the documents containing them, which we donate by $l(n_i)$. The list $n_i : l(n_i)$ has the form

```
CHEBI:36929 : [38]
MESH:E05.598.500 : [9, 11, 18, 36]
ENTREZGENE:3630 : [29, 37]
MESH:C10.597.742 : [9]
ENTREZGENE:387244 : [34]
MESH:C10.597.622 : [14]
MESH:F03.900.675.400 : [12]
MGI:96543 : [6, 15, 29, 32, 33]
CHEBI:6271 : [26]
...
```

We will now use a novel set covering approach. Following [12] the labels for distinct subsets can be seen as potential candidates for cluster labels. For example we can cover r with n terms:

```
 1   ;MESH:F01.700.039
 1   ;MESH:C10.597.606.057
 1   ;SWISSPROT:HUMAN
 1   ;MESH:D059445
 1   ;MESH:C23.888.592.604.039
 2   ;CHEBI:36929   ;MESH:F01.700.039
 2   ;CHEBI:36929   ;MESH:C10.597.606.057
...
```

If we use sim_1 as an evaluation measure, we are nearly done. But we might get more documents in $\mathbb{D}$ by querying these labels.

We can now construct a hierarchical tree using the logical operators *and* and *or* in $\mathbb{X}$. We will do this by considering a graph $G = (V, E)$ with nodes $V = N$, in our example $V = \{n_1, ..., n_{654}\}$. We add weighted edges between two nodes n_i, n_j if $l(n_j) \subset l(n_i)$. The weight is set to zero if $\nexists n_k \in N$ such that $l(n_j) \subset l(n_k) \subset l(n_i)$. Otherwise we set the weight $w(n_i, n_j)$ to the largest number $\mathfrak{w}$ so that $\mathfrak{w}$ elements in N exist with $l(n_j) \subset l(n_1) \subset ... \subset l(n_{\mathfrak{w}}) \subset l(n_i)$. Thus the weight is zero if the document set is a direct subset of the other document set. Otherwise, it is the largest number of subsets that lie in between. Finding the minimum spanning tree(s) in this graph G and connecting all nodes with `AND` and the `OR` of their child nodes lead to the solution μ:

```
MESH:F01.700.039
AND MESH:C10.597.606.057
AND SWISSPROT:HUMAN
AND MESH:D059445
AND MESH:C23.888.592.604.039
AND (
  MESH:E05.598.500 AND (MGI:95574 OR MESH:C10
      .597.742)
  OR ENTREZGENE:3630 AND (MGI:96542 OR MESH:C19
      .246.300)
  OR
  ...)
```

The description function f is a heuristic finding one spanning tree on G with the named entities covering the domain subset R.

This is both: a correct solution of clustering labeling of R on $\mathbb{X}$ obtained by f as well as a possible solution of a search query so that $q(\mu) = R$.

As we can see, even this simple approach needs a complex heuristic. Although finding minimum spanning trees is usually in $\mathcal{FP}$, we can construct more complex examples that are $\mathcal{NP}$-complete. It would be very beneficial to find problems that are in $\mathcal{P}$.

This approach can now very easily be transferred into natural language, although a very complex boolean algebra might not be very helpful to human readers. This leads to another problem we already discussed: How is the space $\mathbb{X}$ defined? Is it a metric space? Is q an injective mapping? This reformulation of both problems is very helpful to discuss and proof the complexity and the real underlying problems and to find more suitable heuristics and algorithms. But it also leads to new questions and problems.

IV. More information extraction problems

Information extraction problems can be transferred into $p = \mathbb{D}|R|\mathbb{X}|sim|\emptyset$ with a result information description image set $\mathbb{X}$ for R. Here f is a function that extracts some information out of a document d: This may be natural language $f : \mathbb{D} \to L$ or another subset of $\mathbb{D}$ or a mapping to ontologies or terminologies.

We will discuss some examples and point out the benefits of our new approach.

A. Named Entity Recognition

Named Entity Recognition (NER) was initial proposed as the task to identify names, location and temporal constructs in text [19]. Over the decades this initial definition expanded to detect arbitrary concepts, defined as things of thought [20]. Different Algorithms developed and adapted to NER over time from simple gazetters [21] over rule based engines [18] and probabilistic context-free grammar [22] to Conditional Random Field [23] and neuronal networks [24]. No matter on how the algorithm solves the problem in the end it needs to link a sequence of character to one or many concepts. Therefor a model, a function, is constructed that encode how this should happen. This model is either generated by manual labour or my machine learning approaches or mixtures in between.

The evaluation of NER applications is often done by comparing an obtained result to a reference (gold) standard [25]. For this comparison a function $e : \{\mathbb{S} \times [0, 1]\} \times \{\mathbb{S} \times [0, 1]\} \to \{0, 1\}$ is needed that assesses if a Named Entity (NE), a Concept, is correctly detected or not. Based on the discrete values of the function Precision, Recall and a F-score are computed and are used as a performance indicator [25]. For the definition of e we assume that the target set of the description function, performing the NER, matches the definition of a reference standard. This allows to use a result of a description function a as a reference for a different function.

As initial mentioned NER is the task to link a sequence of character to concepts. Subsequent the description function looks like $f : \Sigma^* \to Concept$. If we assume a Concept is encoded as an SDA the function is a mapping from a sequence of character to SDAs. We also encode sequences of characters as SDAs so the function can be refined as $f : \mathbb{S} \to \mathbb{S}$. To also encode the uncertainty of such a mapping the final function is $f : \mathbb{S} \to \{\mathbb{S} \times [0, 1]\}$. To fit into the proposed schema, the

function needs to map from a document $\mathbb{D}$, thus we define $\mathbb{D} = \mathcal{P}(\mathbb{S})$ and $f : \mathbb{D} \to \{\mathbb{S} \times [0, 1]\}$

We transferred the application $p = \mathbb{D}|\emptyset|f|\emptyset|\emptyset$ and evaluation $p = \mathbb{D}|\emptyset|f|e|\{\mathbb{S}\times[0, 1]\}$ phase of NER into the initial proposed five-tuple. Likewise we can decode a learning phase, therefore we define a training set $T = \{\mathbb{S}^1 \times [0, 1]\}, \mathbb{S}^1 \subset \mathbb{D}$. The function f than can be learned resp. optimized by using the training set T and the evaluation function e. Subsequent a learning phase is expressed as $p = \mathbb{D}|T|\mathbf{f}|e|\emptyset$. Combing all three phases the initial proposed five tuple is constructed:

$$p = \mathbb{D}|T|\mathbf{f}|e|\{\mathbb{S} \times [0, 1]\}$$

We like to illustrate this with a gazatter based approach. We choose gazetter as an example because it is simple to understand and it eases discussion about extending to more advanced methods. Assume we have a gazetter, a list, with n Concepts $g = g_1, g_2, ..., g_n$. Each Concept $g_i = \{g_{i1}, g_{i2}, g_{im_i}\}$ is a set of alternative names (pairwise disjoint), where g_{i1} is the Representative. The description function f maps SDAs that encode sequences of character to a SDA that encode the Representative of a Concept. For a simple gazetter the function could look like the function below where an exact string match [26] between the character sequence and an alternative name is required.

$$f(n) = \begin{cases} \{S(g_{11}), 1\} & \text{if } \exists g_{1j} = n, 1 \leq j \leq m_1 \\ \{S(g_{21}), 1\} & \text{if } \exists g_{2j} = n, 1 \leq j \leq m_2 \\ \vdots & \vdots \\ \{S(g_{n1}), 1\} & \text{if } \exists g_{nj} = n, 1 \leq j \leq m_n \end{cases}$$

This function can be extend to a fuzzy string matching [27]. The fuzziness can be encoded via a normalized edit distance [27] in the second argument. Orthogonal an extension on the used data is possible. The function could work on Token , Stems or Lemmas [25] instead of character sequences. This requires a preprocessing of the gazetter as well as a transformation of SDA. Or alternatively a minor refinement of the tuple, switching from character SDAs to e.g. Token SDAs. However this is possible within the proposed five tuple and shows the generality of our systematization. The same transformation approach can be used to directly incorporate various machine learning methods, like [23], [24], [22]. The SDA, of the training set can be decoded into a better suited feature representation for the used method and the results can also be transformed into SDAs. Alternatively the method could directly use SDA as features.

As it can be see the systematization is a common base for various methods. Various methods can be transferred into this tuple representation by a transforming the data from and into an SDA. This shows the strength and flexibility of SDAs and the proposed tuple.

B. Text summarization

Text summarization is the task of assigning a short summary in natural language L to a document $d \in \mathbb{D}$. Thus our description set $\mathbb{X} = L$ and the complete problem has the form

$$p = \mathbb{D}|R|\mathbf{L}|sim|\emptyset$$

with a result information description $f(R)$ for R. Here f is a function that extracts some information in form of language out of a document d: $f : \mathbb{D} \to L$. Once again we have to ask how our evaluation function sim works. Is sim just the vector distance in a vector-space representation of L? If we limit L to a list of terms, a terminology or ontology summarising the document, this might be suitable. But considering the context of text might be more helpful. We can find several examples in literature: Demner et al. [13], who generated extractive summaries for abstracts of documents in MEDLINE. Barzilay et al. [28] did use lexical chains, without considering the semantic interpretation. Gong et al. [29] rather explicitly considered the semantic of the texts.

Thus all these approaches differ in the definition of the domain set $\mathbb{D}$. It may contain a simple list of texts or abstracts, but it may as well contain semantic digital assets $\mathbb{S}$ considering the semantics and context. In addition the description function is another criterion for distinction. We may add them as additional index to $\mathbb{X}$. This leads to $p = \mathbb{D}|R|\mathbf{L}_{lexical\_chains}|sim|\emptyset$ for [28] or $p = \mathbb{S}|R|\mathbf{L}_{semantics}|sim|\emptyset$ for [29].

This novel problem description provides a helpful framework for sorting the approaches found in literature.

C. Relation Extraction

The task of extracting relational facts – or relations – from a text is the combination of two or multiple entities in a computable format. This is usually either done by manual curation (see [30]) or by supervised or unsupervised learning (see [31] or [32]). Relation extraction is widely used in biomedical research, biology or toxicology to handle the growth of publications and data available. In addition it is used in medical research, see [33]. After relation extraction computable networks are created, see [32].

Let $\mathbb{D} = \mathcal{P}(\mathbb{S})$ be a set of documents denoted by SDAs. Our description set contains all relations between SDAs. Thus it is the set of all functions from SDAs to SDAs: $\mathbb{X} = \mathbb{D}^{\mathbb{D}}$. In addition we need a similarity measure that maps relations from documents to SDA terms:

$$sim : \mathbb{X} \to [0, 1]$$

Then doing Relation Extraction is the task of finding an optimal description function

$$f : \mathbb{D} \to \mathbb{X}$$

that either maps SDAs for sentences to one or more relations between SDAs or to 0 if a SDA has no NE. Thus we find

$$p = \mathbb{D}|R|\mathbb{D}^{\mathbb{D}}|sim|\emptyset$$

A lot of question have to be left open: How can we define f according to the solutions found in literature? Could $\mathbb{X}$ be reduced to a subset without loosing information? Once again a systematization of approaches found in literature could be made, although this will be part of future work.

V. Discussion and further research

We proposed a novel formal schema for information retrieval and natural language processing problems and reformulated several well-known problems. Our schema is helpful to sort NLP-problems according to their underlying and inherent structure and to identify the complex parts to solve the problem.

Discussing the equivalence between cluster labelling and finding search query we proofed that they are – obviously – equivalent if they share the same description set and the same evaluation function. This directly leads to the conclusion, that most NLP-problems have a core problem that can be solved with distinct heuristics and algorithms. Finding an evaluation function and a description set is not a core problem of computer science, but deeply related to linguistics and applied computer science. Our new approach will help to group problems and foster synergies for optimization and offer a better description with terms of theoretical computer science. Here we already reduced a simplified search query problem to a graph problem.

We left several open questions. Further research has to be done with focus on time and space complexity – what is the computational complexity in these natural language problems? Here the integration of formal language theory will be the next step. Also unsupervised and supervised learning can be expressed with our novel approach, more research has to be done regarding this. In addition, our paper is based on text data. But we can also express binary data such as speech and images in $\mathbb{D}$.

In this paper we could only discuss some early work on the preliminaries and provide a few short examples. We hope that the impact of our schema is a better categorization of NLP-problems and a better communication between application and theoretical informatics, leading to more efficient algorithms and heuristics.

References

[1] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] A. Clark, C. Fox, and S. Lappin, *The handbook of computational linguistics and natural language processing*. John Wiley & Sons, 2013.

[3] M. Hagen, M. Michel, and B. Stein, "What was the query? generating queries for document sets with applications in cluster labeling," in *International Conference on Applications of Natural Language to Information Systems*. Springer, 2015, pp. 124–133.

[4] D. Babeanu, A. A. Gavrila, and V. Mares, "Strategic Outlines: Between Value And Digital Assets Management," *Annales Universitatis Apulensis: Series Oeconomica*, vol. 11, no. 1, p. 318, 2009.

[5] J. P. Hopkins, "Afterlife in the Cloud: Managing a Digital Estate," *Hastings Science and Technology Law Journal*, vol. 5, p. 209, 2013.

[6] H. Malissa, "Automation in und mit der Analytischen Chemie IV," *Fresenius' Zeitschrift für analytische Chemie*, vol. 256, no. 1, pp. 7–14, Feb. 1971.

[7] M. Jacobs, S. Hodapp, and J. Dörpinghaus, "SDA: Towards a novel Knowledge Discovery Model for Information Systems," in *Proceedings of the 11th IADIS International Conference Information Systems 2018*. IADIS, 2018, pp. 300–302.

[8] J. Dörpinghaus, M. Jacobs, and J. Fluck, "Graph based Discovery in biomedical Information Systems connecting scientific Texts with structured Expoert Knowledge," in *Proceedings of the 11th IADIS International Conference Information Systems 2018*. IADIS, 2018, pp. 297–299.

[9] D. Suryanarayana, S. M. Hussain, P. Kanakam, and S. Gupta, "Natural language query to formal syntax for querying semantic web documents," in *Progress in Advanced Computing and Intelligent Engineering*. Springer, 2018, pp. 631–637.

[10] D. Melo, I. P. Rodrigues, and V. B. Nogueira, "Semantic web search through natural language dialogues," in *Innovations, Developments, and Applications of Semantic Web and Information Systems*. IGI Global, 2018, pp. 329–349.

[11] P. Borkowski, K. Ciesielski, and M. A. Kłopotek, "Semantic classifier approach to document classification," *arXiv preprint arXiv:1701.04292*, 2017.

[12] A. Kanavos, C. Makris, and E. Theodoridis, "Topic categorization of biomedical abstracts," *International Journal on Artificial Intelligence Tools*, vol. 24, no. 01, p. 1540004, 2015.

[13] D. Demner-Fushman and J. Lin, "Answer extraction, semantic clustering, and extractive summarization for clinical question answering," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 841–848.

[14] E. Younesi, L. Toldo, B. Müller, C. M. Friedrich, N. Novac, A. Scheer, M. Hofmann-Apitius, and J. Fluck, "Mining biomarker information in biomedical literature," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 148, 2012.

[15] M. A. E. K. Emon, R. Karki, E. Younesi, M. Hofmann-Apitius *et al.*, "Using drugs as molecular probes: A computational chemical biology approach in neurodegenerative diseases," *Journal of Alzheimer's Disease*, vol. 56, no. 2, pp. 677–686, 2017.

[16] A. Iyappan, E. Younesi, A. Redolfi, H. Vrooman, S. Khanna, G. B. Frisoni, and M. Hofmann-Apitius, "Neuroimaging feature terminology: A controlled terminology for the annotation of brain imaging features," *Journal of Alzheimer's Disease*, vol. 59, no. 4, pp. 1153–1169, 2017.

[17] J. Dörpinghaus, S. Schaaf, J. Fluck, and M. Jacobs, "Document clustering using a graph covering with pseudostable sets," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017, pp. 329–338.

[18] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "ProMiner: rule-based protein and gene entity recognition." *BMC bioinformatics*, vol. 6 Suppl 1, p. S14, 2005.

[19] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, vol. 1, 1996.

[20] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[21] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 147–155. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596374.1596399

[22] J. R. Finkel, A. Kleeman, and C. D. Manning, "Efficient, feature-based, conditional random field parsing," *Proceedings of ACL-08: HLT*, pp. 959–967, 2008.

[23] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius, "Named entity recognition with combinations of conditional random fields," in *Proceedings of the second biocreative challenge evaluation workshop*, 2007.

[24] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[25] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.

[26] C. Charras and T. Lecroq, *Handbook of exact string matching algorithms*. Citeseer, 2004.

[27] G. Navarro, "A guided tour to approximate string matching," *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.

[28] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," *Advances in automatic text summarization*, pp. 111–121, 1999.

[29] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 19–25.

[30] J. Fluck, S. Madan, S. Ansari *et al.*, "Belief-a semiautomatic workflow for bel network creation," in *Proc. 6th Int. Symp. Semant. Min. Biomed*, 2014, pp. 109–113.

[31] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 1003–1011.

[32] J. Fluck, S. Madan, S. Ansari, A. T. Kodamullil, R. Karki, M. Rastegar-Mojarad, N. L. Catlett, W. Hayes, J. Szostak, J. Hoeng *et al.*, "Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (bel)," *Database*, vol. 2016, p. baw113, 2016.

[33] F. Rinaldi, T. R. Ellendorff, S. Madan, S. Clematide, A. Van der Lek, T. Mevissen, and J. Fluck, "Biocreative v track 4: a shared task for the extraction of causal network information using the biological expression language," *Database*, vol. 2016, 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 479–488
DOI: 10.15439/2018F140
ISSN 2300-5963 ACSIS, Vol. 15

A neural framework for online recognition of handwritten Kanji characters

Małgorzata Grębowiec, Jarosław Protasiewicz
National Information Processing Institute, Warsaw, Poland

***Abstract*—The aim of this study is to propose an efficient and fast framework for recognition of Kanji characters working in a real-time during their writing. Previous research on online recognition of handwritten characters used a large dataset containing samples of characters written by many writers. Our study presents a solution that achieves fine results, using a small dataset containing a single sample for each Kanji character from only one writer. The proposed system analyses and classifies the stroke types appearing in a Kanji and then recognises it. For this purpose, we utilise a Convolutional Neural Network and a hierarchical dictionary containing Kanji definitions. Moreover, we compare the histograms of Kanjis to solve the problem of distinguishing character having the same number of strokes of the same type, but arranged in a different position in relation to each other. The proposed framework was validated experimentally on online handwritten Kanjis by beginners and advanced learners. Achieved accuracy up to 89% indicates that it may be a valuable solution for learning Kanji by beginners.**

I. Introduction

KANJI, along with the syllabic kana - hiragana and katakana, belong to the Japanese writing system. They are adopted logographic Chinese characters and literally mean "Han characters" (漢字). The number of Kanji characters is vast and amounts to over 500,000 [1]. However, in order to understand Japanese newspapers and books, it is usually enough to know 2,136 of *jouyou kanji* from the official list of Kanji determined by the Ministry of Education of Japan in 2010.

A large number of Kanji characters and a complex structure could require a lot of time to learn them. One of the solutions that may help the students in that may be an intelligent application. Its main idea is that it analyses online Kanji characters when a student is writing them stroke by stroke. The analysis relies on checking the correctness of the drawn characters or suggesting their meaning. To make this possible, such a system has to recognize the handwritten Kanjis online.

There are several approaches to this issue. Early solutions for online recognition of handwritten Kanji presented in [2] emphasize the similarity of the sequence of pen movements to the problem of speech recognition. Due to this similarity, they utilised a Hidden Markov Model (HMM) for a sub-stroke of Kanji and built a hierarchical dictionary defining the characters' structure. For further improvement of the recognition accuracy, [3] used pen pressure to propose writer-independent handwriting recognition. The further studies of [4], [5] have expanded this solution, taking into account the relative position of the strokes in Kanji. Study of [6] also is based on HMM recognition for structured character pattern representation.

We have to underline that the approaches mentioned above deal very well with the issue of Kanji recognition. However, some improvements may be required to gain better efficiency, especially when considering a solution that is able to work in real-time. One of the above approaches refers to a hierarchical dictionary containing definitions of Kanji characters. Creating such a dictionary required manual preparation of rules that could contain errors, and it was a time-consuming task. Further work on its construction could include automatic generation of rules that would be helpful in adding new definitions to the dictionary. Current research assumes that strokes in Kanji characters are drawn in the correct order, considering that the errors could occur while writing a character. Thus, new studies should include this problem to develop algorithms that are less sensitive to strokes order.

Having in mind these two problems, namely (i) the definition dictionary of Kanji characters and (ii) sensitivity of algorithms to strokes order, we propose a neural framework for online recognition of handwritten Kanji characters. More specifically, the main objectives of this study are as follows:

1) To propose a framework for online recognition of handwritten Kanji characters utilising convolutional neural networks, which are currently one of the state-of-the-art approaches in image recognition.
2) To automatically generate a dictionary containing definition of Kanji characters, which simple construction allows to easily supplement it with new definitions and its further development.
3) To assess the effectiveness of proposed solutions by experiments carried out using own implementation of the framework.

The novelty of our research is based on the implementation of a framework that allows to suggest Kanji characters in real time. The user can draw a character from the first stroke to the last and the system returns a sorted list of proposed Kanji characters after each line drawn. The results returned by the system are sorted by a measure of similarity between the character being drawn and the characters contained in the Kanji dictionary. This makes it possible to find the target character even before finishing the drawing of all the strokes.

The remainder of this paper is as follows. Section II introduces the problem to solve and presents the proposed framework for Kanji recognition. Next, Section III validates

the framework performance with illustrative examples. Finally, the findings are concluded, and references are provided.

II. KANJI RECOGNITION FRAMEWORK

In this section, we define the problem of online recognition of handwritten Kanji characters, and we propose and describe the framework that solves the indicated issue.

A. Problem definition

Let us assume that a single Kanji character, K consists of n strokes:

$$K = \{s_1, ..., s_i, ..., s_n\}. \quad (1)$$

Information about the number of strokes which compose a character is not sufficient to recognise the Kanji correctly. For a given stroke s_i, it is required to specify its type s_t, order s_o in which it appears in the sign K, and its position s_p in relation to other strokes:

$$s_i = (s_t, s_o, s_p) \quad (2)$$

The examples illustrating the described problem are the signs 犬 (dog) and 太 (fat) . Each of them consists of four lines. There are many more signs that are built by using the same number of strokes. Additional information about the strokes type is also not sufficient to indicate correctly the character. Although they have the same number of dashes, and they are of the same type written in the same order, the last short slashing line is in a different position in each of these Kanjis. Therefore, information about the placement of each stroke in a character is an issue that must be considered.

The task is to recognise the character, K during writing its strokes s_i, $i = 1, ..., n$, including their parameters s_t, s_o, s_p.

B. Framework overview

The simplified construction of our framework that tries to solve the above task is presented in Figure 1.

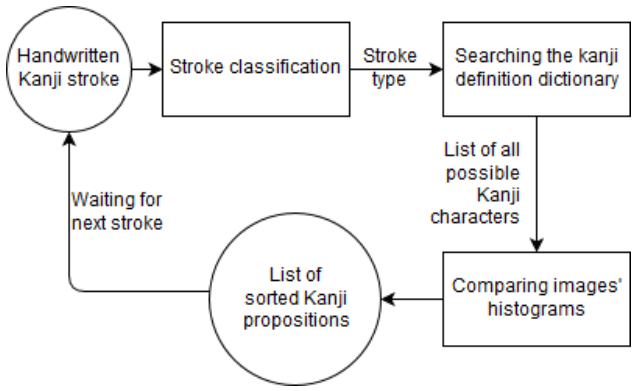


Figure 1. Overview of the framework for online recognition handwritten Kanji characters.

In the first step, after receiving the first handwritten stroke, the system saves it to a PNG file and tries to classify it to a proper type. The classification of the stroke type is done by a convolutional neural network. Having classified the stroke type, the system searches the Kanji definitions dictionary. This dictionary covers definitions of all Kanjis from our dataset. The result of that is a list of all Kanjis that start with the previously classified type. Since the list of such possible characters is extensive, in the next step, the system tries to sort them by the most likely ones. For this purpose, we compare the histograms of Kanjis' images returned by the dictionary with the image containing the handwritten strokes. Based on these comparisons, the list of Kanjis is sorted by the most similar to the drawn character. When the next stroke appears, the whole process is repeated from the beginning as described above by keeping the order of drawing lines. Finally, a sorted list of the proposed Kanji characters is returned.

Detailed implementation of individual modules are presented in the following subchapters.

C. Classification method

We utilise a Convolutional Neural Network (CNN) to classify Kanji strokes. The CNN contains several layers processing signals feed-forward. In the input layer ($INPUT$), neurons are arranged in three dimensions (width, height, depth) to process pictures. They are transferred through the set of hidden layers which are of three types, namely: (i) a convolutional layer ($CONV$), (ii) a pooling layer ($POOL$), and (iii) a fully-connected layer (FC) with an ReLU activation function. These layers produce class scores z. The architecture of the networks can be simplified as follows:

$$\begin{array}{c} INPUT \rightarrow (CONV * N \rightarrow POOL) * M \rightarrow \\ FC * K \rightarrow softmax \end{array}, \quad (3)$$

where the asterisk, $*$ indicates repetition N, M, K times respectively. In the final layer, $softmax$ the vector of class scores, z returned from the last fully-connected layer are integrated using the following softmax function:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}. \quad (4)$$

D. Kanji dictionary

The Kanji dictionary is constructed from the definitions of 7,365 Kanjis appearing in the dataset of files describing the characters. We extract all strokes from each file that maintain information about the appearance order of strokes. The dictionary is based on a tree structure, in which the number of trees is determined by the number of stroke types. The first stroke in the character determines the selection of trees to which its definition will be saved. Each subsequent stroke forms the next node in this tree. Only unique stroke types can exist at the same depth of the tree. An example of Kanji definitions is presented in Figure 2.

E. Histogram comparison

The list of suggested Kanjis returned by the dictionary may be extensive. Moreover, they are not sorted according to the similarity to the character under examination. To solve this problem, we decided to use the comparison of histograms as the similarity measure of images.

A histogram is the graphical representation of the tonal distribution in a digital image. To compare two histograms (H_1,H_2), we choose Chi-Square as a distance measure $d(H_1, H_2)$ which evaluates how well both histograms match:

$$d(H_1, H_2) = \sum_I \frac{(H_1(I)) - H_2(I))^2}{H_1(I)} \quad (5)$$

The smaller the distance between histograms is, the more similar they are. By calculating the distance between the image containing drawn lines and the images representing Kanjis returned by the list, we can sort the Kanji list according to their similarity to the tested image.

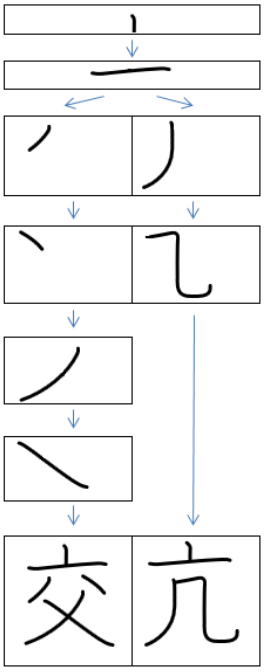


Figure 2. An example of Kanji definitions stored in the dictionary. The straight vertical stroke at the top is the root of a tree. Each subsequent stroke is added to the dictionary as a new node. If a stroke duplicates on a given depth of the tree, it is not added again to the tree.

III. Experiments

In this section, we (i) characterise the dataset for experiments, (ii) evaluate various configurations of the CNN network classifying the stroke types, and (iii) test online recognition of Kanjis written out by two groups of testers.

A. Dataset

The experiments are based on the KanjiVG database, which contains descriptions of Kanji, hiragana, and katakana. In this study, we use only Kanji, namely the version, kanjivg-r20160426 containing 7,365 characters. Each of them is described as an SVG file, which is a universal format of two-dimensional vector graphics. Essential information in this database is the order of strokes from which a particular Kanji character is constructed. Each stroke type is labelled by an identifier. Figure 3 shows 26 stroke types available the database, and samples size for each type. Unfortunately, the distribution of strokes over their types is unequal.

B. Parameters of the Convolutional Neural Network

To select an optimal classifier of Kanji strokes, we decided to test three configurations of a CNN. They are based on the architectures proposed by [7], [8] and our pre-experiments. Table I depicts the architectures in detail, namely CNN1, CNN2, and CNN3. Each of them contains several convolutional layers followed by an activation layer and a max-pooling layer. Next, there are up to three fully-connected layers in which the final result is calculated by a softmax classifier.

Table I
CONFIGURATIONS OF CNN WHICH WHERE EVALUATED. EACH COLUMN CORRESPONDS TO A LAYER OF THE NETWORK, E.G. CONV3-32 STANDS FOR THE CONVOLUTIONAL LAYER WITH KERNEL 3X3 AND 32 FILTERS, FC-1024 STANDS FOR THE FULLY CONNECTED LAYER WITH SIZE 1024.

CNN1	**CNN2**	**CNN3**
input (108 x 108 gray-scale image)		
conv3-32	conv3-64	conv3-64
maxpool	maxpool	maxpool
conv3-64	conv3-128	conv3-128
maxpool	maxpool	maxpool
	conv3-192	conv3-192
	maxpool	maxpool
	conv3-256	conv3-256
	maxpool	maxpool
		conv3-512
		maxpool
FC-256	FC-1024	
FC-n classes		
softmax		

An input layer holds the raw normalised pixel values of an image with the the size of $108x108$ pixels with one colour channel. All convolution layers compute by using the kernel with the size of $3x3$ with the stride equal to 1. The number of filters varies from 32 to 512. All max-pooling layers utilise kernels with the size of $2x2$ with the stride of two downsamples. The fully-connected layers are the size of 256 or 1024 outputs. However, the last fully-connected layer has the same number of outputs as the number of classes. It is followed by the softmax classifier. Each of fully-connected and convolution layers contains a ReLU non-linear layer with a dropout equal to 0.5 [9].

All the networks were trained by using the Adam optimiser with the parameters equal to $\beta_1 = 0.9$, $\beta_2 = 0.999$, ϵ= 10^{-8} and the learning rate equal to 10^{-4} [10]. The weights were initialised by the Glorot uniform initializer, which is also called the Xavier uniform initializer [11]. The size of mini-batch of 16 was performed. For training, we use 80 of all samples in dataset, 20% for validation at each epoch, and an untrained 20% of examples were used for testing. The models were implemented by using the Keres interface.

C. Stroke classification

The first set of experiments involved the selection of an optimal model for Kanji strokes classification to their types. The results of experiments are covered in Table II, which includes typical classification quality indicators such as by accuracy, precision, recall, and F1-score.

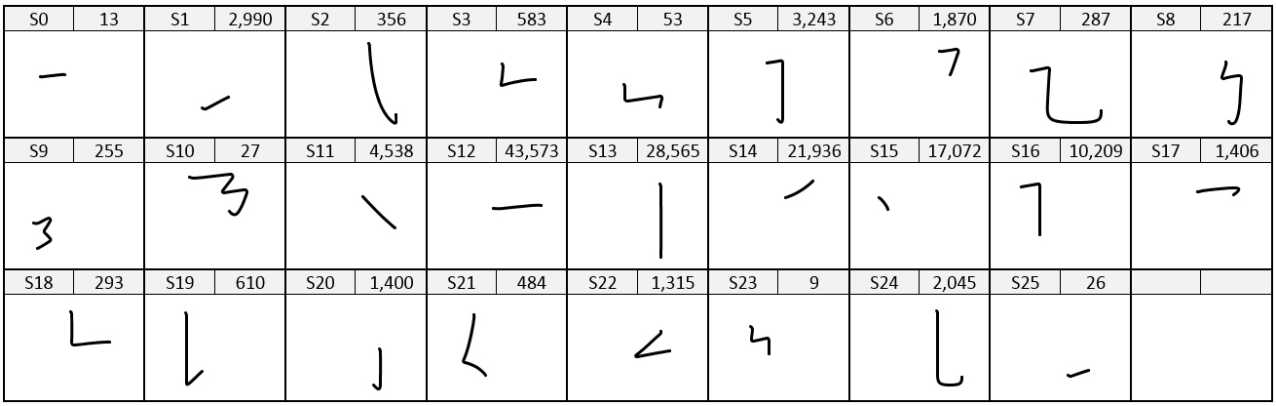


Figure 3. Kanji stroke types available in the KanjiVG database, and distribution of stroke examples over the stroke types.

We implemented and tested three network architecture configurations, namely CNN1, CNN2, and CNN3, which are described in Subsection III-B. We also used a callback function for monitoring the training process. It works as follows: if the loss function does not improve in two epochs, the process is interrupted, and the algorithm does not execute the next epoch. The results of these experiments are covered in the section, "Configuration selection" in Table II. A simple construction of the Kanji character strokes suggested starting from a simple architecture of the convolution neural network (CNN1). However, we decided to increase the number of network parameters by adding more layers, in order to see if this would improve the classification results. Subsequent attempts to deepen the network (CNN2, CNN3) did not bring any improvement. Moreover, the additional complexity of the network configuration adversely affected the network learning time. Comparing the obtained results, we decided to use the CNN1 network in the succeeding experiments.

Initially, the only element of data preprocessing was normalisation to the range of (0,1). Additionally, we utilised a ZCA (Zero Components Analysis) whitening transformation of the input images. It is a linear algebra operation that reduces redundancies in the matrix of pixel images. The operation is intended to better highlight structures and features in images for the learning algorithm. However, in our case, the improvements were insignificant (see the results in the section, "Additional ZCA whitening transformation" in Table II). Thus, we decided to skip this technique not to increase the complexity of the algorithm.

Due to the relatively small number of samples in the case of some types of strokes, e.g. S0, S4, S10, S23, we tried to extend the dataset by more samples. Unfortunately, the image augmentation techniques may not produce the expected results. This is due to the characteristic construction of Kanji characters, in which even a slight rotation of the image could change a stroke type. A good example highlighting this problem is the case of S11 and S12. Turning the stroke S11 in too high angle to the left can cause it to become too similar to the stroke S12. A similar case appears in the pair of strokes S0 and S1. These concerns have been confirmed in the research. Setting the rotation range parameter to 20 degrees degraded the results compared to the original dataset (see the results in the section, "Sensitivity to rotation" in Table II).

Looking at the shape of 26 classes, we noticed that several pairs of strokes are very similar in regard to each other. We combined the most similar stroke types by reducing the number of classes from 26 to 12. Table III presents the stroke types that we decided to merge into new classes. This approach allowed to achieve the best results in comparison to all previous experiments (see the results in the section, "Classes reduction" in Table II).

Table II
RESULTS OF THE EXPERIMENTS INVOLVING THE SELECTION OF AN OPTIMAL MODEL FOR KANJI STROKES CLASSIFICATION TO THEIR TYPES.

Architecture	Image augmentation	Accuracy	Precision	Recall	F1-score
Configuration selection					
CNN1	- normalization	96,14	96,14	96,14	96,11
CNN2	- normalization	95,79	96,23	95,25	95.72
CNN3	- normalization	94,91	95,34	94,55	94.93
Additional ZCA whitening transformation					
CNN1	- normalization - ZCA whitening	95,97	95,96	95,98	95.93
Sensitivity to rotation					
CNN1	- normalization - range rotation	93,24	94,00	92,31	93.12
Classes reduction					
CNN1	- normalization - 12 classes	**96,89**	**96,92**	**96,89**	**96,89**

Table III
LISTING OF THE NEW MERGED CLASSES.

S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S0 S12 S17	S1 S14 S25	S2	S3 S18	S4 S8 S23	S5 S6 S16	S7	S9 S10	S11 S15	S13 S19 S20	S21 S22	S24

D. Kanji recognition

The second set of experiments involved Kanji characters recognition in real-time during their writing by an individual. In order to perform these tests, we have created a simple web interface. It contains a space, in which a user can draw Kanji by using a mouse cursor. The real-time tests involved two groups of individuals, namely (i) ten people who have never studied Japanese before, (ii) one expert familiar with Japanese

language and Japanese calligraphy. Each tester had to draw 100 randomly selected Kanji characters. Prior to that, each non-expert individual had performed a small tutorial, which instructed how to write ten Kanjis.

During the tests, especially with non-expert group, we encountered two problems. The first issue resulted from unfamiliarity of Kanji structure by testers. Some individuals were not able to notice differences between some strokes, because they were so similar in regard to the others. The second issue was of technical nature. Our tests were carried out on a laptop with a mouse device. Testers complained about the shape of the mouse device, because it was too small and uncomfortable to use. It affected the quality of drawn lines. Users could not draw in the way that they wanted to. We were suggested that it would be more convenient to do tests on a tablet with a pen device.

It has to be noted that we assumed that a target Kanji was correctly identified if the following conditions had been met:

1) The target character had to appear in the first five characters proposed by the system.
2) A tester had only two attempts to draw the mark; any subsequent attempts were not taken into account.

In spite of the problems mentioned above, we managed to obtain the following results. 79% of the drawn Kanji characters by the non-experts were correctly recognised by the system. For the expert's drawings, the system recognised correctly 89% of characters. The detailed results showing the number of correctly recognised Kanjis are in Table IV.

Table IV
THE NUMBER OF CORRECTLY RECOGNISED KANJIS IN RESPECT TO A POSITION ON WHICH THEY WERE PROPOSED BY THE SYSTEM (IT PROPOSES FIVE THE MOST LIKELY KANJI CHARACTERS).

Group of testers	Position					Sum
	1st	2nd	3rd	4th	5th	
Non-experts	43	18	7	5	6	79%
Expert	78	9	1	1	0	89%

In both groups, the correctly recognised character was, in most cases, in the first position of the suggested Kanjis. In only one case for the non-expert group, the correctly identified Kanji was farther than the fifth position. In other cases, the type of stroke was incorrectly classified so that the Kanji could not be recognised. The most frequent mistakes occurred between lines S0 and S1 and S9 and S1. An imprecisely drawn line caused an incorrect classification. Due to this, searching of a dictionary with the definitions of Kanji did not bring the expected result. Another encountered problem, which also affected the results, consisted of incorrectly marked strokes in the KanjiVG database. We had found that some of the characters in our sets for testers had a stroke marked as S8, when it should be marked as S1.

IV. CONCLUSIONS

In the study, we presented and implemented a complete framework that recognises Kanji characters in a real-time, based on successively drawn strokes. In comparison to the previous works, our solution did not assume writing characters in cursive. For this reason, the comparison of prior results with those achieved by us is not entirely reliable. However, we were able to find out that even the characters written by a group that has never before had contact with the Japanese writing system can be correctly recognised by the system. Another advantage of our solution is the uncomplicated construction of the system, as well as an automatically generated dictionary definition of Kanji, which can be easily expanded with new meanings. The problems encountered during the tests included the incorrect classification of strokes, which in the training set were present in a small amount. In order to improve the quality of CNN classification, it would be necessary to expand the collection in the least numerous classes. One of the limitations mentioned in earlier studies was the problem of writing strokes in the wrong order. This will be the subject of our further research.

REFERENCES

[1] T. Morohashi, *Dai Kan-Wa jiten*. Tokyo : Taishukan Shoten, Showa 59-61, 1986.
[2] M. Nakai, N. Akira, H. Shimodaira, and S. Sagayama, "Substroke approach to hmm-based on-line kanji handwriting recognition," in *Proceedings of Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 491–495. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2001.953838
[3] M. Nakai, T. Sudo, H. Shimodaira, and S. Sagayama, "Pen pressure features for writer-independent on-line handwriting recognition based on substroke hmm," in *Object recognition supported by user interaction for service robots*, vol. 3, 2002, pp. 220–223. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2002.1047834
[4] J. Tokuno, M. Nakai, H. Shimodaira, S. Sagayama, and M. Nakagawa, "On-line Handwritten Character Recognition Selectively Employing Hierarchical Spatial Relationships among Subpatterns," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, G. Lorette, Ed., Université de Rennes 1. La Baule (France): Suvisoft, Oct. 2006, http://www.suvisoft.com. [Online]. Available: https://hal.inria.fr/inria-00104751
[5] I. Ota, R. Yamamoto, S. Sako, and S. Sagayama, "Online handwritten kanji recognition based on inter-stroke grammar," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, Sept 2007, pp. 1188–1192. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2007.4377103
[6] M. Nakagawa, J. Tokuno, B. Zhu, M. Onuma, H. Oda, and A. Kitadai, "Recent results of online japanese handwriting recognition and its applications," in *Arabic and Chinese Handwriting Recognition*, D. Doermann and S. Jaeger, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 170–195.
[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556
[8] C. Tsai, "Recognizing handwritten japanese characters using deep convolutional neural networks," 2015.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://dx.doi.org/10.1145/3065386
[10] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980
[11] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: http://dx.doi.org/10.1.1.207.2059

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 485–488
DOI: 10.15439/2018F186
ISSN 2300-5963 ACSIS, Vol. 15

Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus

Nina Khairova*, Svitlana Petrasova*, Włodzimierz Lewoniewski†, Orken Mamyrbayev‡ and Kuralai Mukhsina§
*National Technical University "Kharkiv Polytechnic Institute",
Kyrpychova str., 61002, Kharkiv, Ukraine
Email: khairova@kpi.kharkov.ua, svetapetrasova@gmail.com
†Poznań University of Economics and Business,
Al. Niepodległści 10, 61-875 Poznań
Email: wlodzimierz.lewoniewski@ue.poznan.pl
‡Institute of Information and Computational Technologies,
125, Pushkin str., 050010, Almaty, Republic of Kazakhstan
Email:morkenj@mail.ru
§Al-Farabi Kazakh National University, Kazakhstan,
71 al-Farabi Ave., Almaty, Republic of Kazakhstan,
Email: kuka_aimail.ru

***Abstract*—Automatic extraction of synonymous collocation pairs from text corpora is a challenging task of NLP. In order to search collocations of similar meaning in English texts, we use logical-algebraic equations. These equations combine grammatical and semantic characteristics of words of substantive, attributive and verbal collocations types. With Stanford POS tagger and Stanford Universal Dependencies parser, we identify the grammatical characteristics of words. We exploit WordNet synsets to pick synonymous words of collocations. The potential synonymous word combinations found are checked for compliance with grammatical and semantic characteristics of the proposed logical-linguistic equations. Our dataset includes more than half a million Wikipedia articles from a few portals. The experiment shows that the more frequent synonymous collocations occur in texts, the more related topics of the texts might be. The precision of synonymous collocations search in our experiment has achieved the results close to other studies like ours.**

I. Introduction

OVER the last few years, there has been an upsurge of interest in the research which focuses on ways to the retrieval and identification of semantic similarity for textual elements of various levels (words, collocations, and short text fragments). One of the main reasons for this is the expansion of the boundaries of the use of semantically similar texts fragments in various natural language processing applications. Nowadays, words similarity can be processed in Information retrieval systems, Question answering systems, Natural language generation systems, Plagiarism detection systems, Automatic essay grading systems and some others. The second reason for the growth of interest in the identification of a semantic similar element in texts is that on social media billions of small text messages are made public every day, each of which is comprised of approximately thirty words. Whereas very popular traditional algorithms, such as, for example Tf-idf used to compare texts, often fail in very short texts [1]. For this reason, sometimes semantic algorithms and techniques are more needed than statistical ones.

Now there exist enough studies concerning the problems related to the search of words with similar meaning. We could divide all the existing approaches into two groups. The first group of studies is based on the relations of the concepts in a thesaurus. The second group of methods for computing word similarity is based on the appliance of distributional models of meaning.

Measuring the semantic similarity between sentences or collocations is a more challenging task than searching words with similar meaning. Since the task of deciding whether two sentences or two collocations express a similar or identical meaning requires a deep understanding of the meaning of the text fragment. Increasingly, this task is being integrated into the common challenges of the paraphrases [2].

II. Related work

The most explored level of text similarity for the different languages is the level of words. There are a lot of different approaches and methods of computing words similarity. Some of them use thesaurus relations of hyponyms or hypernyms to compute word similarity; the others use distributional similarity of words in a corpus.

However, automatic synonymous collocation pairs extraction from corpora is the more challenging task of NLP. As the task involves two simultaneous operations. The first operation is the collocations extraction from a corpus and the second one is the acquisition of their synonymous pairs.

Wu and Zhou [3] suggested a method that firstly gets candidates of synonymous collocation pairs based on a monolingual corpus and then selects the appropriate pairs of the candidates using their translations in a second language. Pasca and Dienes

This work was supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan

[4] offered to utilize the alignment of two sentences fragments in order to retrieve small phrases with the same meaning. Barzilay and McKeown [5] like [3] built upon the methodology developed in Machine Translation. They presented an unsupervised learning algorithm for identification of similar phrases from a corpus of multiple English translations of the same source text.

Increasingly, the task of the synonymous collocation pairs extraction is being integrated into the common challenge of the paraphrases, which is interpreted as the search of the various textual realizations of the same meaning. Typically, n-gram models [2], annotated corpora and bilingual parallel corpora [6], [7] are used for paraphrases in such studies. Han et al. [8] and Kenter [9] are some of the most recent studies that concern determining the semantic similarity between short fragments of texts. Han et al. [8] combined lexical similarity features, Latent Semantic Analysis (LSA) similarity using WordNet knowledge, alignment algorithm and support vector regression model and n-gram models in order to establish the semantic text similarity. Kenter performed semantic matching between words in two short texts and used the matched terms to create a saliency-weighted semantic network [9].

In our study, we propose using logical equations in order to search collocations of similar meaning in English texts. These equations are based on conjunctions of morphological and semantic characteristics of the words that constitute the collocations. In order to correctly identify the grammatical characteristics, we exploit Stanford POS tagger and Stanford Universal Dependencies (UD) parser[1]. Additionally, in order to pick synonymous words which constitute the collocation we use WordNet synsets[2].

In order to evaluate our approach, we use Wikipedia articles from a few projects. Traditionally, articles of Wikipedia cover various subjects. However, depending on a topic and language versions, Wikipedia community has different numbers of experienced authors or experts [10]. Such groups of users often work together within some subject area of Wikipedia project. Articles related to the projects can have a specific writing style and quality standards, which are defined by the user community of these projects. Therefore, we can expect there is a lot of synonyms and synonymous collocations in texts related to similar topic.

III. Logical-Linguistic Model

According to previous studies [11], [12], the proposed logical and linguistic model formalizes semantically similar elements of a text by means of grammatical and semantic characteristics of words in collocations.

The semantic-grammatical characteristics determine the role of words in substantive, attributive and verbal collocations. Defining a set of grammatical and semantic characteristics of collocation words, we use two subject variables a^i and c^i. In substantive, attributive and verbal collocations, a set of possible semantic and grammatical characteristics for the main collocation word is defined by the predicate *P(x)*, for the dependent collocation word it is defined by the predicate *P(y)*.

The two-place predicate *P(x,y)* describes a binary relation which is a subset of the Cartesian product of $P(x) \wedge P(y)$ and so determines a correlation of semantic and grammatical information of collocation words *x* and *y*:

$$\begin{aligned} P(x,y) = {} & y^{NObjAtt} x^{NSubAg} \vee (x^{NObjOfAg} \vee \\ & x^{NObjOfAtt} \vee x^{NObjOfPac} \vee x^{NObjOfAdr} \vee \\ & x^{NObjOfIns} \vee x^{NObjOfM}) y^{NObjAtt} \vee \\ & x^{VTr} y^{NObjPac} \vee y^{AAtt} (x^{NSubAg} \vee \\ & x^{NObjAtt} \vee x^{NObjPac} \vee x^{NObjAdr} \vee \\ & x^{NObjIns} \vee x^{NObjM}) \vee x^{NSubAg} y^{APr} \end{aligned} \tag{1}$$

Using the algebra of finite predicates, we define the value of the predicate of semantic equivalence for three main types of collocations:

$$\begin{aligned} & \gamma(x_1, y_1, x_2, y_2) = (x_1^{NSubOfAg} \vee x_1^{NSubAg}) \wedge \\ & y_1^{NObjAtt} (x_2^{NSubOfAg} \vee x_2^{NSubAg}) y_2^{NObjAtt} \vee \\ & x_1^{VTr} y_1^{NObjPac} x_2^{VTr} y_2^{NObjPac} \vee x_1^{NSubAg} \wedge \\ & (y_1^{AAtt} \vee y_1^{APr}) x_2^{NSubAg} (y_2^{AAtt} \vee y_2^{APr}) \end{aligned} \tag{2}$$

IV. The Stages of Our Methodology

In order to show the correctness of our synonymous collocations extraction model we have used methodology that comprises a few steps. Fig. 1 shows the structural scheme of the methodology, which includes POS-tagging phase, Stanford UD parser and exploitation of the lexical database WordNet.

In the first phase, we employ POS-tagging and UD parser to define the grammatical and semantic characteristics of words in sentences.

The main reason to use UD parser is that its treebanks is centrally organized around notions of subject, object, clausal complement, noun determiner, noun modifier, etc. [13]. Therefore the syntactic relations which connect words of a sentence to each other can express some semantic content.

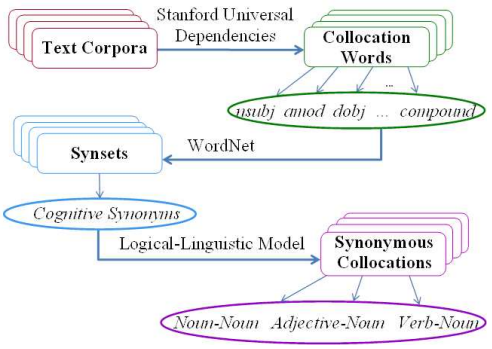


Fig. 1. The structural scheme of our experiment methodology

[1]http://universaldependencies.org/

[2]http://www.nltk.org/_modules/nltk/stem/wordnet.html

We took six types of syntactic relations tags *(compound, nmod, nmod:possobj, obj (dobj), amod* and *nsubj*) from a fixed inventory of UD grammatical relations to denote directed relations between two nouns, a verb and a noun, a noun and an adjective.

Grammatical and semantic characteristics realized through the syntactic relations tags correspond to the variables value in the two-place predicate of equation (2).

Next phase, we use WordNet in order to obtain synonyms of words connected with these types of the syntactical relations. For each collocation (substantive, attributive and verbal), synonyms are searched in WordNet synsets.

If a synonymous word is found, conformity of grammatical and semantic characteristics of a collocation and a potential synonymous word combination is being checked using the proposed logical-linguistic model.

Table I shows the examples of identified synonymous collocations in *Art* and *Biography* Wikipedia portals.

V. Source data and experimental results

Our dataset includes more than half a million articles (502 274) from Wikipedia belonged four thematic projects related to two portals. We focused on projects and portals of Wikipedia because they constitute a huge corpus of texts, which are combined by a common subject, and, at the same time, these texts are written by various authors and, consequently, may contain a lot of different synonyms.

In our studies, we choose two of the biggest portals: *Art* and *Biography*. Each portal can consist of different Wikiprojects. For our experiments, we choose four Wikiprojects (two projects from each selected Wikipedia portals)[3].

In order to estimate our synonymous collocations extraction model, we focus on three approaches. In the first approach, we identify synonymous collocations in any Wikiproject. In the second approach, we identify synonymous collocations in two different projects of the same portal. In the third of our experiments, we identify synonymous collocations in two different projects of two different portals. Our hypothesis is that two projects of the same portal may have the higher number of synonymous collocations than two projects that belong to different portals. The hypothesis is based on an idea that synonyms are occurring more often in related topics texts.

[3]List of the articles of each Wikiproject was extracted in April 2018 https://tools.wmflabs.org/enwp10/cgi-bin/list2.fcgi

Based on the Corpus Linguistics approaches [14], in order to have the opportunity to compare the synonyms occurrence frequency in the Wikiprojects of different sizes, we normalized the frequencies per ten thousand words. Additionally, we devoted attention to synonymous collocations distribution by three types.

Tables II - IV show relative frequencies of synonymous collocations that occur in four different Wikiproject, two different projects of the same portal and two different projects of two different portals, respectively.

TABLE II
Relative frequencies of synonymous collocations that occur in the Wikiproject

Wikiproject	The relative frequency of synonymous collocations		
	Substantive	Attributive	Verbal
Album	214.4	144.8	12.9
Film	277.5	281.7	10.9
Politics and government	200.9	175.4	3.5
Science and academia	280.2	210.4	210.4 4.6

TABLE III
Relative frequencies of synonymous collocations that occur in two different projects of two different portals

Wiki-projects / portal	The relative frequency of synonymous collocations		
	Substantive	Attributive	Verbal
Album – Film / Art	199.3	166.3	5.9
Politics and government - Science and academia /Biography	200.8	162.5	3.9

The results of Tables II - III show that the number of synonymous collocations in articles belonging to one Wikiproject is more than the number of synonymous collocations in articles belonging to two different Wikiprojects.

The results of Tables III - IV show that the number of synonymous collocations in articles belonging to the one portal is more than the number of synonymous collocations in articles belonging to two different portals. The articles of one project are closer to one subject than the articles of two different projects.

However, Wikiprojects can also have similar fields of knowledge. Due to it, articles from different projects of the same portal might have enough synonymous collocations.

TABLE I
The Examples of synonymous collocations extracted from Art and Biography portals

Collocations	Syntactic relation tags	Synonymous collocations	Syntactic relation tags	Collocation types
history of land	nmod:of	nation's story	nmod:poss	Substantive
soul power	compound	ability of person	nmod:of	Substantive
spectacular progression	amod	outstanding advance	amod	Attributive
restoration is incompetent	nsubj:cop	restitution is incapable	nsubj:cop	Attributive
qualify place	dobj	modify position	dobj	Verbal
preserve fire	dobj	maintain flame	dobj	Verbal

TABLE IV
RELATIVE FREQUENCIES OF SYNONYMOUS COLLOCATIONS THAT OCCUR IN TWO DIFFERENT PROJECTS OF THE SAME PORTAL

Wiki-projects (portal)	The relative frequency of synonymous collocations		
	Substantive	Attributive	Verbal
Album (Art)–Politics and government (Biography)	128.7	117.2	2.8
Album (Art)–Science and academia (Biography)	172.5	144.4	3.8

VI. EVALUATION RESULTS AND CONCLUSIONS

In our experiments, we use precision to assess the validity of our approach. The main reason why we could not evaluate recall of experiment results that we did not have a training corpus with correctly identified synonymous collocations.

In order to obtain the number of correctly found semantically similar collocations, we use an expert opinion. About 1000 synonymous pairs of collocations were randomly extracted from each list of three types of collocations and presented for judgment. The purpose of the evaluation was to obtain judgments on how synonymous collocations found in the texts were similar in meaning. The experts were asked to compare the similarity of meaning of the collocation pairs on the scale of from 0 to 2. The experts needed to assess the pair of collocations as 2 if these collocations had not any semantic similarity, as 1 if the pair of collocations had some semantic similarity and as 0 if they obviously found it difficult to answer.

Table V shows the values of the average precision of our approach calculated for three types of collocations.

TABLE V
THE CALCULATION OF AVERAGE PRECISION OF OUR APPROACH FOR THREE TYPES OF COLLOCATIONS

Type of collocations	Average precision
Substantive	0.781
Attributive	0.644
Verbal	0.627

Such precision is close to results of the other studies [4], [5]. In our opinion, the reason why the data show relatively low results is mistakes of the POS tagging and UD-parser.

In the future research we intend to broaden the scope of the study on semantic equivalence. In particular, there is a need for calculating the recall of our experiment and extending the approach to some other languages. Multilingualism of Wikipedia on the one hand, and the independence of each language version of this encyclopedia on the other, give the opportunity to create models that can help to identify content with the highest quality [15]. Therefore, presented approach can be used to define new metrics for the tasks of the quality texts assessment.

ACKNOWLEDGMENT

This research is supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (project No. AP05131073 – Methods, models of retrieval and analyses of criminal contained information in semi-structured and unstructured textual arrays).

REFERENCES

[1] C. De Boom, S. V.Canneyt, S. Bohez, T. Demeester, B. Dhoedt, "Learning Semantic Similarity for Very Short Texts," *Pattern Recognition Letters,* vol. 80, 2016, pp. 150–156. DOI: 10.1109/ICDMW.2015.86

[2] J. Ganitkevitch, B. V. Durme, C. Callison-Burch, "PPDB: The paraphrase database," *in Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2013, pp. 758–764.

[3] H. Wu, M. Zhou, "Synonymous Collocation Extraction Using Translation Information," *in Proc. of the 41st Annu. Meeting on Association for Computational Linguistics,* Stroudsburg, PA, USA, vol. 1, 2003, pp. 120–127. DOI: 10.3115/1075096.1075112

[4] M. Pasca, P. Dienes, "Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web," *in Proc. of the Second Int. Joint Conf.: Natural Language Processing,* Korea, 2005, pp. 119–130. DOI: 10.1007/11562214_11

[5] R. Barzilay, Kathleen R. McKeown, "Extracting Paraphrases from a Parallel Corpus," *in Proc. of the 39th Annu.Meeting on Association for Computational Linguistics,* Stroudsburg, PA, USA, 2001, pp. 50–57. DOI: 10.3115/1073012.1073020

[6] J. Ganitkevitch, C. Callison-Burch, C. Napoles, B. V. Durme, "Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation," *in Proc. of the Conf. on Empirical Methods in Natural Language Processing,* 2011, pp. 1168–1179.

[7] B. Dolan, C. Quirk, C. Brockett, "Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources," *in Proc. of the 20th Int. Conf. on Computational Linguistics,* Geneva, Switzerland, 2004. DOI: 10.3115/1220355.1220406

[8] L. Han, A. Kashyap, T. Finin, J. Mayfield, J. Weese, "UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems", *in Proc. of the Second Joint Conf. on Lexical and Computational Semantics,* vol. 1, 2013, pp. 44–52.

[9] T. Kenter, M. de Rijke, "Short Text Similarity with Word Embeddings, "*in Proc. of the 24th ACM Int. Conf. on Information and Knowledge Management,* 2015, pp. 1411–1420. DOI: 10.1145/2806416.2806475

[10] W. Lewoniewski, K. Węcel, W. Abramowicz, "Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles," *Informatics,* 2017. DOI: 10.3390/informatics4040043

[11] S. Petrasova, N. Khairova, "Automatic Identification of Collocation Similarity," *in Proc. of 10th Inter. Scientific and Technical Conf.: Computer Science & Information Technologies,* Lviv, 2015, pp. 136–138. DOI: 10.1109/STC-CSIT.2015.7325451

[12] S. Petrasova, N. Khairova, "Using a Technology for Identification of Semantically Connected Text Elements to Determine a Common Information Space," *Cybernetics and Systems Analysis,* Springer, vol. 53 (1), 2017, pp. 115–124. DOI: 10.1007/s10559-017-9912-z

[13] Joakim Nivre Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, et al., "Universal Dependencies v1: A Multilingual Treebank Collection," *in Proc. of the Tenth Int. Conf. on Language Resources and Evaluation,* Paris, France, 2016

[14] T. McEnery, A. Hardie, "Corpus Linguistics: Method, Theory and Practice," Cambridge University Press, 2012.

[15] Lewoniewski W. "Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis," *Lecture Notes in Business Information Processing,* vol 303. Springer, Cham, 2017, pp 216–227. DOI: 10.1007/978-3-319-69023-0_19

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 489–495
DOI: 10.15439/2018F110
ISSN 2300-5963 ACSIS, Vol. 15

Evaluating Combinations of Classification Algorithms and Paragraph Vectors for News Article Classification

Johannes Lindén, Stefan Forsström, Tingting Zhang
Departement of Information Systems and Technology
Mid Sweden University
851 70 Sundsvall, Sweden
Email: johannes.linden@miun.se, stefan.forsstrom@miun.se, tingting.zhang@miun.se

***Abstract*—News companies have a need to automate and make the process of writing about popular and new events more effective. Current technologies involve robotic programs that fill in values in templates and website listeners that notify editors when changes are made so that the editor can read up on the source change on the actual website. Editors can provide news faster and better if directly provided with abstracts of the external sources and categorical meta-data that supports what the text is about. In this article, the focus is on the importance of evaluating critical parameter modifications of the four classification algorithms Decisiontree, Randomforest, Multi Layer perceptron and Long-Short-Term-Memory in a combination with the paragraph vector algorithms Distributed Memory and Distributed Bag of Words, with an aim to categorise news articles. The result shows that Decisiontree and Multi Layer perceptron are stable within a short interval, while Randomforest is more dependent on the parameters best split and number of trees. The most accurate model is Long-Short-Term-Memory model that achieves an accuracy of 71%.**

I. Introuction

THERE are several approaches to extracting the key points of non-formatted text to be able to retell the most important information to the reader. A common problem is the over all descriptive word of the text, such as this text is about Kultural arts. In this article we will call this information a category. Other problems involve retrieving shorter summaries of text documents and computing other meta data describing the text content. The purpose is to make the text more available to readers/writers, and from there link the text the appropriate audience by for example personalization and document search algorithms.

Swedish journalists categorize their news articles manually. It is a time-consuming task and yields inconsistent results. A previous study by Oscar Hjelmstedt and Mats Sellfors shows that journalists needs to take advantage of algorithms that can manage news content to get a better understanding of how they work and move on from old habits of news paper press that are very different from digital media [1]. In the future, news will to a greater extent be written by using deep learning algorithms to write news faster and at a lower cost [1]. It is therefore important that the journalists have an understanding and know how to work with the new working conditions. Hence, this research seeks to answer the following research questions:

1) Will a combination of classification and paragraph vector algorithms improve the results of the categorizations?
2) To which extent and which combinations of classification and paragraph vector algorithms shows the best accuracy for new articles?

In this article, the focus will be the paragraph vectors distributed memory and distributed bag of words as described by Mikolov et al. [2]. As an additional layer of algorithms we categorize the paragraph vector using the standard data mining algorithms: decision tree, random forest and multi layer perceptron. A comparison between the result of our work and other categorization algorithms like Fasttext and Lai Siwei et al.'s classification algorithm will be presented and the f-score metric will be evaluated [3], [4].

A. Outline

This article will first go through some related work that already have been done in the research field. Secondly a approach/models description about the algorithms used and how they are combined in the experiments conducted described in the next section. The scores that are based on the model evaluation experiments are then presented in the result section, followed by the discussion and conclusions of the project. Finally suggestions for future work are presented.

II. Related Work

In the field of text categorization, there is already existing research that should be considered. Facebook announced their own categorization algorithm called fasttext a few years ago, which shows good performance in speed [3]. In a matter of seconds, a trained fasttext model is ready to categorize texts in comparison to other algorithms like Gensim with the same dataset setup this is fast, there by the name. The reason for the speed increase is most likely their n-gram implementation that mainly introduces good results for the syntactical parts of the text, but weakens the semantic parts. The fasttext algorithm gets a way with less computational complexities and still performing well on syntactic problems [4]. In this article

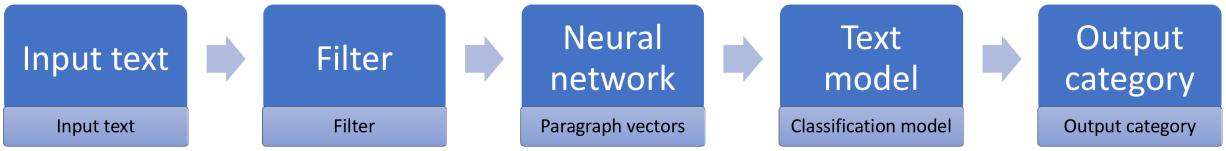


Fig. 1. An overview of the approach and model

several steps to categorize the text are performed. An LSTM classification algorithm using a multi-timescale approach to prolong the long-term memory in the network is proposed by Liu et al [5]. Liu uses English data sources with question-answer and binary data.

Standford University conducts research in the field of natural language processing, which includes text categorization of Twitter feeds. They consider a large number of Twitter posts and categorizes the posts into positive and negative. They also consider emoticons noisy labels. Although they are showing good results they categorize the posts of two categories and emoticons are quite unstable facit. The Standford work is in line with this article in that it too uses predefined categories for training, but positive and negative categories are more abstract which implies human error is greater. [6]

A completely different approach to text categorization is term association, associates certain rules and patterns of the text with certain categories. This approach relies on the pruning phase of the model, since a lot of fake rules will emerge from a large dataset. Therefore Maria-Luiza Antonie and Osmar R. Zaïane [7] came up with three pruning rules that reduce the amount of rules and increases the accuracy of the model in 2002. The paper shows that an association model for categorization can be both accurate and fast.

The algorithms used when dealing with natural language processing are commonly also used in image processing. For example a paper about practical study of network image based classification by Dabrowski, Marek et al uses a convolutional network to categorize images which can be compared to a very deep convolutional network approach to character based language classification by Conneau Alexis et al [8], [9]. These algorithms depending on the dataset takes often long time to converege given the initial weights the time of the result could vary between different training runs Polap, Dawid et al shows one method to use multi-threaded learning with a multi-core solution to achieve faster training time [10].

Google released a data source platform called GDELT that stores a lot of news metadata from all over the world. The system has the computer power to store and monitor world news on the internet from certain news sources, new events as well as events reaching as far back in time as 1979. Over 200 million events are recorded from over 240 countries and available for live requests. In 2013, a comparison between the GDELT and ICEWS was made that compared the popularity and scale of the two data sources. [11], [12]

Other competitive algorithms that provide a document vector for a given text are LDA algorithms and text ranking algorithms. In an article by Thanda et al. [13] they compare the different algorithms in a systematic matter to find relations between math queries.

III. Approach and Model

We propose a four step model that predicts categories of arbitrary text paragraphs. See Figure 1 for an overview of our implementation. The input in Figure 1 is the algorithm parameters θ and a single document D, which is interpreted as a sequence of words $w_1, w_2, \cdots, w_n$. The output of the model is a set of category probabilities $c_1, c_2, \cdots, c_i, \cdots, c_m$ where

$$c_i = P\left(\text{ith category}|D, \theta\right) \approx P\left(\text{ith category}|D\right) \quad (1)$$

. Before the actual training, the data is filtered from text paragraphs that only consists of a link to another article and that does not represent any categorical value. The combination of step three and four is the machine learning part, which will answer the research questions. The algorithms used are described in the following sections.

A. Input Text

The input of the proposed algorithm is an unstructured sequence of words forming a text paragraph. The text should be in Swedish and can be of any length. Although in this article the tested the text sizes have a length between 5 to 600 words.

B. Input Filter

Before the text can serve as the input to the model the text needs to be filtered to remove special characters. Exclamation marks and question marks are replaced by full stops. Commas, references and document links are removed. The purpose of the filter layer is to make the paragraph uniform, so that the model can be processed with as few exceptions as possible. In this step, one scenario was to filter on verbs and nouns to make the input data more precise to the point and thus describe the category using narrow information without noise words. To filter on these words a part-of-speech tagger was used.

Part-of-speech taggers (POS-tagger) are used to extract the sentence structure in the form of a dependency tree and the cooresponding word's tags [14]. A tag indicates if the word is a verb, noun, preposition or any other type. The dependency tree has a root word node and child words that directly relates to the parent word, an example is shown in Algorithm III-B. Google released a POS-tagger called SyntaxNet with state-of-the-art performance, and one year later announced an

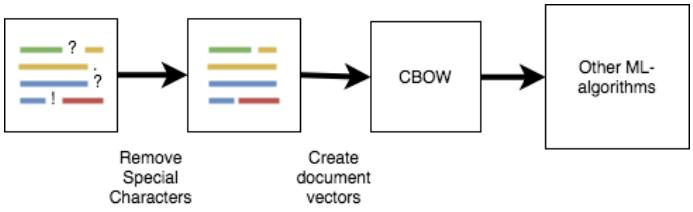


Fig. 2. The input vector of the algorithms: decision tree, random forest and multi layer perception is generated as shown above. It applies the filter layer conditions, and by using a CBOW algorithm directly on the document it produces the document vectors that can be categorized with the classification algorithms.

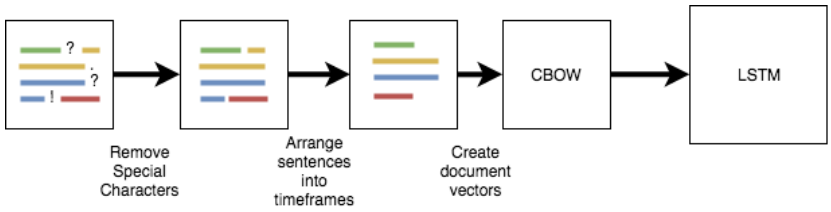


Fig. 3. The input vector of the LSTM algorithm is shown above. It first applies the filter layer conditions, then divides the document into sentences to be used as input data for the CBOW algorithm that produces the document vectors.

improved version [15], [16]. In the experiments in this article SyntaxNet is used with a Swedish training set (also called a treebank). Out of the resources mentioned in Nilson et al, we selected a treebank made by Jan Einarsson's project, which is well documented [17], [18], [19]. In the experiments we will try to select only nouns and verbs to predict a category.

Algorithm 1 A part-of-speech example sentence parsed by SyntaxNet.

Input sentence: I found a website to post AI tutorials .
Parsed dependency tree:

```
found VBD ROOT
+– I PRP nsubj
+– website NN dobj
|   +– a DT det
|   +– post VB infmod
|       +– to TO aux
|       +– tutorials NNS dobj
|           +– AI NNP nn
+– . . punct
```

C. Paragraph Vectors

The third step in Figure 1 is a neural network model that is constructed and trained to predict paragraph vectors when given the text form in the input or filter step. The paragraph vectors are unique vectors that describe the relation between the words in the document and a likely word to appear with them [2]. Computing the cosine similarity between two paragraph vectors yields a positive value when the documents are sharing similar contexts, a value close to zero when no relation could be found, and a negative number when a relation with opposite meaning [2]. With this knowledge, it is common to carry out paragraph operations such as you could for word vectors, for example Equation 2 [2].

$$king - man + woman = queen \tag{2}$$

The paragraph vectors do have context awareness, and are therefore believed to contain information about what makes a document category. The paragraph vectors are computed using the PV-DM algorithm which is an extension of the known word2vec algorithm bag of words (WV-BOW) [20].

The PV-DM algorithm tries to map all word vectors in a paragraph to a unique vector. The unique vector and the word vectors are averaged into the hidden layer h in our implementation. The rest of PV-DM algorithm follows the continuous bag of words (CBOW) algorithm [4], [21]. The unique paragraph vector can be considered an additional word in the context of a CBOW network. The idea of this extra vector is to have a form of memory about the topic of the paragraph, which explains the name PV-DM. The training of a PV-DM uses stochastic gradient descent [22] and neural network back-propagation by calculating the derivate of the vector from the next layer and applying it to compute the previous layer vector.

In our experiments, the PV-DBOW paragraph algorithm is implemented by the distributed memory vector concatenated with the distributed bag of words vector described by Mikolov et al [2].

D. Text Model

When a paragraph representation has been established it is time to go to step four: the categorization step. Therefore, we continue with the assumption that the paragraph vectors are properly and uniquely defined with good paragraph relationships in the previous step. The categorization algorithms we propose in these experiments are decision trees, random forest, multi layer perception and long-short term memory (LSTM).

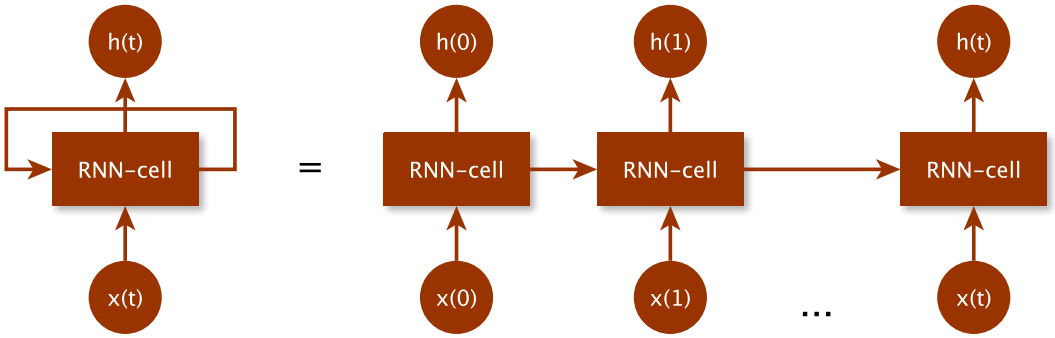


Fig. 4. A recurrent neural network cell. The cell to the left is the general notation of the hidden layer in an RNN model. The right unfolded version is the representative of the RNN with a short-term length of t.

Each algorithm has its own parameters that will be evaluated in the experiments. The input of the categorization step is the paragraph representation of the text. The output is the category belonging probabilities for each considered category described in Equation 1.

For the machine learning algorithms decision tree, random forest, multi layer perception are using the document vector from doc2vec CBOW algorithm was used directly as input, as shown in Figure 2. For the LSTM model, the time parameter was used to separate the paragraph into sentences, and then calling the CBOW algorithm in each sentence input into each LSTM time-slot as shown in Figure 3.

LSTM networks are based on recurrent neural network techniques. A recurrent neural network (RNN) is constructed like a neural network with an input, hidden layers and a output layer. The size of the input and output layer depends on the objective of training. The RNN-cell can be visualised as shown in Figure 4. The activation function of an RNN is usually the $tanh$ function. For each iteration, the model is trained by backpropagation through the network. The purpose of RNN is to have a short-term memory that remembers previous neurons. One of the first and simple constructions of RNN is the recurrent neural network language Model (RNN-LM). The hidden layer of an RNN-LM algorithm remembers the neurons one time-step back in the training history. [23]

Today, there are different variations of RNNs depending on the goal of the model. Andrej Karpathy summarises the different networks that are used into different mappings. One-to-one mapping is the original algorithm, for example the RNN-LM algorithm. One-to-many mapping when there is one input and several RNNs connecting to several outputs. This mapping can, for example be, used for image prediction with one image and several words that predict the image. Many-to-one mapping is where there are several inputs mapping to one output, this mapping can for example be used for classification. Many-to-many mapping is what Karpathy describes as two different mappings: one mapping that maps to an equal number of input and output RNNs (N-N), and another mapping that maps to a different number of inputs and outputs (N-M). The N-N mapping can, for example, be used to predict video sequences over time, while N-M mapping can be used for translation problems. [24]

LSTM networks are a special case of RNN that solve a fatal problem in the original RNN. The long-term problem that LSTM solves is introduced in RNN where the gradient descent exponentially diverges to infinity or converges to zero. On an ordinary RNN, the most simple solution that is used is to clamp the value between zero to one, but that still leaves the convergence to zero problem. The way LSTM networks work is to introduce three sigmoid layers and certain gates that only let parts of the information through to compensate for the vanishing gradient. The first sigmoid layer determines what information that is important from the previous LSTM-cell, the second sigmoid layer determines what information is important from the $tanh$ layer in the current cell and the third sigmoid layer determines what information will be passed to the next LSTM-cell. The gates that open or close based on the input from the previous LSTM-cell either remove or add information to a cell state that is also passed through to the next LSTM-cell. The third sigmoid layer extracts a piece of information from the cell state to the output value. [25], [26]

E. Output Category

As mentioned in the beginning of this chapter, the output layer interprets the output of the classification model and determines the number of categories, the probability of the prediction, and finally returns the category probabilities of the given text paragraph. The output category probabilities are normalized such that the highest probability of a category is one and the lowest one is zero according to Equation 3.

$$\frac{value - min(value)}{max(value) - min(value)} \tag{3}$$

IV. Evaluation and Experiments

This section presents the experiments conducted in the evaluation of the model. The dataset that we will train the models on consists of Swedish texts from the MittMedia article

TABLE I
THE PARAMETERS USED TO TRAIN THE CLASSIFIER MODELS. THERE ARE 2848 TRAINED MODELS IN TOTAL, E.G. EACH COMBINATION OF THE PARAMETERS FOR EACH ALGORITHM.

Algorithm	Parameter	Values
MLP	Dimensionality of the feature vectors	100x1, 100x2, 100x3, 100x4, 100x5 and 100x6
	Activation function	Identity, Logistic sigmoid, tanh and relu
	Solver function	LBFGS, SGD and Adam optimizer
	L2 penalty	0.005, 0.010, 0.015, 0.020
Decisiontree	Criterion	Gini and Entropy
	Max features	20%, 40%, 60%, 80% and 100% of the training data
	Max depth	10, 20, 30 and 40
	Minimum sample split	2, 4, 6 and 8
	Minimum leaf samples	2 and 4
Randomforest	Criterion	Gini and Entropy
	Max depth	10, 20, 30 and 40
	Minimum sample split	2, 4, 6 and 8
	Minimum leaf samples	2 and 4
	Number of trees	5, 10, 15, 20 and 25
	Features count for best split	2, 4, 8, 10, auto, $\sqrt{\text{number of features}}$, $\log_2$ (number of features)
LSTM	No of hidden layers	2, 3, 4
	LSTM neurons	10, 32, 50
	LSTM Timesteps	10, 20, 40
	Filter	Stop words, non-nouns and non-verbs
	Training epochs	2, 5, 15

database, including metadata. The metadata for categorization contains tags and categories that are attached to each training instance. Each instance can have more than one category. When the article was written, the categories were attached manually by the editors. During the experiments 5 to 30 categories were used to train, test and validate the models. The implementation of the experiments where made in Python. A tensorflow model was constructed for each model [27].

The data-instances are not always valid, therefore a prepossessing step is necessary to filter outlier texts. The dataset was divided into three groups to train, test and validate. First, one large filtered set was fetched from the database. The training and testing groups were separated into 60% training (5597 articles) and 40% testing (8396 articles) data after filtering of invalid outliers. Next a new non-seen filtered data-set was fetched and used for the validation group.

A. Experiment Settings

The following pre-processing was done before starting the actual training. The dataset used was filtered due to some odd outliers. The outliers are the result of different guidelines from the company Mittmedia, at different times, such as links to other articles or dynamically loading content. The restrictions were removed by ignoring the instance body content shorter than 10 words. If the instance body content contained more than 10 words, there were still outliers and unusable document-instances. The unusable instances sometimes contained JavaScript code that loaded contents from another URI onto the page dynamically when loading the page. Since these instances from the database are usually displayed in a web browser, this was not a problem. However, when the instances were directly fetched to the algorithm, the JavaScript content had to be removed.

In the experiments, the parameters of the classification algorithms consisted of all combinations of values for each algorithm as shown in Table I. The document count and categories were also changed independently of the algorithm parameters to evaluate impact on the result. For a full report on the results for the document and categories variation we recommend that you read the full report [20]. The F_1-score measurement were used to compare, validate and test the models. The measurement was developed in 1992 and gives an objective result of the harmonic mean between the precision and recall with equal weights [28].

The categories used for the experiment are labelled in Swedish: Blåljus, Ekonomi, Kultur, Nöje, Släkt och familj and Sport. The categories could be roughly translated as Accidents, Economy, Culture, Entertainment, Family and Sport, respectively. Accidents are texts about car chases, fires, injuries and so on. Economy is about financial issues such as business deals, the stock market and so on. Culture is mostly about art, museum or movie premiers, the nobelprice and so on. Entertainment is similar to Culture, as also this category potentially could include movie reviews, popular events, and other fun activities in the society. It is a fine line what would be defined as Culture and what is defined as Entertainment and different editors could have slightly overlapping definitions. Family is about newborns, the royal family, or family activities. The Sport category covers all kinds of sports, such as tennis, hockey, horse riding and so on. A majority of news

TABLE II
CONFUSION MATRIX OF THE CBOW AND LSTM COMBINED PREDICTIONS OF THE NEWS ARTICLE DATASET

Real \ Predicted	Accidents	Economy	Culture	Entertainment	Family	Sport	
Accidents	536	35	3	2	18	6	600
Economy	1	465	39	4	38	2	600
Culture	50	30	426	35	106	2	600
Entertainment	15	33	200	272	62	18	600
Family	14	25	112	29	413	7	600
Sport	14	17	10	25	81	453	600
	632	605	790	367	718	488	3600

TABLE III
CONFUSION MATRIX OF THE SINGLE LSTM-NETWORK PREDICTIONS OF THE NEWS ARTICLE DATASET

Real \ Predicted	Accidents	Economy	Culture	Entertainment	Family	Sport	
Accidents	287	51	16	27	46	173	600
Economy	154	287	27	34	57	41	600
Culture	32	100	262	13	121	72	600
Entertainment	92	128	106	139	58	77	600
Family	85	70	62	41	242	100	600
Sport	148	20	13	50	90	279	600
	798	656	486	304	614	742	3600

are written for the Sport category. Therefore it is important that we consider equal amount of text documents for each category so that the model isn't biased to, for example the Sport category.

TABLE IV
ACCURACY OF THE EVALUATED MODELS

Algorithm	Test Score	Validation Score
LSTM	0.74	0.71
LSTM (without CBOW)	0.42	0.37
MLP	0.31	0.14
Decision Tree	0.10	0.05
Random forest	0.08	0.03

The best model will be selected and evaluated without the CBOW vectors but, instead, word identifiers are used to verify that the combination is better than the algorithm alone. For objective fairness, the settings of the additional evaluated model will be the same as the best performing model.

B. Results

The test and validation measurements of each model are presented in Table IV. Only the best performing models are selected and presented in Table IV for each algorithm. The neural network is consistently performing well with a test F-score of about 0.31 and validation score of 0.22. The decision-tree classifier performs with a validation score of 0.05. The F-score of randomforest validation is 0.03. The LSTM network is currently superior to the other algorithms with a test score of 0.74 and validation score of 0.71. The confusion matrix of the combined CBOW and LSTM model show that the categories Culture and Entertainment are frequently mixed up by the algorithm, it is where the majority of miss-predictions occur, see Table II. In Table III the LSTM is compared with indicies as input, which shows that there is a larger uncertainty in this data.

Since the LSTM model performed best out of the the selected models, the additional model was trained using only the LSTM model with the same settings and unique IDs as input data. The additional model was performing with a validation score of 0.37 and thus we can confirm the research question that will investigate the combination of paragraph vectors and classification algorithm.

By running the model with different initial conditions we can evaluate the be model that had highest score value with a statistical approach. This way we check the reliability of the model in case it will be retrained at some point.

C. Discussion

The confusion matrix in Table III indicates that articles within one category are potentially difficult to distinguish from another category's texts. For example, the categories Entertainment, Family and Culture have some prediction overlaps. Most predictions are correct for all categories, which means that there are at least a few articles that characterize each category. Comparing LSTM with CBOW and the network without CBOW yields that the combination has significantly better performance. The reason is likely to have something to do with the vocabulary size, which is many times larger than the dataset that we are using, and thus not all words are present in the training data. This means that it is more difficult for LSTM without CBOW to predict correct categories.

Filtering away all words except nouns and verbs performed poorly compared to using all words in the document as input. The reason could be because the document vector also captures some information about how the text is structured and how the words are used in conjunction with each other. For example, which words tend to be used together, thus more frequent the word in a certain category the easier it is to predict. The time and network sizes did not significantly change the outcome, the score was slightly better using a larger value with any or both parameters. MLP is a good candidate if speed is a concern, although not near as good as fasttexts' performance. Random forest is slightly better than the decisiontree algorithm, but in general they perform similarly.

V. Conclusion

In this article we proposed a combination of classification algorithms and paragraph vector algorithms to improve the results of categorization problems. We aimed to find out if the combination of classification and paragraph vector algorithms improves the categorization, which we found to be true. We also investigated how the algorithm performed on news articles and to what extent it can be used. From the trained categorization models a probability score can be estimated for each available category that can be predicted. Based on the probabilities, a number of categories can be suggested to the editors in the system that, for example, has a probability higher than a certain threshold. The LSTM model is perfoming best in combination with the word vectors when predicting the categories. Based on the confusion matrix we can see that it is not overfitted. It can be concluded that a combination of LSTM and CBOW (classification and paragraph vector) algorithms perform better together (score of 0.71) than using only a classification algorithm such as LSTM (score of 0.37). Although the combination is not enough for the other evaluated algorithms: decisiontree, random forest and MLP with the combined CBOW algorithm achieve better result than a LSTM network with word IDs as input.

Future work for this project could be to extend the domain to other domains outside of the journalists that has a certain way of writing texts, such as common word choices and spelling standards. Although the proposed model should work in any other domain, further exploration has to be made to confirm. A recommended next step is to compare the model with the results of Liu et al, to make this possible we need to look into what data that model is evaluated on and see if we can apply the CBOW combined LSTM model using that data instead.

References

[1] O. Hjelmstedt and M. Sellfors, "Robotjournalistikens nya utmaningar," 2017.

[2] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *arXiv preprint arXiv:1607.01759*, 2016.

[4] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2267–2273, 2015.

[5] P. Liu, X. Qiu, X. Chen, S. Wu, and X. Huang, "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2326–2335.

[6] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.

[7] M.-L. Antonie and O. R. Zaiane, "Text document categorization by term association," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, pp. 19–26.

[8] M. Dabrowski, J. Gromada, and T. Michalik, "A practical study of neural network-based image classification model trained with transfer learning method." in *FedCSIS Position Papers, DOI: http://dx.doi.org/10.15439/2016F211*, 2016, pp. 49–56.

[9] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for natural language processing," *arXiv preprint*, 2016.

[10] D. Połap, M. Woźniak, W. Wei, and R. Damaševičius, "Multi-threaded learning control mechanism for neural networks," *Future Generation Computer Systems, DOI: https://doi.org/10.1016/j.future.2018.04.050*, vol. 87, pp. 16–34, 2018.

[11] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012," in *ISA Annual Convention*, vol. 2. Citeseer, 2013.

[12] M. D. Ward, A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford, "Comparing GDELT and ICEWS event data," *Analysis*, vol. 21, pp. 267–297, 2013.

[13] A. Thanda, A. Agarwal, K. Singla, A. Prakash, and A. Gupta, "A Document Retrieval System for Math Queries," pp. 346–353, 2016.

[14] A. Voutilainen, "Part-of-speech tagging," *The Oxford handbook of computational linguistics*, pp. 219–232, 2003.

[15] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins, "Globally normalized transition-based neural networks," *arXiv preprint arXiv:1603.06042*, 2016.

[16] C. Alberti, D. Andor, I. Bogatyy, M. Collins, D. Gillick, L. Kong, T. Koo, J. Ma, M. Omernick, S. Petrov *et al.*, "Syntaxnet models for the conll 2017 shared task," *arXiv preprint arXiv:1703.04929*, 2017.

[17] J. Nilsson and J. Hall, *Reconstruction of the Swedish Treebank Talbanken.* Matematiska och systemtekniska institutionen, 2005.

[18] J. Einarsson, "Projektet talbanken. i: C platzack (utg), svenskans beskrivning 8, s76-96," 1974.

[19] ——, "Talbankens talspråkskonkordans," 1976.

[20] J. Lindén, "Understand and Utilise Unformatted Text Documents by Natural Language Processing algorithm," vol. 46, no. 0, 2017.

[21] X. Rong, "word2vec parameter learning explained," *CoRR*, vol. abs/1411.2738, 2014. [Online]. Available: http://arxiv.org/abs/1411.2738

[22] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2011, pp. 693–701.

[23] T. Mikolov, S. Kombrink, L. Burget, J. Černocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[24] A. Karpathy, "The unreasonable effectiveness of recurrent neural networks," *Andrej Karpathy blog*, 2015.

[25] C. Olah, "Understanding lstm networks," *GITHUB blog, posted on August*, vol. 27, p. 2015, 2015.

[26] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM Neural Networks for Language Modeling." in *Interspeech*, 2012, pp. 194–197.

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[28] N. Chinchor and B. Sundheim, "Muc-5 evaluation metrics," in *Proceedings of the 5th conference on Message understanding*. Association for Computational Linguistics, 1993, pp. 69–78.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 497–500
DOI: 10.15439/2018F13
ISSN 2300-5963 ACSIS, Vol. 15

Voice control in mixed reality

Dawid Połap
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
Email: Dawid.Polap@polsl.pl

***Abstract*—The gameplay in the augmented or virtual reality is based on the use of external equipment such as glasses/telephones and possibly the use of additional sensors or controllers. In both cases, the interaction involves pressing the keys on the phone or controller. An interesting aspect is the control of objects created in a virtual way using voice commands. In this paper, we propose a solution to manipulate objects in the augmented reality using player's voice. The user can move the object using pre-programmed commands. The solution is based on speech processing and artificial neural networks. The technique has been tested and the results presented and discussed.**

I. Introduction

AUGMENTED, merged or virtual realities are leading technological and research directions in the field of human machine interaction, image processing, games and many more. The rapid development of technology is visible through numerous achievements in these areas. The last examples are getting rid of mobile phones for extended goggles with motion controllers that allow not only to much more interaction, but a much longer immersion (the goggles do not overheat as fast as smartphones). Not only virtual reality, but augmented one is widely used. It is especially used in games, learning or even medicine what is visible in the effects of scientific work of scientists from around the world.

The biggest commercial achievement of these technologies was the mobile application for catching virtual monsters in 2016, i.e. *Pokemon GO*. The game based on searching and catching was based on a combination of reality (using a camera and GPS locator in smartphones) with 3D models that caused young people to leave the houses for fresh air and cross the streets. To this day, game is considered as a good change in the gaming industry when it comes to long time spent in front of the computer [1]. These types of applications not only have a good effect on health, but also make it possible to meet new people and cooperate [2]. Support for augmented reality is performed using a certain application with access to the camera. It is a kind of human–machine interaction [3], [4].

The transition from augmented to virtual reality is quite smooth. It is only necessary to turn off the camera and set up goggles to be able to move into a fully artificially created world. Studies on immersion are dispelled in increasing the feeling of detachment from the environment [5]. Enabling and capturing movements is quite a significant problem. However, it is gradually solved, as shown by the authors of [6], where the virtual keyboard was presented. Health problems such as stress or focusing too close to the display are constantly analyzed so that in the future everyone can safely and fully immerse into virtual world [7]. Nowadays, important issue is Internet of Things and its impact on our life [8], [9]. Similar, this technology and methods can be applied to different problems like some of geological [10], [11].

An important aspect is also research in the field of voice analysis and processing. In [12] deep learning was used to quickly detect the accent for English language. Again in [13], the authors focused on recognizing emotions on the basis of voice samples using classical processing and classification techniques. Moreover, the techniques of voice processing recorded in the sound file to the text version using artificial intelligence methods are presented and widely described in terms of numerous applications in practice [14].

In this paper, we focus on introducing a voice to this type of technology. Selected, recent developments in the signal processing field can be viewed in [15], [16]. My proposition is based on the analysis of the voice and its transformation into a text command that can be realized in an augmented/virtual reality.

II. Voice Controller

The voice controller operates on the basis of downloading a sound sample, its processing, classification and processing of the output decision. Described controller is based on converting a speech signal into its graphical form, cutting and classifying it, and then transferring information to an application supporting augmented or virtual reality.

A. Voice processing

The sound analysis is quite problematic for several reasons. The pronounced sound has the continuous form, by saving it in bit form, it simplifies the signal and is still incapable to analyze. Let $s(n) = (s_0, s_1, s_2, \ldots, s_{N-1})$ is a signal in discrete form. Unfortunately, this number sequence is still not possible to be analyzed. To remedy this, use a selected transformation such as Fourier defined as follows

$$S_k = \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2\pi i n k}{N}\right) \quad 0 \leq k \leq N-1, \qquad (1)$$

where $S_k \in \mathbb{C}$ is a discrete value in $(S_0, S_1, S_2, \ldots, S_{N-1})$. For the purpose of machine calculations, in practice the above

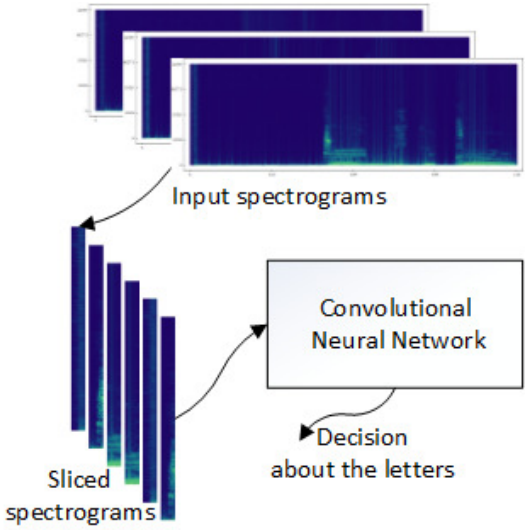


Fig. 1: Model of the proposed technique of converting the voice to the text.

equation is not used, and its recursive form called Fast Fourier Transform is used. It is defined as follows

$$\begin{aligned} S_k &= \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2i\pi nk}{N}\right) \\ &= \sum_{m=0}^{N/2-1} s_{2m} \exp\left(-\frac{2i\pi km}{N/2}\right) \\ &+ \exp\left(-\frac{2i\pi k}{N}\right) \sum_{m=0}^{N/2-1} s_{2m+1} \exp\left(-\frac{2i\pi km}{N/2}\right). \end{aligned} \tag{2}$$

It is possible to present a sound sample with the help of an image so-called spectrogram. It is a flattened three-dimensional graph that is spanned on two axes – time and frequency. The flattened dimension is the intensity, which is depicted by the shadow of a given color. The formula for that is defined as

$$spectrogram\{s(t)\}(t,f) \equiv |S(t,f)|^2, \tag{3}$$

where $S(\cdot)$ is a short-time Fourier transform understood as

$$S(m,f) = \sum_{n=-\infty}^{\infty} s[n]w[n-m]\exp(-jfn), \tag{4}$$

where $s[n]$ is a signal in discrete form, $w(\cdot)$ is a window function like sine window described as

$$w(n) = \sin\left(\frac{\pi n}{N-1}\right) \tag{5}$$

B. Convolutional Neural Network

Convolutional neural network are an example of neural structures adapted to receive graphic images in the input [17]. The idea of operation this type of classifier is based on the cells in the primary cortex. The network structure is composed of three types of layers. First type is convolution layer where some feature are extracted from input image. In practice, some filter ω is applied to the image (for instance blur or sharpening defined as a matrix of 3×3 with step size S). Output from this layer is an image called feature map. The second type of layer is pooling, which reduces the size of the incoming image by calculation the mean, maximum or minimum value in a given neighborhood area. The third type is called fully connected and the architecture of these is similar to classical neural network structure. Each pixel returned from the last layer is considered as a single neuron which form the input layer. Then, hidden and one output layers are created.

These type of structure can be trained using backforward propagation algorithm [18], [19]. Let me introduce some designation – $f(\cdot)$ as an error function, output value from neuron at position (i,j) in l layer as $\frac{\partial f}{\partial y_{ij}^l}$. The error at the end of the network is known and marked as $\frac{\partial f}{\partial y_{ij}^l}$. Algorithm is based on chain rule what is understood as sharing weight with one another can be defined as

$$\frac{\partial f}{\partial \omega_{ab}} = \sum_{i=0}^{N-m}\sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l}\frac{\partial x_{ij}^l}{\partial \omega_{ab}} = \sum_{i=0}^{N-m}\sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^l} y_{(i+1)(j+b)}^{l-1}. \tag{6}$$

Using above equation, error $\frac{\partial f}{\partial x_{ij}^l}$ can be calculated as

$$\frac{\partial f}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\frac{\partial y_{ij}^l}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\frac{\partial\left(\sigma(x_{ij}^l)\right)}{\partial x_{ij}^l} = \frac{\partial f}{\partial y_{ij}^l}\sigma'(x_{ij}^l), \tag{7}$$

where $\sigma(x)$ is activation function in classic reasoning. Having defined a formula for an error on the current layer, it is necessary to define formula for an error in earlier layers. Note that the gradient for the convolutional layer can be determine as

$$\begin{aligned} \frac{\partial f}{\partial y_{ij}^{l-1}} &= \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^l}\frac{\partial x_{(i-a)(j-b)}^l}{\partial y_{ij}^{l-1}} = \\ &\sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^l}\omega_{ab}, \end{aligned} \tag{8}$$

and this equation can be used to define error as

$$\frac{\partial x_{(i-a)(j-b)}^l}{\partial y_{ij}^{l-1}} = \omega_{ab}. \tag{9}$$

It worth to note, that algorithm does not work for pooling layer which are skipped during training process.

C. Proposed technique for object manipulation in augmented reality

The built-in microphone on smartphone or tablet can record the sound in real time. Assume that the obtained audio will be saved every 2 seconds, it is possible to represent such a sample as spectrogram that will be cut every 0.5 seconds (these values are chosen in an empirical way). As a result, four graphics will be obtained, which will be subjected to training

and classification through a convolutional neural network (classical 5x5 filters were used – Gaussian blur and emboss). A schematic model is presented in Fig. 1.

This action allows to classify the analyzed spectrogram in order to change the sound into a text form. The text is simple in order to pass the parameter to the program, and more accurately enables the voice manipulation of the object placed in the augmented reality.

TABLE I: Values of statistical coefficients.

	0.1	0.01	0.001	0.0001
Γ	0.435	0.575	0.755	0.835
Λ	0.362	0.560	0.756	0.832
Ψ	0.221	0.388	0.608	0.713
Υ	0.416	0.581	0.752	0.845
Φ	0.447	0.570	0.758	0.825

III. Experiments

In order to test the proposed speech processing techniques and object manipulation, a simple model was created and placed on the screen of smartphone. The classifier was trained with 150 samples (75 per command) which were arranged in a random manner in the ratio of 70 : 30 (learn:test samples). The classifier was trained to obtain an error of 0.1, 0.01, 0.001 and 0.0001. The correctness of classification with respect to the error has been presented in Fig. 2-5. The best results were achieved for the smallest error, screenshots from application where these proposition was implemented are shown in Fig. 6. These results allow to calculate some statistical coefficient like accuracy Γ, Dice's coefficient Λ, overlap Ψ, sensitivity Υ, specificity Φ and calculated values of these parameters are presented in Tab. I. It shows that the accuracy increases very fast, which is a good indicator. Other values also increase, such as sensitivity, which means the probability of indicating that this is a mistaken command among all erroneous commands. Again, the coefficient of specificity points to value of incorrect samples that had a negative result.

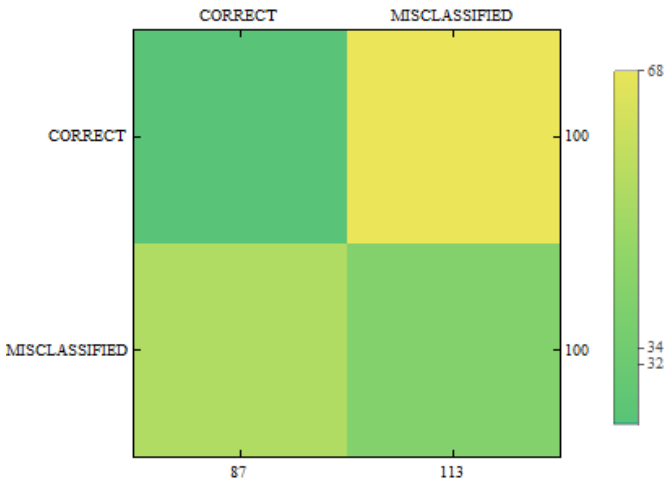


Fig. 2: Accuracy in relation to obtained error 0.1.

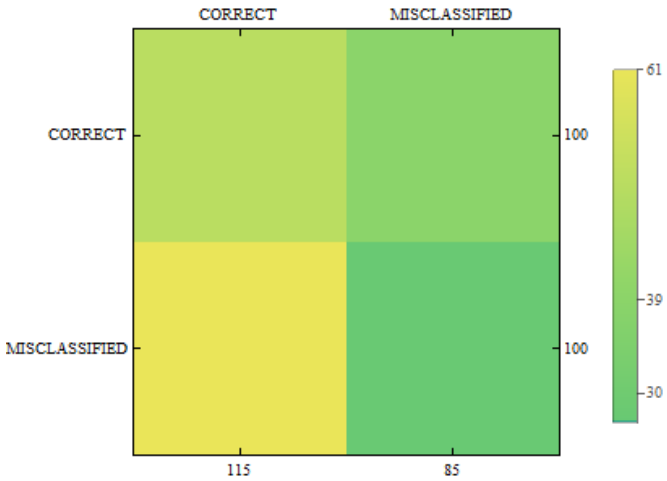


Fig. 3: Accuracy in relation to obtained error 0.01.

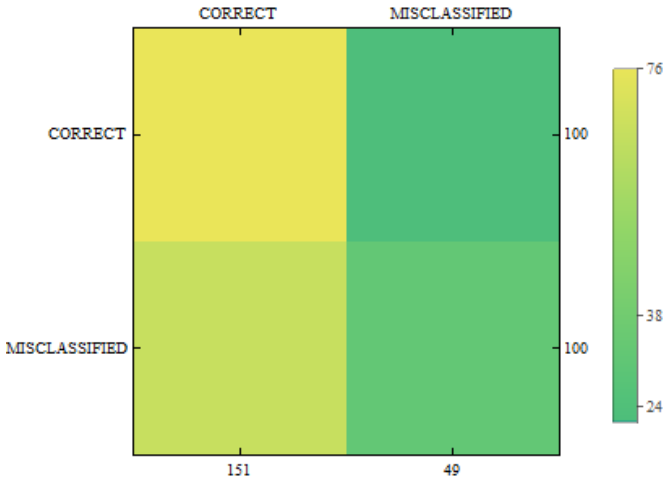


Fig. 4: Accuracy in relation to obtained error 0.001.

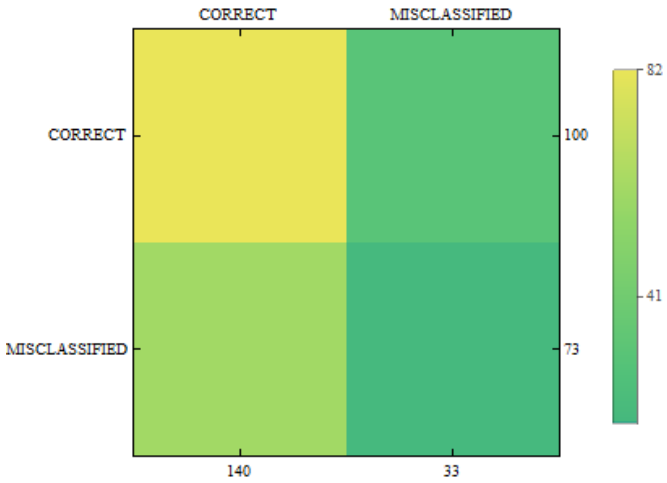


Fig. 5: Accuracy in relation to obtained error 0.0001.

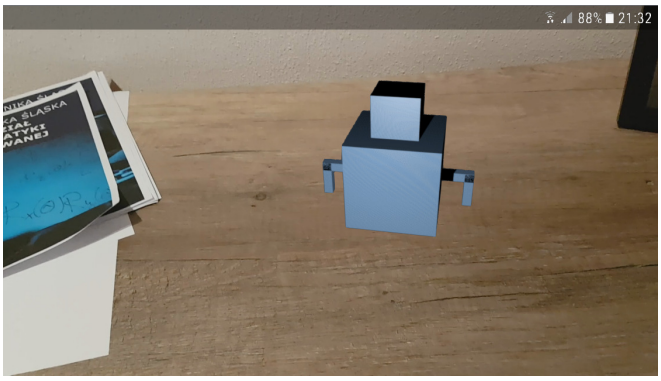
IV. Conclusion

The proposed solution can diversify, and above all, improve user interaction with artificially created objects. This type of solution is the first approach to this type of activity. The obtained results indicate a high potential, however, the proposed technique has several parameters that should have been taken into account. Particularly problematic is saving audio sample every few seconds, loading processing as well the length of sliced elements from spectrograms.

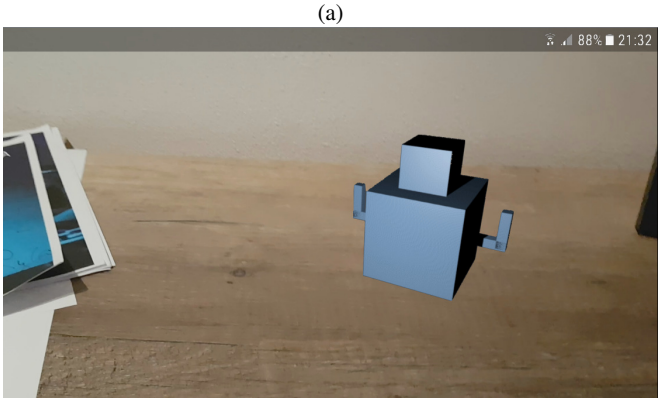
In this paper, two simple commands such us *UP* and *DOWN* were tested. The trained classifier allowed to correct manipulation of the object. The effectiveness of this model indicates the high flexibility of use in games like *Pokemon Go*, which can increase the playability and refresh the classic operation of augmented reality technology. In my future work, we plan to focus on improving these issues as well as improving the classifier's operation on longer text commands.

V. Acknowledgment

Author acknowledge contribution to this project of the "Diamond Grant 2016" No. 0080/DIA/2016/45 from the Polish Ministry of Science and Higher Education and the Rector pro-quality grant No. 09/010/RGJ18/0033 at the Silesian University of Technology, Poland.



(a)



(b)

Fig. 6: Screenshots of the application's operation – (a) object in the initial state, (b) object after executing the sound command "*UP*".

References

[1] A. G. LeBlanc and J.-P. Chaput, "Pokémon go: A game changer for the physical inactivity crisis?" *Preventive medicine*, vol. 101, pp. 235–237, 2017.

[2] B. Morschheuser, M. Riar, J. Hamari, and A. Maedche, "How games induce cooperation? a study on the relationship between game features and we-intentions in an augmented reality game," *Computers in human behavior*, vol. 77, pp. 169–183, 2017.

[3] S. Chodarev, "Development of human-friendly notation for xml-based languages," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 1565–1571.

[4] Z. Sroczyński, "Actiontracking for multi-platform mobile applications," in *Computer Science On-line Conference*. Springer, 2017, pp. 339–348.

[5] L. Wang, F. Forni, R. Ortega, Z. Liu, and H. Su, "Immersion and invariance stabilization of nonlinear systems via virtual and horizontal contraction," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 4017–4022, 2017.

[6] C.-M. Wu, C.-W. Hsu, T.-K. Lee, and S. Smith, "A virtual reality keyboard with realistic haptic feedback in a fully immersive virtual environment," *Virtual Reality*, vol. 21, no. 1, pp. 19–29, 2017.

[7] D. Cho, J. Ham, J. Oh, J. Park, S. Kim, N.-K. Lee, and B. Lee, "Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine," *Sensors*, vol. 17, no. 10, p. 2435, 2017.

[8] S. Deniziak, T. Michno, and P. Pieta, "Iot-based smart monitoring system using automatic shape identification," in *Federated Conference on Software Development and Object Technologies*. Springer, 2015, pp. 1–18.

[9] J. Protasiewicz, W. Pedrycz, M. Kozłowski, S. Dadas, T. Stanisławek, A. Kopacz, and M. Gałężewska, "A recommender system of reviewers and experts in reviewing problems," *Knowledge-Based Systems*, vol. 106, pp. 164–178, 2016.

[10] M. Sołtysiak, D. Dabrowska, K. Jałowiecki, and V. Nourani, "A multi-method approach to groundwater risk assessment: a case study of a landfill in southern poland," *Geological Quarterly*, vol. 62, no. 2, pp. 361–374, 2018.

[11] V. Nourani, G. Andalib, and D. Dabrowska, "Conjunction of wavelet transform and som-mutual information data pre-processing approach for ai-based multi-station nitrate modeling of watersheds," *Journal of hydrology*, vol. 548, pp. 170–183, 2017.

[12] K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, "Automatic lexical stress and pitch accent detection for l2 english speech using multi-distribution deep neural networks," *Speech Communication*, vol. 96, pp. 28–36, 2018.

[13] R. Maskeliunas, V. Raudonis, and R. Damasevicius, "Recognition of emotional vocalizations of canine," *Acta Acustica united with Acustica*, vol. 104, no. 2, pp. 304–314, 2018.

[14] R. Shadiev, T.-T. Wu, and Y.-M. Huang, "Enhancing learning performance, attention, and meditation using a speech-to-text recognition application: Evidence from multiple data sources," *Interactive Learning Environments*, vol. 25, no. 2, pp. 249–261, 2017.

[15] A. Venckauskas, A. Karpavicius, R. Damaševičius, R. Marcinkevičius, J. Kapočiūte-Dzikiené, and C. Napoli, "Open class authorship attribution of lithuanian internet comments using one-class classifier," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017, pp. 373–382.

[16] M. S. Elmahdy and A. A. Morsy, "Subvocal speech recognition via close-talk microphone and surface electromyogram using deep learning," in *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017, pp. 165–168.

[17] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, no. 5-6, pp. 555–559, 2003.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1717–1724.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 501–504
DOI: 10.15439/2018F321
ISSN 2300-5963 ACSIS, Vol. 15

Classification of Computer Network Users with Convolutional Neural Networks

Jakub Nowak, Marcin Korytkowski, Rafal Scherer *Member, IEEE*
Computer Vision and Data Mining Lab, Institute of Computational Intelligence, Czestochowa University of Technology,
Al. Armii Krajowej 36, 42-200 Czestochowa, Poland
Email: {jakub.nowak, marcin.korytkowski, rafal.scherer}@iisi.pcz.pl

***Abstract*—Automatic detection of abnormal behaviour of computer network users is a desirable and hard to achieve feature. We show that convolutional neural networks can classify users in local computer networks based on features of web pages which were requested by a user (e.g. URL address, URL category, the day of week or time when the web page was visited). We demonstrate our approach on data collected from a firewall over an eight-month period. This network traffic meta-data allowed to achieve satisfactory classification accuracy on unseen, future network traffic data.**

I. Introduction

FOR the past twenty years, the Internet and its utilisation have grown at an explosive rate. Moreover, for several years computer network users have been using various devices, not only personal computers. We also have to manage with many appliances being constantly online and small Internet of Things devices. Efficient computer network intrusion detection and user profiling are substantial for providing computer system security. Along with the proliferation of online devices, we witness more sophisticated security threats. It is possible to enumerate many ways to harm networks, starting from password weakness. Malicious software can be illicitly installed on devices inside the network to cause harm, steal information or to perform large tasks. Another source of weakness can be Bring Your Own Device schemes, where such devices can be infected outside the infrastructure. At last, social engineering can be used to acquire access to the corporate resources and data.

Each network user leaves traces, some of them are generated directly by the user, e.g. on social networks, others are closely related to the computer network mechanisms. Thanks to network traffic-filtering devices, network administrators nowadays have an enormous amount of data related to network traffic at their disposal. One of the ways to ensure security is to block traffic based on the categorisation of websites. Edge devices (e.g. firewalls or routers with firewall function) verify requested URLs based on the global URL databases and their category ultimately deciding whether a user can access a given page. An example of such devices and reputation databases can be PaloAlto with the Brightcloud database. In order to increase the security, high-end firewalls, simultaneously to filtering, log all the traffic passing through them, storing it, e.g. in relational databases, SYSLOG systems, etc. Thus, network operators can verify the actions of individual users. Log analysis is a crucial element of the network security diagnosis. Usually, the log content is analysed after the fact of an attack or a possible error. Registration of logs is also one of the basic requirements of the right to conduct telecommunications activities. It is mandatory for Internet providers to record who and when visited or shared network resources. Depending on the authentication methods used in a given network and the class of security devices, logs contain information from a very general level, e.g. user IP address, time of the event (of page visit), the address of the requested page up to the user's name.

In [1] the authors rely on the classification of users with all data stored in network logs. The aim was to identify users for the purposes of forensic applications. A compelling argument about why to identify users using data from network traffic and not using the IP addresses assigned to them is that people use mobile devices more often and identify less and less with one, single network. They do not limit the data, as in our case, to the URLs themselves. They use the meta-data of the traffic. However, that base only includes 46 users. The disadvantage of the system that uses all the data can be its performance. Using only URLs, we have fewer data to process, which contributes to the higher efficiency of our system. Events can also be detected by distributed MapReduce approaches [2]. The authors of [3] used a logger on each computer, which additionally logged applications, mouse movements, how were the keys pressed. The error was only 7.1 per cent for 21 users. However, the disadvantage of the method is the interference in the user's system and continuous logging of its behaviour in the system. We expect that artificial neural networks can improve the results on the problem presented in the paper [4][5][6]. A promising approach can be using space-time features [7]. A comprehensive surveys are presented in [8] and [9]. As we faced the challenge of processing a significant amount of data, it would be beneficial in the future to utilise some big data processing methods [10].

In the paper we use convolutional neural networks [11], [12] to classify computer network users based on URL requested by their devices.

II. Computer Network Data

This article is based on data collected from a WAN network infrastructure, which is used by residents of four districts in Poland, as well as network users who are employees of the local government offices and their organisational units,

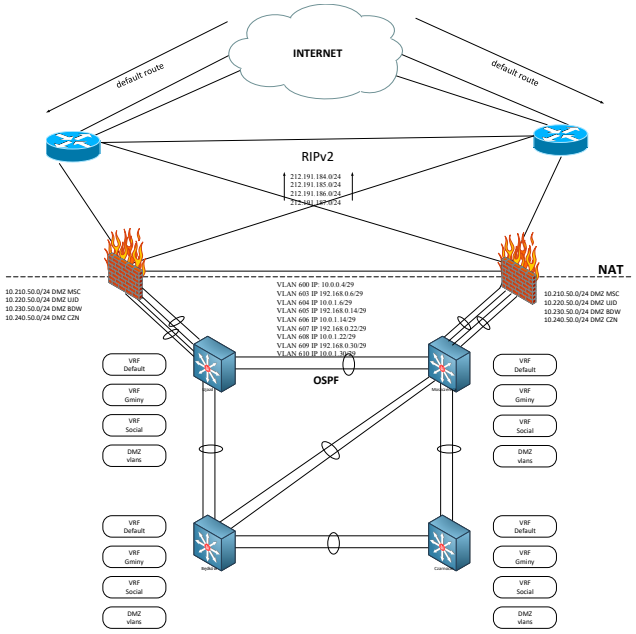


Fig. 1. Schema of the network infrastructure used in the paper to collect traffic data. The Internet is accessed by two routers, and the local network is protected by firewalls.

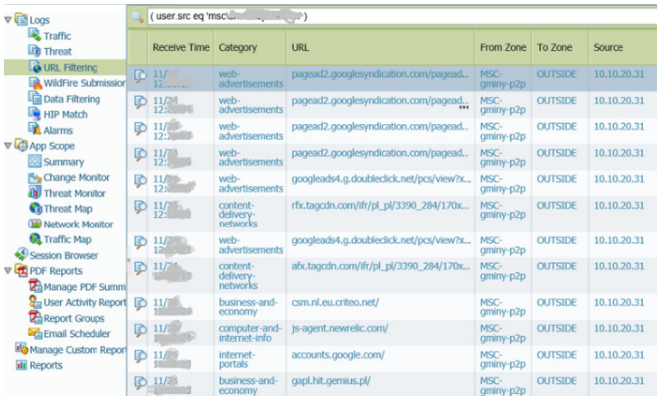


Fig. 2. Part of the network log used to create the training data.

e.g. schools, hospitals, etc. Internet access to the analysed network is done with the help of two CISCO ASR edge routers that route packets using RIP version 2. A cluster of PaloAlto devices working in an active-active mode takes care of the network security. The network is routed by the Open Shortest Path First (OSPF) algorithm with virtual routing and forwarding (VRF). In each of the four districts, there is one CISCO core switch. The network is shown in Fig. 1. In the analysed network infrastructure, users are authenticated using accounts in Active Directory services. A RADIUS service is configured for users using the wireless network. PaloAlto devices integrate with the list of accounts contained in domain controllers, thanks to which each user's network traffic is logged using its Active Directory name. The data was acquired in the form of logs from the MS SQL database, then sorted

Fig. 3. Example of a part of the input image created from a set URLs.

by user names, by the dates when they were recorded, and by the name of the URL address. Sorting by date is designed to reflect the order of websites that were visited. We do not operate on exact dates in this case. It is important to note that additional sorting by name is important. It happened that between two identical URLs registered at exactly the same time a foreign address was requested by another program of that user. If the two URLs next to each other were the same, only one was left. This treatment had a big influence on the results and was able to improve the results.

III. Experiments

A URL is an address that allows locating a website on the Internet. The user encounters it mainly when using a web browser. However, the computer network logs we work on in the paper also contain URLs that have been used by programs running in the background such as antiviruses, system updates, etc. Each user's computer uses different applications and at a different time, which allows to even better distinguish them. As we have observed in our data, certain addresses can be typed in a characteristic way for a given person, e.g. www.google.pl, google.com, www.google.com are three different URLs from our point of view but referring to the same site.

We were inspired here by Zhang et al [12] to present text data in the form of a one-hot vector at the character level. The dictionary consisted of 70 characters:

```
abcdefghijklmnopqrstuvwxvyz_0123456789
-;.!?:/\|#$%&ˆ'+=<>()[],"'|^
```

From URL sequences, we created sessions consisted of 8 to 300 URL addresses. The session in the paper is a set of user's URLs where the interval between requests does not exceed 30 minutes. We choose 300 as the maximal value because it was the longest session in the data where there was no 30-minute break.

URLs are concatenated into one string (string), with one URL maximum being 45 characters long. We did not assume a minimum URL length. There was no such need in the collected data. If one URL next to the other was exactly the same, then only one remained. In our experiments, we were only interested in transitions between addresses. The maximum pessimistic length of the training vector is therefore 300 * 45 characters = 13,500.

After creating the input sequences, because only a few vectors had a length of 13,500, we truncated all the sessions to 8014, which allowed to speed up the learning of the network.

The data were collected from June 2017 till February 2018, where the last ten days of February 2018 were used as testing data. The data we collected allowed to divide it into 36,937 sessions, with 1,684,704 for training data, and 9,208 sessions

TABLE I
TOP MOST FREQUENT IP NUMBERS IN THE TRAINING DATA

IP	Count	Domain
212.77.101.148	15918	AS12827 Wirtualna Polska S.A.
172.217.20.174	14846	AS15169 Google LLC
40.77.226.250	13959	AS8075 Microsoft Corporation
172.217.22.14	13137	AS15169 Google LLC
86.111.241.163	10511	elara.iq.pl
185.184.8.30	10118	AS60558 PHOENIX NAP
65.55.44.108	9510	AS8075 Microsoft Corporation
212.77.100.82	8569	AS12827 Wirtualna Polska S.A.
173.241.240.143	8550	AS36089 OPENX TECHNOLOGIES
127.0.0.1	15182	localhost

TABLE II
TOP MOST FREQUENT IP NUMBERS IN THE TESTING DATASET

IP	Count	Domain
217.74.66.216	10639	AS16138 INTERIA.PL Sp z.o.o.
172.217.22.14	10326	AS15169 Google LLC
212.77.101.148	7773	AS12827 Wirtualna Polska S.A.
172.217.23.174	6751	AS15169 Google LLC
40.77.226.250	6126	AS8075 Microsoft Corporation
185.14.253.220	6069	s11.smartsupp.com
185.184.8.30	5005	AS60558 PHOENIX NAP
172.217.22.110	4903	AS15169 Google LLC
172.217.22.3	4636	AS15169 Google LLC
216.58.208.46	4559	AS15169 Google LLC
86.111.241.163	4502	IQ PL Sp. z o.o.

with 788,086 URLs for testing data. The testing dataset was created from the last data in the aforementioned period, thus the testing was performed on the future, unseen URLs. Table I presents top IP numbers present in the training data.

We built convolutional networks [11], [12] and trained them with the backpropagation algorithm [13]. To improve the accuracy we added the linear embedding layer [14] that was also trained from the data. The first network (CNN1), shown in Fig. 4 obtained 33% classification accuracy. It had 32-output linear embedding layer, convolution, max pooling and finally three-layer full-connected network with 62 outputs. The second network (CNN2) is shown in Fig. 5 and obtained 27% classification accuracy. It had the same structure and the only difference was adding additional two input channels encoding time of the week (work days vs. weekends). It allowed improving the accuracy. The next experiment was performed with the same structure with increased the number of outputs in the linear embedding layer to 42 (Fig. 5). It reduced slightly the classification error to 26%. The network training is shown in Figures 7-9.

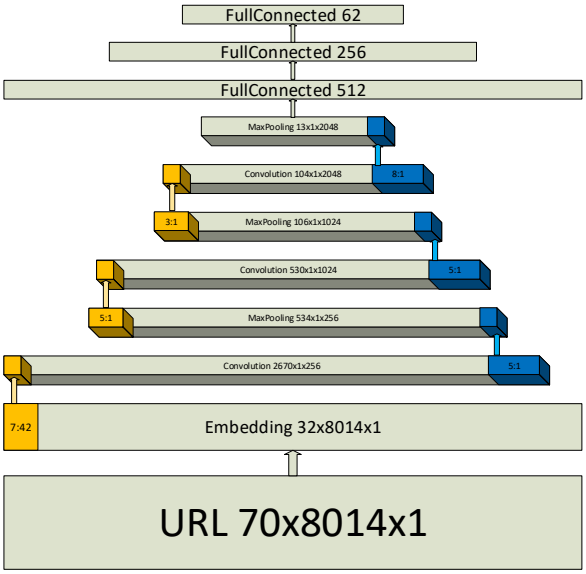


Fig. 4. Convolutional network architecture (CNN1) with 32-output embedding layer without work day information.

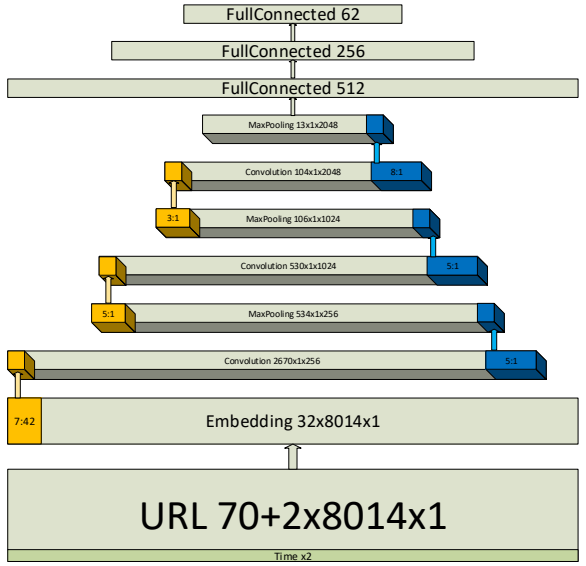


Fig. 5. Convolutional network architecture (CNN2) with 32-output embedding layer with work day information encoded in two last characters.

IV. CONCLUSION

In this article, we proposed a method to classify computer network users based on URLs they have visited. To this end, we encoded URLs as one-hot vectors and presented them as inputs to convolutional neural networks. The obtained results show that the use of additional input data channels with information on users' work days (working days or weekends) resulted in improvement of profiling quality by over 6%. Also, very good effects brought the addition of the embedding layer in the structure of the convolution network. However, increasing the size of this layer does not significantly improve the quality of classification results. The limitation of the presented method is, of course, limited possibility to accurately detect users solely basing on the requested URLs.

REFERENCES

[1] N. Clarke, F. Li, and S. Furnell, "A novel privacy preserving user identification approach for network traffic," *Computers & Security*, vol. 70, pp. 335 – 350, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404817301384

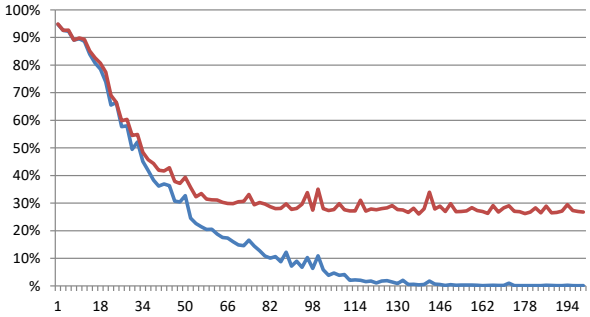


Fig. 9. Convolutional network training (blue line) and validation (red line error (y axis) through epochs (x axis) for CNN3 (Fig. 6).

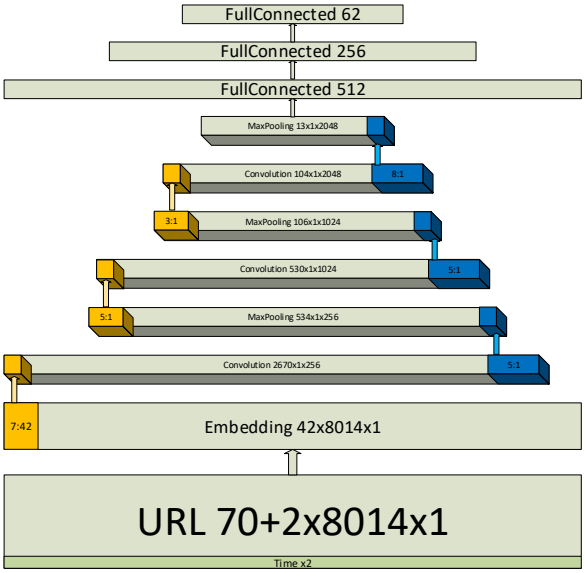


Fig. 6. Convolutional network architecture (CNN3) with 42-output embedding layer with work day information encoded in two last characters.

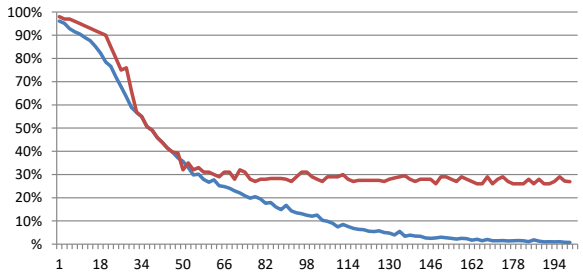


Fig. 7. Convolutional network training (blue line) and validation (red line error (y axis) through epochs (x axis) for CNN1 (Fig. 4).

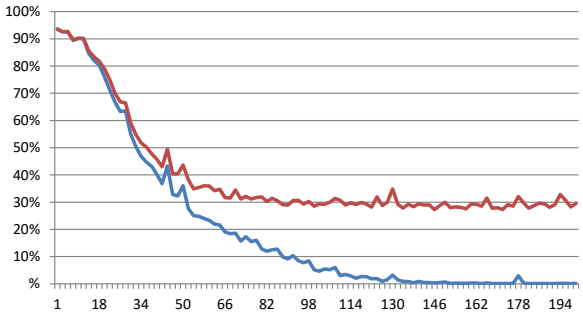


Fig. 8. Convolutional network training (blue line) and validation (red line) error (y axis) through epochs (x axis) for CNN2 (Fig. 5).

[2] P. Yan, "Mapreduce and semantics enabled event detection using social media," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 3, pp. 201–213, 2017.

[3] A. Aupy and N. Clarke, "User authentication by service utilisation profiling," in *Proceedings of the ISOneWorld 2005, Las Vegas, USA*, 2005.

[4] G. Bologna and Y. Hayashi, "Characterization of symbolic rules embedded in deep dimlp networks: a challenge to transparency of deep learning," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 4, pp. 265–286, 2017.

[5] Y. Ke and M. Hagiwara, "An english neural network that learns texts, finds hidden knowledge, and answers questions," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 4, pp. 229–242, 2017.

[6] T. Minemoto, T. Isokawa, H. Nishimura, and N. Matsui, "Pseudo-orthogonalization of memory patterns for complex-valued and quaternionic associative memories," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 4, pp. 257–264, 2017.

[7] O. Chang, P. Constante, A. Gordon, and M. Singana, "A novel deep neural network that uses space-time features for tracking and recognizing a moving object," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 7, no. 2, pp. 125–136, 2017.

[8] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[9] B. Lee, S. Amaresh, C. Green, and D. Engels, "Comparative study of deep learning models for network intrusion detection," *SMU Data Science Review*, vol. 1, no. 1, p. 8, 2018.

[10] Z. Marszalek, M. Wozniak, G. Borowik, R. Wazirali, C. Napoli, G. Pappalardo, and E. Tramontana, "Benchmark tests on improved merge for big data processing," in *2015 Asia-Pacific Conference on Computer Aided System Engineering*, July 2015, pp. 96–101.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 649–657. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969239.2969312

[13] D. E. Rumelhart, G. E. Hinton, R. J. Williams *et al.*, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[14] A. Conneau, H. Schwenk, Y. Cun, and L. Barrault, "Very deep convolutional networks for text classification," in *Long Papers - Continued*, vol. 1. Association for Computational Linguistics (ACL), 1 2017, pp. 1107–1116.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 505–514
DOI: 10.15439/2018F69
ISSN 2300-5963 ACSIS, Vol. 15

Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora

Nina Rizun
Gdansk University of Technology
Gabriela Narutowicza 11/12, 80-233
Gdansk, Poland
Email: nina.rizun@zie.pg.gda.pl

Yurii Taranenko
Alfred Nobel University, Dnipro
Naberezhna Lenina Str., 18, Dnipro,
49000, Ukraine
Email: taranenkonew@gmail.com

***Abstract* — Methodology of Constructing and Analyzing the Hierarchical structure of the Contextually-Oriented Corpora was developed. The methodology contains the following steps: Contextual Component of the Corpora's Structure Building; Text Analysis of the Contextually-Oriented Hierarchical Corpus. Main contribution of this study is the following: hierarchical structure of the Corpus provides advanced possibilities for identification of the Morphological and Structural features of texts of different tonalities; Contextual, Morphological and Structural specificity of texts with tonality, originally assigned by the authors, has significant differences; exist the certain thought and writing style Templates, under the influence of which the formation of texts of various tonalities takes place. As basic features of such templates for the texts of the two basic (positive/negative) tonalities could be used: Contextual Structure, Morphological Types, Emotional Features, Writing Style and Vocabulary Richness. For verification of the proposed methodology, a case study of Polish-language film reviews Dataset was used.**

INTRODUCTION

In recent years the sentiment analysis has been one of the hottest research areas in natural language processing [1]. The challenges to the researchers are both theoretical aspects, such as the objective laws of the sentiment expressions in the natural language, and the practical aspects, for example, the analysis of consumer opinions and reviews [2].

There are two main approaches to the sentiment analysis [3]: *lexicon-based* and *machine learning*. The first of them determines the text sentiment by means of individual words polarity in the text. The latter considers the task of sentiment analysis as the problem of text categorization. Both approaches require high quality sentiment lexicons: even in the text categorization methods the word weights are often proportional to word polarity and strength.

There are many studies on the problem of Sentiment Lexicons creation. They generally use three main approaches [4]: manual approach, dictionary-based approach, and corpus-based approach. In the manual approach the sentiment lexicons are constructed by human annotators. In the dictionary-based approach the sentiment lexicons are created with the help of the universal dictionaries and thesauri, e. g., WordNet [5].

In the corpus-based approach the sentiment lexicons are built based on the analysis of text corpora. Also, the various hybrid combinations of these approaches are used. Though the problem of Sentiment Lexicon creation is very important, little attention is paid to the evaluation of the quality and in-depth analysis of the generated lexicons, especially for Polish language.

Main direction of this research is to design the methodology for constructing and analyzing the Hierarchical Corpora, which allows improving the quality of the algorithms for the Sentiment Lexicon building by offering the additional tools for determining the tonality based on the availability of data about the semantic properties of the text. As the main research tool, text mining methods and algorithms will be used.

THEORETICAL BACKGROUND

Under the notion of texts mining we understand the application of methods of texts computer analysis and presentation in order to achieve the quality, which corresponds to the "manual" processing for further usage in various tasks and applications. One of the actual tasks of automatic texts mining is their clustering (definition of groups of the similar documents). More and more often statistical topical methods are being applied [6].

The topics are presented as discrete distributions on a number of words, and the documents – as discrete distribution on a number of topics [6]. Topical methods perform a "non-precise" clustering of words and documents, which means that a word or a document can be referred to a few topics with different probabilities simultaneously. The synonyms with higher probability will appear in the same topics since they are frequently used in the same documents. At the same time, the homonyms (words different in meaning, but similar in writing) will be placed in different topics because they are used in different contexts [7].

A. Preprocessing Procedure

Topical methods, as a rule, apply the method of a "bag-of-words", where each document is considered as a set of words not connected to each other. Before the topics are defined, the text is preprocessed – its Graphematic and Morphologic

This work was not supported by any organization

analysis is conducted with the objective to define the initial form of words and their meanings in the speech context.

Graphematic Analysis

To start the preprocessing procedure of a text it is necessary to divide the original unstructured text into sentences and words. At first sight, it is a very simple task, but it has its own specificities and plays an important role in the further analysis of a text.

Graphematic analysis includes:

– division of the original text into elements (words, separators);

– elimination of non-text elements (tags, meta-information);

– extraction and formalization of non-standard elements: structural elements: headlines, paragraphs, notes; numbers, dates, complexes of letters and numbers; names, patronymics, surnames; extraction of e-mail addresses;

– extraction of files' names;

– extraction of sustained phrases, words that are not used separately from each other.

In English sources, we can meet the definition of tokenization, which, by its content, is similar to the graphematic analysis. Tokenization – is a process of dividing the text stream into tokens: words, collocations and sentences [8]. Thus, the graphematic analysis is the initial analysis of an unstructured text, presented as a chain of symbols in any coding, elaborating information, which is necessary for further text processing.

There are almost no tools specializing exceptionally on graphematic analysis. Basically, graphematic is included into integrated packages of text analysis: NLTK, Stanford CoreNLP, Apache NLP, AOT, MBSP etc. The function of division into tokens is also included into programs of text markup, for instance into the part-of-speech taggers.

Morphologic Analysis

Morphologic analysis provides definition of the normal form, from which the word-form was created, and of the set of parameters, assigned to this word-form [9].

Stemming has been the most widely applied morphological technique for information retrieval. With stemming, the searcher does not need to worry about the correct truncation point of search keys. Stemming also reduces the total number of distinct index entries. With short queries and short documents, a derivational stemmer is most useful, but with longer ones the derivational stemmer brings in more non-relevant documents. Stemming increases search key ambiguity. Stemming may, however, is a non-optimal approach to the clustering of documents in agglutinative languages. Firstly, stemmers do not conflate compounds whenever the first components do not match exactly. Secondly, they are unable to split compounds, which typically have the head-modifier structure and the headword is the last and more important component for clustering [10]. The most widely-spread algorithm of stemming is the Porter's algorithm. Except for that algorithm there exists the Lancaster's algorithm (for English language) and the algorithms, working by the principle of a "snowball" (snowball stemmers) for other languages.

Lemmatization is another normalization technique: for each inflected word form in a document or request, its basic form, the lemma, is identified. The benefits of lemmatization are the same as in stemming. In addition, when basic word forms are used, the searcher may match an exact search key to an exact index key. Such accuracy is not possible with truncated, ambiguous stems. Homographic word forms cause ambiguity (and precision) problems – this may also occur with inflectional word forms [11]. Another problem is that words cannot be lemmatized, because the lemmatizer's dictionary does not contain them.

B. Vector Space Models of the Semantic Relations Analysis

The method of processing words in a machine-readable natural language, as a rule, is based on the vector-space method of data description (Vector Space Model) [12], suggested by [13]. Within the framework of the method each word in a document has its particular weight. Thus, each document is presented as a vector and its dimension is equal to the total number of words in the document.

Similarity of a document and a topic is evaluated as a scalar product of a few information vectors. The weight of separate words (terms) can be calculated both applying the absolute frequency of a term appearing in the text and the relative (normalized) frequency:

$$F_{w_i} = TF \times IDF = tf(w,t) \cdot \frac{\log_{10} D}{df} \qquad (1)$$

$tf(w,t)$ – relative frequency of the w-th term occurrence in document t:

$$tf(w,t) = \frac{k(w,t)}{df} \qquad (2)$$

$k(w,L_t)$ – the number of w-th term occurrences in the document t; df – the number of documents in the collection that contain the w-th term; D – total number of documents in the collection.

Then, for solving the problem of finding the similarity of documents (terms) from the point of view of the relation to the same topic, the different metric can be applied, for example:

– *Euclid's* measure:

$$dist_{t_i} = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}\,, \qquad (3)$$

where x – vector of the document, y – point of reference words vector;

– *Cosine* of the edge between the vectors:

$$dist_{t_i} = \cos\theta = \frac{x \cdot y}{\|x\| \cdot \|y\|}, \qquad (4)$$

where $x \cdot y$ – scalar product of the vectors, $\|x\|$ and $\|y\|$ – quota of the vectors, which are calculated by the formulas:

$$\|x\| = \sqrt{\sum_{i=1i}^{n} x_i^2}\ ,\ \|y\| = \sqrt{\sum_{i=1i}^{n} y_i^2} \qquad (5)$$

A further algorithm is to divide the source data into groups corresponding to the events, as well as in determining whether a text document describes a set of any topic. The main idea of the solution is the use of clustering algorithms [14] (e.g., k-means method, etc.). It is assumed that each cluster contains documents that describe an event.

Latent Semantic Analysis (LSA) is a Discriminant theory and method for extracting context-dependent word meanings by statistical processing of large sets of text data [15-17]. It uses the "bag-of-words" for modelling, begins with transforming text corpora into term-document frequency matrices, reduces the high dimensional term spaces of textual data to a user-defined number of dimensions by singular value decomposition (SVD), produces: weighted term lists for each concept or topic; concept or topic content weights for each document; outputs that can be used to compute document relationship measures [18].

According to the theorem on singular decomposition, any real rectangular matrix can be decomposed into a product of three matrices:

$$X_{t\times d} \approx X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T \qquad (6)$$

where $\Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T$ – represents terms in *k-d* latent space; $U_{K_{t\times d}} \Sigma_{K_{t\times d}}$ – represents documents in *k-d* latent space; $U_{K_{t\times d}}$, $V_{K_{t\times d}}$ – retain term–topic, document–topic relations for top *k* topics.

But, as [19, 20] proved, there are three *limitations* to apply LSA: documents having the same writing style; each document being centered on a single topic; a word having a high probability of belonging to one topic but low probability of belonging to other topics. The limitations of LSA is based on orthogonal characteristics of dimension factors as well as on the fact, that probabilities for each topic and the document distributed uniformly, which does not correspond to the actual characteristics of the collections of documents [13, 21, 22]. That is why, LSA tends to prevents multiple occurrences of a word in different topics and thus LSA cannot be used effectively to resolve polysemy issues.

C. Probabilistic Topic Models

To get rid of the above-mentioned disadvantages the probability LSA is conducted, based on the multinomial distribution – in particular, on the algorithm of Latent Dirichlet Allocation (LDA) [23, 24]. The *probabilistic topic modelling* – a set of algorithms to analyze the words in large sets of documents and from the retrieve the threads that connect into topics [25, 26]. In this case document is regarded as a set of words, the order of which does not matter. For each document to determine the distribution $\boldsymbol{\theta_d}$ of its words on topics, that probability for each topic meets it herein. This topic is presented in the form of distributions $\boldsymbol{\varphi_t}$ of words from a fixed vocabulary, i.e. each word included in the subject with a certain probability

The next text mining technique that was developed to improve upon LSA was the Probabilistic topic modeling techniques. Probabilistic topic modeling is a set of algorithms that allow analyzing words in textual corpora and extract from them topics, links between topics [23, 24, 27]. Latent Dirichlet Allocation (LDA) is a generative model that explains the results of observations using implicit groups, which allows one to explain why some parts of the data are similar. It was proposed by David Blei [23, 24] and it uses a Bayesian model that treats each document as a mixture of latent underlying topics, where each topic is modeled as a mixture of word probabilities from a vocabulary.

The algorithm of the method is the following: Each document is generated independently: randomly select for document its distribution on topics θ_d for each document's word; randomly select a topic from the distribution θ_d, obtained in the first step; randomly select a word from the distribution of words in the chosen topic φ_k (distribution of words in the topic *k*). In the classical model of LDA, the number of topics is initially fixed and specifies the explicit parameter *k*. In the process of assigning the topics to documents usually LDA uses the maximal from possible (not always very high) level of probability of documents belonging to the topic.

According to [28] – words in a topic from LDA (as an extended LDA method) are more closely related than words in a topic from LSA. For polysemy, words in a topic from LDA can appear in other topics simultaneously: topics are Dirichlet multinomial random variables, each word is generated by a single topic, and different words may be generated from different topics. The *limitation* of LDA is that there is no probability distribution model at the level of documents. Thus, the larger the number of documents, the larger the LDA model.

METHODOLOGY FOR CONSTRUCTING AND ANALYZING THE HIERARCHICAL CORPORA

A. Novelty and Motivation

The *purpose* of this research is development of the methodology of constructing and analyzing the *Hierarchical Corpora* intended for subsequent use in the creation of the Sentiment Lexicon using text mining tools.

In this research the following scientific research questions (RQ) were raised:

RQ: *Using what methods and algorithms it is possible to increase the quality of the formation of the Corpus intended for the analysis of text tonality*?

RQ_1: *Does creation of the Contextual Structure of the Corpus provides advanced possibilities for identification of the morphological and tonal specifics of analyzed texts?*

RQ_2: *Does the preliminary Morphological and Structural analysis of the Corpus allow to reveal specific characteristics of Corpus content in the light of improving the quality of texts Sentiment recognition?*

For finding the answers for these questions, the following assumptions (A) were formulated:

A1: Taking into account the specificity of the chosen case study [16, 17, 29], assume that each paragraph could be interpreted as a topically completed textual component (TCTC).

A2. Classified texts are characterized by their initially known subjective (author's) evaluations of their tonality.

On the basis of the research questions and assumptions, the following scientific hypotheses (H) were formulated:

H1. *Contextual structure of the certain type of texts writing does not depend on its tonality, initially assigned to it by the author.*

H2. *Morphological structure of the certain type of texts writing does not depend on its tonality, initially assigned to it by the author.*

H3. *Writing style of certain type of texts does not depend on the tonality, initially assigned to it by the author.*

H4. *Vocabulary richness of certain type of texts does not depend on the tonality, initially assigned to it by the author.*

As a *case study* for testing the basic workability and proposed Methodology quality, the Polish-language Film Reviews Dataset was used.

B. Contextual Component of the Corpora's Contextually-Oriented Hierarchical Structure Building

At the first stage of the methodology development, the authors take into account specificity of the chosen case study and the results of previous research results [29]. These results suggest the possibility of building the Hierarchical structure of the Contextually-Oriented Corpus (COHC) via application of the Discriminant and Probabilistic Methods of the Latent Dirichlet Allocation (LDA) and Latent Semantic Relations Analysis (LSA) [28, 30]. In our case COHC is the two-point (Positive/Negative Classes) structure of the sets of paragraphs, semantically close to Topics, identified as the main Contextual Framework of the analyzed initial Dataset. The process of Hierarchical structure of COHC *Building* involves the following stages (figure 1):

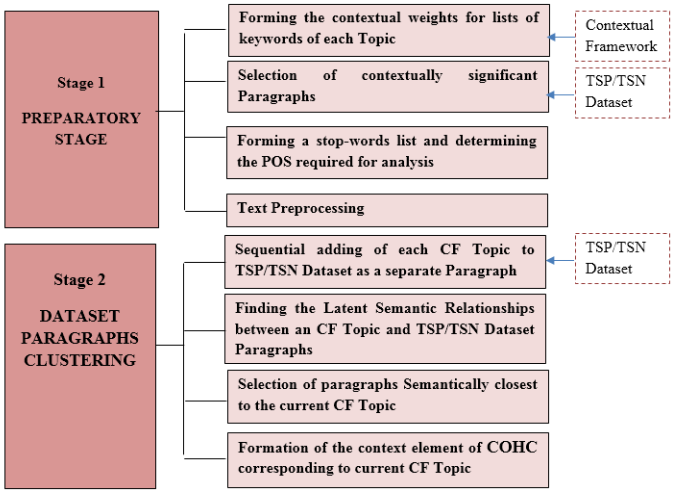


Fig 1. Steps of the Corpora's Contextually-Oriented Hierarchical Structure Building

Formation and analysis of the Contextual components of the COHC are based on the following concepts:

Concept A. Obtaining an adequate sentiment description of positive and negative texts tonality is possible only via formation of the Corpus on the basis of Truly Subjectively Positive (TSP, the subjective evaluation by reviewes is more than 8 points) and Truly Subjectively Negative (TSN, the subjective evaluation is less than 4 points) Dataset.

Concept B. As a Contextual Framework (CF) for COHC building the hierarchical structure of Topics (with list of keywords) for TSP and TSN dataset are used. Applied methods for CF creating – combination of LSA and LDA methods [16, 17, 28].

Concept C. As a quantitative measure of the degree of influence of each keyword from the CF Topics on the process of COHC building the contextual keyword weights (CKW) are used. Applied methods for CKW creation – combination of LSA and LDA methods, measure – the probability of occurrence of each word in the topic [16, 17, 28].

Concept D. As a tool for determining the belonging of each paragraph to a CF topic, the LSA is used [16, 17, 28].

Concept E. As a TCTC a paragraph of at least 100 characters should be used. The possibility to determine the topic of such paragraph with sufficient accuracy is experimentally proved [16, 17, 28].

C. Text Analysis of the Contextually-Oriented Hierarchical Corpus

The process of Text Analysis of COHC involves the following steps (figure 2).

Step 1. Morphological Analysis of the Contextually-Oriented Hierarchical Corpus

The purpose of this first stage is to conduct the COHC analysis and create the Hierarchical Morphological Framework (HMF) for each element of each COHC level.

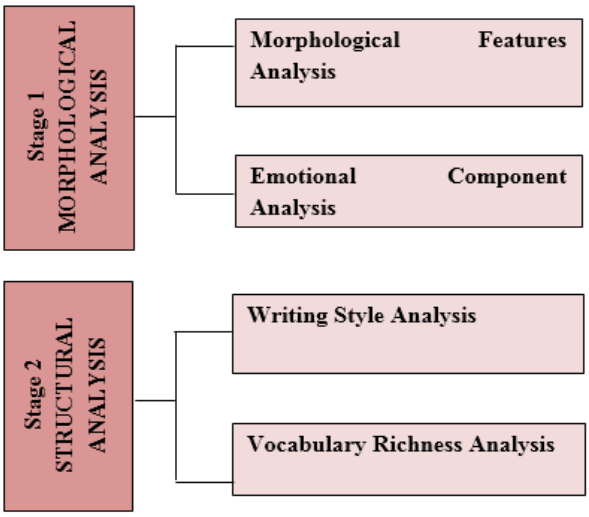


Fig 2. Steps of the Text Analysis of Contextually-Oriented Hierarchical Corpus

The objective of creating this HMF is to accumulate the Hierarchy of specific morphological types and emotional

features of the COHC texts to identify their differences depending on: the text tonality, initially assigned to it by the author, and belonging to a particular CF Topic. As main indicators to carry out this stage of analysis, the measures with following interpretations are proposed:

– the part of speech (POS) distribution for each COHC Element (*M1*) – determining the authors morphological types of the expression of positive or negative judgments;

– the percentage of new (unique) words in each COHC Element (*M2*) – characterizing the emotional component specificity of positive or negative judgments expression.

Step 2. Structural Analysis of the Contextually-Oriented Hierarchical Corpus

The purpose of this first stage is to conduct the COHC analysis and create the Hierarchical Structural Framework (HSF) for each COHC element of each level. The objective of creating this HSF is to accumulate the Hierarchy of specific writing styles and vocabulary richness features of the COHC texts to identify their differences depending on: the text tonality, initially assigned to it by the author, and belonging to a particular CF Topic. As main indicators to carry out this stage of analysis, the measures with following interpretations are proposed:

– the specificity of first Zipf's "rank-frequency" law for each element of each COHC level (*M3*), determining the authors *writing styles* of the expression of positive or negative judgments and classically characterized by a constant value of C as the ratio [31, 32]:

$$C = F*R, \quad (7)$$

where: F – frequency of occurrence of a term in the text; R – Rank of the word (the most commonly used word gets rank 1, the next – 2, etc.); C – constant;

– the specificity of the second Zipf's "quantity-frequency" law for each element of each COHC level (*M4*), determining the authors *vocabulary richness* of the expression of positive or negative judgments [5,6].

CASE STUDY RESULTS AND DISCUSSION

For testing and evaluating the adequacy of the author's Methodology realization, as a *case study* were used the training samples: 3000 Polish-language films reviews (1500 TSP and 1500 TSN) from the filmweb.pl.

All words/terms of film reviews in this paper will be presented in English language. The experimental part of all steps of author's Algorithm *was* technically realized in Python 3.4.1.

D. Contextual Component of the Corpora's Contextually-Oriented Hierarchical Structure Building

In the process of implementing this stage, about 30% of contextually insignificant paragraphs, and about 20% of paragraphs, for which topic could not be identified, were separated.

The quality indicator – recall rate as the ratio of the number of topically recognized paragraphs (probability of belonging the paragraph of topic >0,7) to the total number of paragraphs – is within 90-95% (table I).

As a method of evaluating the quality of probabilistic topic models the calculation of *Perplexity* index on the test data set [2-4] is used. In information theory, perplexity is a measurement of how well a probability model predicts a sample. A low perplexity indicates that the probability distribution is good for predicting the sample.

TABLE I.
STRUCTURE OF THE CONTEXTUAL SUMMARY

Corpora Samples	Number of paragraphs	Number of CF topics	Average Number of topics in Document	Average Number of terms in	Average Perplexity Value
TSP	10239	36730.0	5.1	5.7	1 182 169
TSN	10934	41015.0	4.2	6.1	1 342 155

As a result, the following structure of the two-level two-point *Contextually-Oriented Hierarchical Corpora of Polish-Language Film Reviews* [29] was obtained (tables II- III).

TABLE II.
CONTEXTUAL STRUCTURE OF THE SUBJECTIVELY POSITIVE (SP) ELEMENTS OF THE COHC OF POLISH-LANGUAGE FILM REVIEWS

Element of 1st level CF l	Element of 2nd level CF	% of paragraphs
"Hero"	Actor / Play	24%
	History / Film	43%
	Picture / Scene	30%
	Director / Creator	3%
"Director"	Film / Director	30%
	Scene / Story	10%
	Style	6%
	Creator / Author	54%
"Script"	Film / Director	8%
	Story / Hero	58%
	Author / Creator	13%
	Role / Actors	21%
"Plot"	Film / Effects	5%
	Portrait / Image	31%
	Director / Production	24%
	Script / History	40%
"Spectator"	Hero / Fan	40%
	Film / Aspects	20%
	Role / Formulation	16%
	Scene / Director	24%

TABLE III.
CONTEXTUAL STRUCTURE OF THE SUBJECTIVELY NEGATIVE (SN) ELEMENTS OF THE COHC OF POLISH-LANGUAGE FILM REVIEWS

Element of 1st level CF	Element of 2nd level CF	% of paragraphs
"Hero"	Action / History	49%
	Director / Cinema	21%
	Scene / Actor	31%
"Actor"	Hero / Image	24%
	Role / Scene	58%
	Script / History	18%
"Creator"	Hero / Scene	23%
	Film / Script	60%
	Picture / Actor	18%
"Plot"	Story / Hero	39%
	Director / Image	18%
	Creator / Film	43%

The results obtained at this stage of the experiment indicate that the tonality, initially assigned by the authors specificity of texts with Persuasive writing type, affects the Contextual Structure of the analyzed content. As can be seen from the tables II-III, the Contextual Structure of Subjectively Positive and Subjectively Negative elements of COHC differs both in content and in the variety (amount) of topics covered in the texts (H2 is rejected).

E. Text Analysis of the Contextually-Oriented Hierarchical Corpus

Step 1. Morphological Analysis of the Contextually-Oriented Hierarchical Corpus

As a result of the specific morphological types and emotional features of the COHC texts identification, the following initial statistics were obtained (Table IV)

TABLE IV.
FREQUENCY CHARACTERISTICS OF THE PARTS OF SPEECH DISTRIBUTION IN SP / SN HIERARCHICAL CORPORA'S ELEMENTS

Negative Hierarchical Corpora's Elements				**Positive Hierarchical Corpora's Elements**			
Morphological Types (M1)		**Emotional Features (M2)**		**Morphological Types (M1)**		**Emotional Features (M2)**	
% of adjectives	**Frequency**	**% of Unique adjectives**	**Frequency**	**% of adjectives**	**Frequency**	**% of Unique adjectives**	**Frequency**
18,09003	1	46	1	12.50	2	38	1
19,61974	1	53.5	2	15.02	0	50.4	2
21,14945	4	61	3	17.55	0	62.8	2
22,67915	9	68.5	4	20.07	0	75.2	9
More	1	More	6	22.59	12	87.6	6
				More	11	More	5
% of nouns	**Frequency**	**% of Unique nouns**	**Frequency**	**% of nouns**	**Frequency**	**% of Unique nouns**	**Frequency**
0.906111	1	38	1	52.88136	1	31	1
15.52333	0	46.5	2	56.16872	5	42.6	1
30.14056	0	55	5	59.45609	12	54.2	3
44.75778	0	63.5	4	62.74345	4	65.8	8
More	15	More	4	66.03082	1	77.4	7
				More	2	More	5
% of verbs	**Frequency**	**% Unique verbs**	**Frequency**	**% verbs**	**Frequency**	**% of Unique verbs**	**Frequency**
20.02801	1	51	1	13.43874	1	43	1
21.01446	4	60.75	2	15.22556	0	54.4	1
22.00092	6	70.5	3	17.01239	2	65.8	1
22.98737	2	80.25	6	18.79922	3	77.2	4
More	3	More	4	20.58605	5	88.6	9
				More	14	More	9

In table IV the "% of part of speech" is the border of % of these types of words in the whole number of words in particular Corpora's Element; "Frequency" – the number of COHC elements in which this "% of part of speech" occurs.

The results of comparative analysis of differences in the part of speech distribution and percentage of new words in the different COHC elements could be interpreted in the following way:

1. The law distribution of adjectives in the positive and negative COHC *Elements* indicates that:

– % of adjectives used in *positive* elements of the Hierarchical Corpora is slightly higher than this percentage in *negative* Elements.

This result can be interpreted as the presence of a general tendency to make reviews more intonational in expressing positive emotions;

– % of new (unique) adjectives used in positive elements of the Hierarchical Corpora is significantly higher (by 20%) compared to this indicator in negative Elements.

Thus, the need and realization of the emotional component of the authors' positive judgments through the use of different adjectives (characteristics) is much higher than in the expression of negative emotions.

2. The distribution of nouns in the positive and negative COHC Elements indicates that:

– % of the nouns, used in positive Hierarchical Corpora's Elements, obeys the classical normal distribution law and indicates the average weightiness of the judgments expressed. Negative judgments are characterized by extremes – either many, or very few nouns;

– % of new (unique) nouns used in positive Hierarchical Corpora's Elements is higher (about 10%) compared to this indicator in negative reviews.

Since the nouns primarily serve to ascertain the facts, the existence of objects (entities, etc.), on the whole these facts can indicate that negative judgments are more based on emotions rather than facts.

3. The distribution law of verbs in the positive and negative COHC Elements indicates that:

– % of verbs used in *positive* elements of the Hierarchical Corpora is insignificant, and even lower than this percentage in *negative* Elements.

This can be interpreted as the desire of reviewers to characterize the actions that caused the particular emotions, in a greater degree;

– % of new (unique) verbs used in *positive* elements of the Hierarchical Corpora is higher (about 10%) compared to this indicator in negative reviews.

This again testifies to a more creative approach to writing reviews by authors of positive reviews.

In general, these facts may indicate that the expression of negative emotions is characterized by greater stinginess of emotional coloring (from the point of view of linguistic evaluation).

When generalizing the results obtained at this stage of the experiment, it can be argued that the tonality, initially assigned by the authors specificity of texts with Persuasive writing type, affects the Morphological Structure of the analyzed content. As it can be seen from table V, the Morphological Structure of Subjectively Positive and Subjectively Negative elements of COHC differs both in the morphological types and emotional features (H2 is rejected).

TABLE V.
MORPHOLOGICAL TYPES AND EMOTIONAL FEATURES IN SP / SN HIERARCHICAL CORPORA'S ELEMENTS

Part of speech	SP Elements		SN Elements	
	M1	M2	M1	M2
Adjectives	High	High	High	High
Nouns	Average	High	Polar	High and Average
Verbs	High	High	Average	Average

Step 2. Structural Analysis of the Contextually-Oriented Hierarchical Corpus

a) First Zipf's "Rank-Frequency" Law

As a result of comparative analysis of writing styles of the expression of positive or negative judgments, the following types of internal structure of the COHC Elements from the point of view of the specific of the Rank-Frequency distribution (*writing styles*) of the words usage were identified:

– the *Classical* structure of the COHC Elements, in which the proportion of terms with a high frequency of usage (in the authors' algorithm it is intended to remove the often-used and not load-bearing stop-words of the Polish language at the stage of preprocessing) account for no more than 0.25% of all terms (C≈0.06-0.07);

– the *Medium normalized* structure of the COHC Elements, in which terms with a high frequency of use account for about 10% of all terms (C≈0.02-0.05);

TABLE VI.
THE RANK-FREQUENCY DISTRIBUTION STRUCTURE OF THE 1ST LEVEL OF HIERARCHICAL CORPORA'S ELEMENTS

Element of 1st level CF	Classical				Medium normalized				Non-standard			
	All POS	Adjective	Nouns	Verbs	All POS	Adjective	Nouns	Verbs	All POS	Adjective	Nouns	Verbs
Positive Hierarchical Corpora's Elements												
"Hero"	18.75%	12.50%	25.00%		6.25%	6.25%		18.75%	6.25%			6.25%
"Spectator"					18.75%	6.25%	25.00%		6.25%	18.75%		25.00%
"Script"	25.00%	12.50%	25.00%					12.50%		12.50%		12.50%
"Director"	12.50%		6.25%		6.25%	12.50%	12.50%	12.50%	6.25%	12.50%	6.25%	12.50%
"Plot"	6.25%		6.25%		12.50%	6.25%	12.50%	6.25%	6.25%	18.75%	6.25%	18.75%
Negative Hierarchical Corpora's Elements												
"Hero"	25.00%		25.00%	8.33%		25.00%		16.67%				
"Actor"	16.67%		25.00%		8.33%	16.67%		8.33%		8.33%		16.67%
"Creator"	25.00%		25.00%			25.00%		16.67%				8.33%
"Plot"	16.67%		16.67%		8.33%	25.00%	8.33%					25.00%

– the *non-standard* COHC Elements structure, in which about 90% of terms have a frequency of use no more than 1 time (C≈0.03-0.04).

Characteristics of Rank-Frequency distribution, obtained during this stage of experiment for the 1st level of COHC, are presented in table VI.

Interpretation of these results as a characteristic of specific Writing Style, could be the following:

– negative reviews are characterized by a small part of opinions, expressed with the help of unique words. And the most non-standard from the point of view of the use of unique words and the brevity of presentation are the paragraphs characterizing the topic Plot of the film with the help verbs.

Adjectives used in negative reviews mainly refer to the second type of Hierarchical Corpora structure, which can be interpreted as a fairly high percentage of frequently repeated definitions (terms).

The predominant type of COHC Elements structure are standard reviews, in which the most commonly used words account for 0.25% of all COHC Element vocabulary;

– structure of the *positive* part of the COHC is fairly uniform – all the texts represented in it are equally structured. However, from the point of view of distinctive features, it is necessary to note the percentage of non-standard (unique) adjectives.

The structure of the COHC that characterizes the Spectator is especially different – in this part of the COHC there is no standard Corpora Elements structure, and there is a high percentage of both repeating and unique verbs.

Generalizing the results obtained at this stage of the experiment (table VII) it again can be argued that the tonality, initially assigned by the authors specificity of texts, affects the Writing Style of the analyzed content (H3 is rejected).

TABLE VII.
WRITING STYLES STRUCTURE OF THE HIERARCHICAL CORPORA OF POLISH-LANGUAGE FILM REVIEWS

Writing Style (M3)	SP Hierarchical Corpora's Element	SN Hierarchical Corpora's Element
Classical	30.00%	45.83%
Medium normalized	35.00%	39.58%
Non-standard	35.00%	14.58%

b) Second Zipf's "Quantity-Frequency" Law

As a result of the experiments, the initial statistics, which describes the specificity of second Zipf's "quantity-frequency" law for Polish-Language Film Reviews COHC were obtained.

The basic coefficients for the analysis were the coefficients of the approximation in the equation for the second Zipf's law:

$$y = a + \frac{b}{x} \qquad (8),$$

where, a – determining the average frequency of occurrence of most part of the terms in the COHC Element; b – determining the average speed of appearance of new words in the text – the *Vocabulary Richness* (M4) of text's author.

The *lower* the value of the coefficient b, the higher the richness of the vocabulary of the COHC Element, since the curve of the dependence of the occurrence frequency of each word in the number of these words decreases more quickly, accordingly a smaller number of terms appears frequently (that is, the same words are used more often).

In order to ensure the adequate comparability of the conducted studies results, the corrected (taking into account the number of unique words in the COHC Element) coefficients of equations a and b were used.

The characteristics of Quantity-Frequency distribution (Zipf's law coefficients), obtained during this stage of experiment, are presented in the figures 3-4.

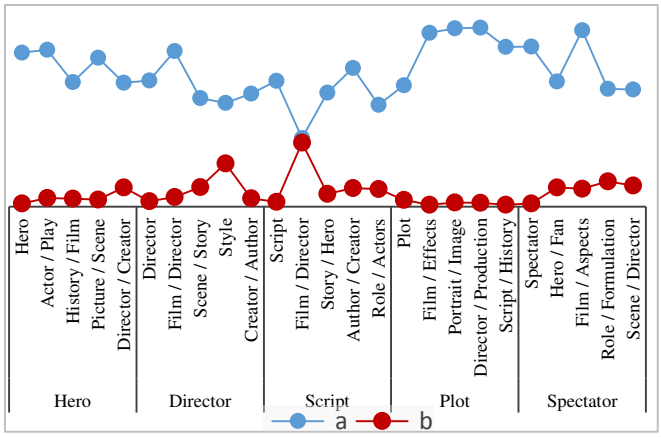


Fig 3. Second Zipf's law coefficients for Positive Hierarchical Corpora's Elements

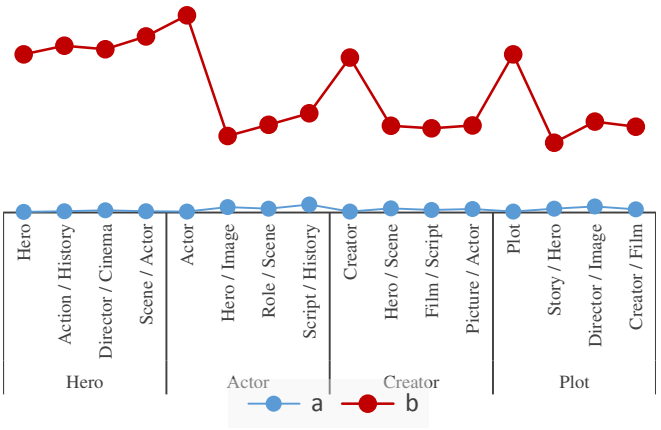


Fig 4. Second Zipf's law coefficients for Negative Hierarchical Corpora's Elements

Based on the obtained data, the comparative analysis of the specificity of second Zipf's "quantity-frequency" law coefficients distribution of the Positive and Negative Elements of Polish-Language Film Reviews COHC was carried out (table VIII), where "Frequency" – the number of COHC elements in which such value of coefficients occurs:

TABLE VIII.
STRUCTURE OF THE SECOND ZIPF'S LAW COEFFICIENTS DISTRIBUTION (M4)

Positive COHC Elements				Negative COHC Elements			
Coefficient a	Frequency	Coefficient b	Frequency	Coefficient a	Frequency	Coefficient b	Frequency
0.016	3	0.004	22	0.001	9	0.029	9
0.022	12	0.007	2	0.001	5	0.041	0
0.028	10	0.010	1	0.002	2	0.052	7

Interpretation of the specificity of Polish-Language Film Reviews COHC from the point of view of the vocabulary richness of the expression of positive or negative judgments (table VIII) could be the following:

– *positive* reviews are characterized by an initially high (in comparison with negative) values of the corrected coefficient ***a*** (a≈0.010-0.027), which indicates a high average level of frequency of most part of the terms in the COHC Element. At the same time, this part of the case is characterized by relatively low values of the corrected coefficient ***b*** (b≈0.0003-0.0009), on the one hand, testifying to a sufficiently rich (in comparison with negative reviews) vocabulary of the text. That is, in general, positive reviews characterized by a greater proportion of terms that are used *uniformly often* throughout the text. This, in turn, may indicate a rather *highly semantic structured opinion*, expressed in a carefully and balanced manner;

– the *negative* reviews are characterized by sufficiently low (in comparison with positive) values of the corrected coefficient ***a*** (a≈0.0002-0.0021), which indicates a lower (i.e., more unique) average level of frequency of most part of the terms in the COHC Element. In this case, this part of the COHC is characterized by sufficiently high values of the

corrected coefficient ***b*** (b≈0.0183-0.0517), which may indicate that in the expression of the main negative emotions, authors use the same words quite often, and the rest of the words are used *randomly*, depending on the context of the film or the specific expression of the author's thoughts. This, in turn, can testify to the *average level of semantic structure of the opinion*, expressed more spontaneously and under the influence of emotions.

Generalizing the results obtained at this stage of the experiment (table IX) it again can be argued that the tonality, initially assigned by the authors specificity of texts, affects the Vocabulary Richness of the analyzed content (H4 is rejected).

TABLE IX.
VOCABULARY RICHNESS STRUCTURE OF THE HIERARCHICAL CORPORA OF POLISH-LANGUAGE FILM REVIEWS

Vocabulary Richness (M4)	**SP COHC Element**	**SN COHC Element**
Average frequency of words occurrence	High	Low
Average speed of new words appearance	Low	Polar
Vocabulary Richness	Sufficiently High	Random
Semantic Structuredness	Highly Semantic Structured	Medium Semantically Structured

CONCLUSIONS

In this paper, authors present the methodology of constructing and analyzing the *Hierarchical Corpora* for the Purpose of Further Forming and Training the Sentiment Lexicon. The main contribution of the paper and the authors' study is finding the answers to the main research questions:

1. The hierarchical structure of the Corpus allows for more flexible and clear identification of the Morphological and Structural features of texts of different tonalities and contexts, originally assigned by the authors.

These differences should contribute to improving the quality of algorithms development for the Sentiment Lexicon creation, allowing the introduction of additional tools for determining the tonality based on the availability of data about semantic properties of the text being studied.

2. The Contextual, Morphological and Structural specificity of texts that have a tone, originally assigned by the authors, has significant differences. The basic features of the Sentiment Patterns for the texts of the two basic tonalities, which were obtained, are the following (table X):

Table X.
SENTIMENT PATTERNS OF THE TEXTS WITH THE DIFFERENT TONALITY

Features	**Subjectively Positive Texts**	**Subjectively Negative Texts**
Contextual Structure	Wide	Moderate
Morphological Types	More Adjectives and Verbs	More Adjectives and Partly Verbs
Emotional Features	Very Emotional	Restrained Emotionally
Writing Style	Colorful	Monochrome
Vocabulary Richness	Structured and rich	Medium Structured and Random

The frameworks obtained by the authors testify to the existence of certain thought and style templates, under the influence of which the formation of texts of various sentiments takes place.

ACKNOWLEDGEMENTS

The research results, presented in the paper, are partly supported by the Polish National Centre for Research and Development (NCBiR) under Grant No. PBS3/B3/35/2015, the project "Structuring and classification of Internet contents with the prediction of its dynamics".

REFERENCES

[1] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at International Conference on Learning Representations* (ICLP), 2013. http://arxiv.org/abs/1301.3781

[2] Feldman, R. Techniques and applications for sentiment analysis. *Communications of the ACM*, 2013. 56(4), pp. 82-89.

[3] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Word and Phrases and their Compositionaly. *Proceedings of Workshop at The Twenty-seventh Annual Conference on Neural Information Processing Systems*. 2013 (NIPS) http://arxiv.org/abs/1310.4546

[4] Mikolov T., Le Q. Distributed Representations of Sentences and Documents. *Proceedings of Workshop at The 31st International Conference on Machine Learning* (ICML). 2014. http://jmlr.org/proceedings/papers/v32/le14.pdf.

[5] Elias P. Interval and recency rank source encoding: two on-line adaptive variable-length schemes. *IEEE Trans. Inform. Theory*. 1987. V. 33, N 1. P. 3–10.

[6] Popescu, I.-I., Altmann, G., Čech, R. The Lambda-structure of Texts. *Lüdenscheid: RAM-Verlag,* 2011.
(1) Vocabulary Richness Measure in Genres. Available from: https://www.researchgate.net/publication/258518594_Vocabulary_Richness_Measure_in_Genres [accessed Jul 10 2018].

[7] Dempster, A.P., Laird, N.M., & Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. 1977. Series B., 39(1), 1-38.

[8] Feinerer, I., Hornik, K. & Meyer, D. Text mining infrastructure in R. *Journal of statistical software*, 2008. 25(5). American Statistical Association.

[9] Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications*. 2003.

[10] Koreniu T., Laurikkala J., Järvelin K., & Juhola M. Stemming and Lemmatization in the Clustering of Finnish Text Documents. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004. Washington, DC, USA, 625-633.

[11] Alkula, R. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval*, 2001. 4, 195-208.

[12] Nokel, M. A. & Lukashevich, N.V. Thematic Models: Adding Bigrams and Accounting Similarities Between Unigrams and Bigrams. *Computational methods and programming*, 2015. 16, 215-217

[13] Salton G., Wong A., Yang C.S. (A vector space model for automatic indexing. *Communications of the ACM*. 1975. Volume 18. Issue 11, pp. 613-620

[14] Jain A.K., Murty M.N. & Flynn P.J. Data Clustering: A Review; *ACM Computing Surveys*, 1999. 31 (3), 264-323. http://dx.doi.org/10.1145/331499.331504

[15] Papadimitrious, C.H., Raghavan, P., Tamaki, H., and Vempala, S. Latent semantic indexing: A probabilistic analysis. Journal of Computer and System Sciences, 2000. 61, 217-235

[16] Rizun N., Ossowska K., Taranenko Y. Modeling the Customer's Contextual Expectations Based on Latent Semantic Analysis Algorithms. *Information Systems Architecture and Technology*: 38th

International Conference on Information Systems Architecture and Technology. 2018, pp.364-373.

[17] Rizun N., Taranenko Y., Waloszek W. The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models. *Knowledge Engineering and Semantic Web. 8th International Conference*, 2017, pp.53-68.

[18] Patricia J. Crossno, Andrew T. Wilson and Timothy M. Shead, Daniel M. Dunlavy. Topic View: *Visually Comparing Topic Models of Text Collections*. 2011.

[19] Leticia H. Anaya. (2011). *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers*, Doctor of Philosophy (Management Science), 2011. 226 pp

[20] Papadimitrious, C.H., Raghavan, P., Tamaki, H., and Vempala, S. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 2000. 61, 217-235.

[21] Deerwester S., Susan T. Dumais, Harshman R. *Indexing by Latent Semantic Analysis*. 1990. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

[22] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. Using latent semantic analysis to improve information retrieval. *Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: 1988. ACM, 281-285

[23] Blei, D. M. Introduction to Probabilistic Topic Models. *Communications of the ACM*, 2012. 55 (4), 77-84.

[24] Blei, D. M., Ng, A., & Jordan, M. Latent Dirichlet Allocation. *International Journal of Advanced Computer Science and Applications (3):* 2003. 147-153.

[25] Anagha R Moosad, Aiswarya V., Subathra P and P.N. Kumar. Browsing Behavioural Analysis Using Topic Modelling. *International Joural of Computer Technology and Applications*. 2015. No.8. Issue No. :5. Pp. 1853-1861

[26] Alghamdi R. Alfalqi K. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications* (IJACSA), 2015. Volume 6 Issue 1.

[27] Daud Ali, Li Juanzi, Zhou Lizhu, Muhammad Faqir. Knowledge discovery through directed probabilistic topic models: a survey. *Proceedings of Frontiers of Computer Science in China*. 2010. pp. 280-301

[28] Lee, S., Song, J., and Kim, Y. An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, 2010

[29] Rizun N., Taranenko Y., Waloszek W. The Algorithm of Building the Hierarchical Contextual Framework of Textual Corpora. *Eighth IEEE International Conference on Intelligent Computing and Information System,* 2017, pp.366-372.

[30] Rizun N., Kucharska W. Text Mining Algorithms for Extracting Brand Knowledge from Facebook. The Fashion Industry Case. *International Business Information Management Conference*. 2018.

[31] Mandelbrot B. On recurrent noise limiting coding. *Lab. d'Electronique et de physique appliquces*. 1954. Paris.

[32] Mandelbrot B. In the theory of word frequencies and on related markovian models of discourse. The structure of language and its mathematical aspects. Providence, RI: Amer. Math. Soc. 1961. pp. 190–219. *Proceeding Symposium on Applied Mathematics*; V. 12.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 515–522
DOI: 10.15439/2018F227
ISSN 2300-5963 ACSIS, Vol. 15

A Comparative Study of Classifying Legal Documents with Neural Networks

Samir Undavia, Adam Meyers, John E. Ortega
New York University
60 5th Avenue
New York, New York 10011, USA
Email: {su478,meyers,jortega}@cs.nyu.edu

***Abstract*—In recent years, deep learning has shown promising results when used in the field of natural language processing (NLP). Neural networks (NNs) such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used for various NLP tasks including sentiment analysis, information retrieval, and document classification. In this paper, we the present the Supreme Court Classifier (SCC), a system that applies these methods to the problem of document classification of legal court opinions. We compare methods using traditional machine learning with recent NN-based methods. We also present a CNN used with pre-trained word vectors which shows improvements over the state-of-the-art applied to our dataset. We train and evaluate our system using the Washington University School of Law Supreme Court Database (SCDB). Our best system (word2vec + CNN) achieves 72.4% accuracy when classifying the court decisions into 15 broad SCDB categories and 31.9% accuracy when classifying among 279 finer-grained SCDB categories.**

I. Introduction

Legal court opinions are lawful statements written by a judge providing the justification and legal reasoning for a court ruling. This paper describes an automated document classification model implemented as our Supreme Court Classifier (SCC) system. In theory, SCC could make obsolete many time-consuming manual tasks requiring legal experts. A legal expert would need to read hundreds or thousands of documents in order to place opinions into subject categories, whereas an automatic system like SCC could do this with little or no human effort.[1]

Some document classification efforts, particularly, those using unsupervised approaches, evaluate output based on human evaluation of automatically derived categories. However, when automatic document classification is based on human-defined categories, the results are, arguably, more "natural." Evaluation tends to be more straightforward with human-annotated classifications because it is usually easy for a human being to tell whether or not a document belongs to a human-defined category. In contrast, this determination is harder to make with purely unsupervised methods (e.g., topic modeling [1]), unless a manual component is added. For example, aligning automatically defined categories with some set of human categories will produce clearer results. Human-defined categories have names and notional definitions such as *Criminal Procedures*, *Civil Rights*, and *Federal Taxation* [2].

In contrast, automatically classified categories are usually defined as sets of keywords or other more oblique definitions using words in the corpus. For instance, the case *Roe v. Wade*, 410 U.S. 113 (1973)[2], may fit into a class labeled by a set of keywords like *abortion, reproduction, medical, ...* Although these words describe the case, they do not correctly encapsulate the legal significance of the case. Roe v. Wade would be classified under the *Privacy* legal issue and the *abortion: including contraceptives* sub-issue [2]. Unfortunately, it may be difficult for a human to decide what the boundaries of the classes are defined by these keyword sets. On the other hand, it may be possible to align the output of unsupervised classification with a manual set of categories. In fact, we do this for our baseline system that uses latent Dirichlet allocation (LDA) to model documents and then a logistic regression (LR) [3] model to align the results with the manual categories (see Section IV-A).

While categories can be extracted in many ways, we believe that some sort of human validation is preferable. Moreover, the legal domain often requires documents to be aligned with human-defined legal categories for a majority of legal tasks. For this reason, we have implemented a system that uses pre-defined document categories (from human evaluators) to train a model for classifying legal texts. Specifically, SCC automatically classifies legal opinions from cases seen by the Supreme Court of the United States (SCOTUS), manually classified into topics as part of the Supreme Court Database (SCDB) by Washington University School of Law [2]. Henceforth, we will refer to these as the SCDB categories. The SCDB categories are defined in Section III-A.

Our work systematically tests the application of a variety of machine learning (ML) techniques to automate semantic classification of SCOTUS legal opinions. Our most successful systems are based on neural networks (NNs). NNs are used to solve a variety of natural language processing (NLP) tasks because of their ability to extract relevant information from

[1]SCC can be downloaded from https://github.com/samir1/web_of_law_scotus_classification/ under an Apache 2.0 license (https://www.apache.org/licenses/LICENSE-2.0). In addition to computer code, the repository includes our training/development/test split of the SCDB data, ensuring that our results are both reproducible and comparable to future work.

[2]Roe v. Wade is a supreme court case that is famous in the United States for causing abortion to become legal.

text without having to specify features for any particular domain [4]. We examine two common NN architectures: the convolutional neural network (CNN) [5] and the recurrent neural network (RNN) [6], much like the medical text classification experiments using CNNs conducted by Hughes et al. [7]. Initially used for classifying images, variations of the original CNN architecture are used for NLP tasks [8]. Two main variations of the RNN, long short-term memory (LSTM) [9] and gated recurrent unit (GRU) [10], have recent successful results in their application to sequence modeling [11], [12].

We compare a series of different CNN and RNN architectures to documents represented by word embeddings. We show that our CNN performs better with the legal corpus than the other models implemented. SCC uses neural networks to select one of the SCDB categories for each supreme court case in our validation corpus.

In this paper, we use CNNs and RNNs to classify legal documents with minimal pre-processing, in contrast to other machine learning approaches (e.g., support vector machines) that require manually specifying features for classification or manually determining key words [13]. Any additional pre-processing that could potentially improve performance requires manual editing since each document contains slightly different formatting resulting from OCR errors from scans of printed documents. We measure the efficacy of NN classification techniques applied to our corpus and show that our CNN outperforms RNNs for legal text classification based on SCDB categories (see Table I in Section V). In order to apply our classification models to text, we first represent each word in our corpus as a word embedding (vector representations of words were generated using an unsupervised learning method from Mikolov et al. [14]). We use the publicly available pre-trained word vectors trained on about 100 billion words from part of the Google News dataset.[3] We use 300-dimensional word2vec vectors trained using the skip-gram architecture with negative sampling by Mikolov et al. [14]. We map each word to a word vector and use neural network classifiers on the dataset. Additionally, we present results from using two other word embedding models, fastText [15] and GloVe [16], with our CNN (our best system). We describe the neural neural network models in Section IV and results in Table I.

II. Related Work

We have found a relatively small body of previous work about automatic text classification of legal documents. For example, support vector machines (SVMs) have been used to classify legal documents like legal docket entries [17] and to classify non-English legal texts [4]. Although our work also examines the application of machine learning to a corpus written in the legal context, we focus on classifying SCOTUS legal opinions without significant pre-processing. For example, the Nallapati and Manning [17] system includes several steps of pre-processing before using an SVM to classify documents with human-selected features and labels. We explore more recent automated document classification techniques that do not rely on significant pre-processing and human interaction. Moreover, we present a comparison of different machine learning techniques in order to determine methods with the highest performance for our task.

Our work on SCC is similar to the approach of Wood et al. [1] for classifying medical summaries in that we model our corpus using LDA and classify with pre-defined labels (see Section IV-A). In that study, an initial topic model was derived from some training documents. Then the topic model was modified with pre-labeled data in order to classify the new data. Likewise, we use a combination of LDA and pre-labeled legal opinions to create our baseline classification system. We compare the results of our NN-based classification results to our application of LDA and an LR classification.

Domain-specific automated document classification has been applied to several fields, including electronic medical records. Hughes et al. [7] use convolutional neural networks to detect features for sentence-level classification of medical texts, resulting in a much smaller input compared to our experiments. Unlike SCC, in which we feed an entire document into a neural network, the Hughes et al. [7] model classifies texts by first transposing each document into a matrix of sentences with fixed lengths. Their model also differs from ours in that their model is essentially two sets of two stacked convolutional layers followed by a pooling layer, whereas our model does not have any consecutive convolutional layers. Additionally, one of our experiments (following Hughes et al. [7]) tests the effectiveness of using doc2vec embeddings with an LR model as the classifier. Similarly, Weng et al. [18] use medical texts as the subject of their classification task, although they use a different neural network architecture to classify health documents. Weng et al. [18] apply a complex combination of CNN and RNN architectures to clinical note text classification; their model is summarized by three convolutional and pooling layers followed by a bidirectional LSTM [18]. Moreover, Yin et al. [19] present a comparison of different neural network architectures used to complete a variety of NLP tasks. Such tasks include sentiment analysis, document classification, and part-of-speech tagging. In their text classification experiment, Yin et al. [19] use a pre-labeled set of 10,717 sentences evenly distributed over 19 labels, compared to our unevenly distributed dataset, in which a third of the categories have under one hundred examples and four classes have over 1,000 documents. In contrast, our experiments aim to solve the specific problem of document classification applied to legal texts. Moreover, Kim [20] describes a general CNN used to classify sentences with word2vec word embeddings. Similar to the model we propose, Kim's model also includes three convolutional and pooling layers. We optimize hyperparameters and experiment with a combination of different convolutional, pooling, and dropout layers. We compare applications of the Kim [20] and Hughes et al. [7] models to our legal corpus. We ultimately obtain improved results by using the customized CNN on the SCOTUS legal opinion corpus.

[3]https://code.google.com/archive/p/word2vec/

III. Experimental Setup

A. Dataset

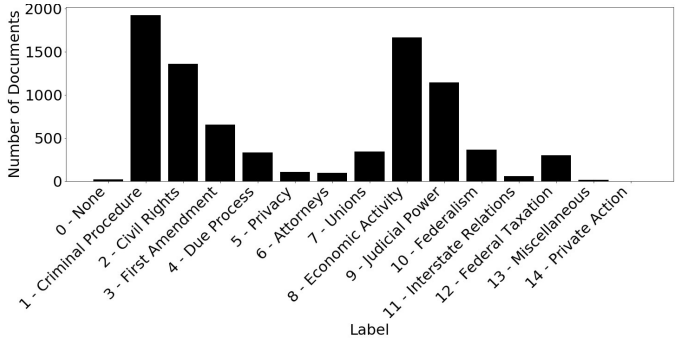


Fig. 1: 15 legal issues

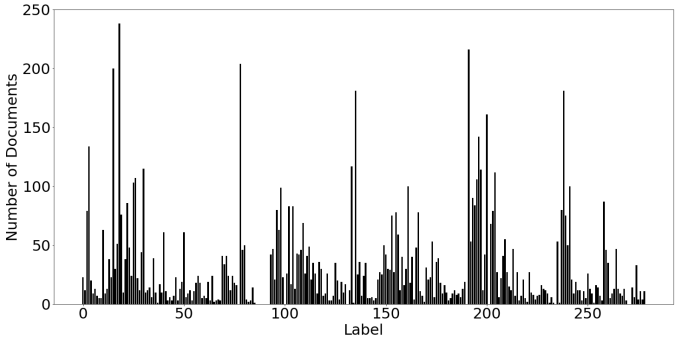


Fig. 2: 279 subtopics

We train and test our system on the manually-categorized SCOTUS legal opinion (Supreme Court Database or SCDB) corpus, from Washington University School of Law [2]. The dataset consists of 8419 US Supreme Court court opinions from "modern" cases (1946-2016), organized into 15 legal categories (Figure 1), which are further divided into 279 subtopics (Figure 2). We chose the modern dataset because both court opinions and pre-defined labels were available through the textacy Python package.[4] Textacy provides a formatted version of modern cases from FindLaw's US Supreme Court legal opinions database. The 8419 documents were randomly divided into training, validation, and test sets with a 80%/10%/10% split. Although the SCDB labels also covered "legacy" cases (1791-1945), FindLaw's database only reliably provided case text from the "modern" era of US law.

B. Initial Processing

Our system first removes (as noise) special characters that refer to footnotes. We also removed a number of characters used in formatting the original printed document. Next, each word in the corpus was mapped to a word2vec vector before being fed into a neural network for classification.

[4]http://textacy.readthedocs.io/en/latest/_modules/textacy/datasets/supreme_court.html

IV. Methods

A. LDA + Logistic Regression

Before the widespread use of neural networks for NLP tasks, probabilistic methods like LDA [21], [22], were used as a standard for a variety of NLP tasks including text classification. We use LDA as a baseline to compare the results of the NN-based classification models. Our process involves using LDA to represent each of the heavily pre-processed legal documents as a series of latent feature vectors. LDA is most commonly used to generate a collection of latent topics for a corpus, and then calculate the probability of a document belonging to a topic. We use the Gensim[5] library to create and train the LDA model. We classify vectors from the implementation of this model using LR.

The LDA Bayesian probabilistic model is an unsupervised machine learning method used to organize documents through topic modeling. In this model, each document is represented as a probability distribution over latent topics. These topics are derived from the assumption that the document's words themselves, modeled as a term frequency-inverse document frequency (tf-idf) matrix, with words represented using the bag-of-words (BoW) model, are distributed over latent topics as defined by the distribution of words in the corpus.[6]

After latent feature vectors are generated to describe each of the documents, we apply an LR to classify the documents into 15 legal issues and 279 legal subtopics. As in Wood et al. [1], we use a combination of LDA and pre-defined labels with corresponding documents to create our baseline classification system.

B. Doc2vec + Logistic Regression

Our first method of document classification using deep learning involves a higher-level application of word2vec. As described in [23], paragraph vectors, often referred to as doc2vec or document vectors, can be used to map semantic meaning from a variable-length document to a fixed-size vector. As in the word2vec learning method, a word is predicted by its neighboring words. The significant difference between the Distributed Memory Model of Paragraph Vectors and other similar learning techniques is that an additional paragraph token (treated like an additional word in the document) is used when learning the paragraph vector. Next, we classify the documents into both 15 and 279 classes using a logistic regression.

C. Bag-of-Words + Support Vector Machine

As another baseline, we represent documents using the BoW model and apply an SVM using Scikit-learn's SVM package[7] with default parameters. We chose SVM as a baseline because it is one of the highest performing traditional ML methods,

[5]https://radimrehurek.com/gensim/

[6]In order to find the ideal number of topics for the LDA-based classification, we conducted the experiment with different numbers of topics ranging from 100 to 600 in steps of 100 and and chose 300 topics because there was not a noticeable improvement in performance with more than 300 topics.

[7]http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

used for lots of different tasks. Similarly, Kim [20] also uses an SVM benchmark.

D. Word2vec + CNN

For our neural network classification approach, we designed a multi-layer model similar to the one described by Kim [20], but with additional layers and modified hyperparameters.

Our model first creates an embedding layer using pre-trained word2vec word embeddings, and then creates a matrix of documents represented by 300-dimensional word embeddings. We include three sets of the following: a dropout of 0.25, a convolution layer of 128 filters with a filter size of 5, and max pooling layer with a pooling size of 5. We also add a dense layer consisting of 128 units between two dropouts of 0.5 to prevent overfitting. The last layer is a dense layer with size equal to the number of labels for the test (15 or 279).

E. Other Embeddings

In addition to using word2vec embeddings with the CNN, we conduct the same experiments with two other pre-trained word embeddings: fastText vectors[8] from Facebook AI Research (FAIR) [15] and GloVe vectors[9] from Pennington et al. [16]. The pre-trained 300-dimensional fastText vectors are trained on Wikipedia using the skip-gram model described in Bojanowski et al. [15]. The pre-trained 300-dimensional GloVe vectors are trained on Wikipedia and the Gigaword 5 dataset using GloVe model [16].

We use the publicly available pre-trained word vectors trained on about 100 billion words from part of the Google News dataset.[10] We use 300-dimensional word2vec vectors trained using the skip-gram architecture with negative sampling by Mikolov et al. [14].

F. Word2vec + LSTM

One of the RNN-based networks we used to classify our legal corpus is the LSTM, which is defined by these equations:

$$\mathbf{i}_t = \sigma(\mathbf{x}_t\mathbf{U}^i + \mathbf{h}_{t-1}\mathbf{W}^i + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{x}_t\mathbf{U}^f + \mathbf{h}_{t-1}\mathbf{W}^f + \mathbf{b}_f) \quad (2)$$

$$\mathbf{o}_t = \sigma(\mathbf{x}_t\mathbf{U}^o + \mathbf{h}_{t-1}\mathbf{W}^o + \mathbf{b}_o) \quad (3)$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(\mathbf{x}_t\mathbf{U}^c + \mathbf{h}_{t-1}\mathbf{W}^c + \mathbf{b}_c) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t) \quad (5)$$

In this model, $\mathbf{x}_t$ represents an input x at time t. The three gates of the LSTM are the input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$. $\mathbf{c}_t$ is the memory cell state. $\mathbf{h}_t$ is the hidden state. The input weights are defined by $\mathbf{W}$, recurrent weights by $\mathbf{U}$, and bias by $\mathbf{b}$.

Our implementation of the LSTM consisted of the embedding layer formed with pre-trained word2vec vectors, an LSTM layer consisting of 128 units, and a dropout of 0.5 to prevent overfitting. Lastly, we had a dense layer representing the number of labels for the experiment.

G. Word2vec + GRU

Our last experiment involves the memory-enhanced GRU [10], a variation of the RNN. The GRU is described by the following equations:

$$\mathbf{z} = \sigma(\mathbf{x}_t\mathbf{U}^z + \mathbf{h}_{t-1}\mathbf{W}^z) \quad (6)$$

$$\mathbf{r} = \sigma(\mathbf{x}_t\mathbf{U}^r + \mathbf{h}_{t-1}\mathbf{W}^r) \quad (7)$$

$$\mathbf{s}_t = tanh(\mathbf{x}_t\mathbf{U}^s + (\mathbf{h}_{t-1} \circ r)\mathbf{W}^s) \quad (8)$$

$$\mathbf{h}_t = (1 - \mathbf{z}) \circ \mathbf{s}_t + \mathbf{z} \circ \mathbf{h}_{t-1} \quad (9)$$

where $\mathbf{x}_t$ represents an input vector x at time t, $\mathbf{r}$ is the reset gate, and $\mathbf{z}$ is the update gate. The input weights are defined by $\mathbf{W}$ and recurrent weights by $\mathbf{U}$.

As with our CNN and LSTM models, our application of the GRU begins with a word2vec embedding layer. Next, we include a GRU layer of size 128 and a dropout of 0.5 before the final dense layer for classification.

H. Hyperparameters and Regularization

In our experiments, we tested our model with a series of different hyperparameters and found that our best NN systems use 128 units for the RNNs and 128 filters for each of the convolutional layers in the CNN. For both these settings, we tried values of 32, 64, 128 and 256 and 128 gave the best results. Basically, the 128 gave better results than the lower settings and it turned out that the 256 setting could not be run effectively when training with an Nvidia GPU. It seems that additional GPU memory would be required (or a more efficient algorithm) to use 256 units. It is probable that 128 is simply the largest (power of 2) setting that is practical to use given the available equipment. This seems to be supported by the fact that many other NN systems (e.g. Kim [20]) use a value around 100.

Additionally, each of the models are regularized with a dropout [24], which works by "dropping out" a proportion p of hidden units during training. We found that a dropout of 0.5 before the final dense layer and batch size of 32 worked best for the LSTM, GRU, and CNN. We also found that the Adam optimizer [25] worked best for both the for CNN and RNN networks.

V. Results

Our goal is to determine the best method for applying automated document classification to legal texts with the hopes of facilitating legal experts in their classification of court documents. Our experiments look not only at comparing existing networks, but also at developing our own superior network. As shown in Figure I, our CNN model achieves the

[8]https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[9]https://nlp.stanford.edu/projects/glove/

[10]https://code.google.com/archive/p/word2vec/

highest accuracy, both for the 15-label and 279-label tasks (72.4% and 31.9% accuracies, respectively). We present an analysis of our results.

TABLE I: Classification Accuracy Results on the Test Set

Model	15 labels	279 labels
Word2vec + CNN	**72.4**	**31.9**
fastText + CNN	67.3	25.1
GloVe + CNN	67.1	17.7
Word2vec + GRU	68.6	14.5
Word2vec + LSTM	43.8	6.5
Word2vec + CNN (Kim [20])	65.9	14.7
Word2vec + CNN (Hughes et al. [7])	54.7	19.8
LDA + LogR [21]	40.3	13.4
Doc2vec + LogR [23]	54.1	28.6
BoW + SVM	64.0	30.5

The accuracy results of correctly classifying documents with our CNN, LSTM, and GRU models compared to other classification methods. Our CNN best overcomes the problem of an uneven distribution of classes, as shown in Figures 1 and 2.

TABLE II: CNN Results on the Development Set by Category

Label	Precision	Recall	F-Score	# of Docs
0 - None	0.33	0.33	0.33	3
1 - Criminal Procedure	0.82	0.91	0.86	182
2 - Civil Rights	0.72	0.70	0.71	138
3 - First Amendment	0.87	0.79	0.82	84
4 - Due Process	0.45	0.33	0.38	30
5 - Privacy	0.56	0.62	0.59	8
6 - Attorneys	0.33	0.40	0.36	5
7 - Unions	0.73	0.76	0.75	29
8 - Economic Activity	0.72	0.72	0.72	172
9 - Judicial Power	0.57	0.56	0.56	116
10 - Federalism	0.41	0.41	0.41	34
11 - Interstate Relations	0.67	0.67	0.67	6
12 - Federal Taxation	0.90	0.79	0.84	33
13 - Miscellaneous	0.50	0.50	0.50	2
14 - Private Action	0.00	0.00	0.00	0
Avg/Total	0.71	0.72	0.71	842

The relation between frequency and f-measure for the development set.

TABLE III: CNN Results on the Test Set by Category

Label	Precision	Recall	F-Score	# of Docs
0 - None	0.20	1.00	0.33	1
1 - Criminal Procedure	0.81	0.85	0.83	183
2 - Civil Rights	0.77	0.81	0.79	121
3 - First Amendment	0.79	0.88	0.83	56
4 - Due Process	0.48	0.30	0.37	33
5 - Privacy	0.57	0.44	0.50	9
6 - Attorneys	0.80	0.73	0.76	11
7 - Unions	0.77	0.70	0.73	33
8 - Economic Activity	0.72	0.74	0.73	145
9 - Judicial Power	0.56	0.54	0.55	102
10 - Federalism	0.48	0.33	0.39	33
11 - Interstate Relations	0.50	0.40	0.44	5
12 - Federal Taxation	0.86	0.96	0.91	25
13 - Miscellaneous	0.00	0.00	0.00	1
14 - Private Action	0.00	0.00	0.00	0
Avg/Total	0.72	0.72	0.72	758

The relation between frequency and f-measure for the test set.

Tables II and III show the CNN's (our best system) performance on individual classes. Although the details are slightly different (e.g., a different number of documents for each category), the relative scores are about the same. We now do a more detailed analysis on the development set results, rather than the test set because we do not want to examine the test results too closely and bias our future work. It is clear that the model's accuracy tends to be higher for the most frequent categories. Categories 1, 2, 3, 8 and 9, all of which are labels on more than 50 documents, mostly have f-measures of over 70%, with one outlier at 56%. It is difficult to generalize about the least frequent categories (frequency < 10), including labels 0, 5, 6, 11, 13 and 14, as there is too little data to analyze. Some of these have f-measures of 0 or near 0, and on average, they do much worse than the high-frequency categories. This is expected since the high-frequency categories have more training data and thus provide more evidence for the model to build on. Thus, as expected, cases with correct labels of 1, 2, 3, 8 and 9 tend to be classified correctly and the categories with little to no training data (0, 5, 6, 11, 13, 14) are most often misclassified.

On the other hand, category labels 4, 7, 10, and 12 each have a similar moderate number of test documents (around 30 documents), but have very different results: the model achieves relatively high results for labels 7 and 12 and relatively poor results for 4 and 10. Thus it would seem that the results for labels 4, 7, 10 and 12 cannot be explained purely on the basis of frequency. Figure 3 is a confusion matrix for our CNN results on the development/validation set. We observe some patterns which may help us understand these results. For labels 7 and 12, where the model performs well, the correct category clearly dominates–no other category is marked for more than 4 documents. However, the poorly performing categories, each have a second (or third) dominant category in addition to the correct one. Label 4 (*Due Process*) is applied to 10 true *Due Process* cases and incorrectly classifies 7 as *Civil Rights* cases, 6 as *Economic Activity* cases and 4 as *Criminal Procedure* cases and another 3 miscellaneous erroneous labels. To the extent that a case may be given multiple classifications (*Due Process* and *Civil Rights*) or (*Due Process* and *Economic Activity*), these errors are understandable and may even reflect a defect in the experiment–perhaps cases should have multiple classifications and the one classification per case assumption is unrealistic. Similarly, label 10 (*Federalism*) is applied 14 times correctly to *Federalism* cases and 11 times incorrectly to *Economic Activity* cases (and rarely to other categories). It is expected that some federalist issues (issues about the power of the federal government) will overlap with economic issues. So these may also be understandable errors.

We now attempt to better understand these errors, focusing our error analysis on *Federalism* classification. We examine three samples from our development set, each sample consisted of four cases. We look at 4 cases that are correctly classified by our CNN as *Federalism* cases (True_Fed); 4 cases that were correctly classified as *Economic Activity* cases (True_Eco); and 4 *Federalism* cases that our system misclassified as *Economic Activity* cases (False_Fed). We compare the False_Fed cases to both True_Fed and True_Eco. Our goal is to understand the sort of factors that might cause a human or a machine learning algorithm to mis-classify the False_Fed

documents.

The True_Fed cases include *Testa et al. v. Katt*; *Bethlehem Steel Co. et al. v. New York State Labor Relations Board*; *Rice et al. v. Santa Fe Elevator Corp. et al.*; and *Rice et al. v. Board of Trade of the City of Chicago*. These all involved the interaction of state and federal authorities and laws, including questions of whether a state authority should be compelled to enforce a federal law or whether a state agency/law takes precedence over a federal agency/law.

The 4 True_Eco cases included *Halliburton Oil Well Cementing Co. v. Walker et al.*; *Champlin Refining Co. v. United States et al.*; *United States v. Howard P. Foley Co., Inc.*; and *Richfield Oil Corp. v. State Board of Equalization*. The Halliburton case is a patent dispute. The Foley case is about the government's liability in a contract dispute. The Champlin case examines whether the Interstate Commerce Commission, a federal entity, could require information from an oil refining company operating across several states. While similar to the True_Fed cases in some ways, there is no conflict between a state and a federal authority. The Richfield case is a dispute about whether a state sales tax applies to a sale to a foreign government. This seems similar to Federalist concerns, but there is no conflict between a state and federal authority. Rather the concern is whether or not a state sales tax effects a possibly-foreign transaction.

The False_Fed cases include *Phillips Chemical Co. v. Dumas Independent School District*; *Panhandle Eastern Pipe Line Co. v. Michigan Public Service Commission et al.*; *Wyeth v. Diana Levine*; and *North Dakota v. United States*. These cases all involve financial transactions, state authorities and federal authorities. The topics covered includes: the legitimacy of state taxes on federal land leased to a company; state regulation of alcohol and other goods procured for sale on a federal military base; liability of a drug company (in state civil court) for damages from harm by their drug, even though the FDA, a federal agency, granted them clearance for the drug; and whether the sale of natural gas was subject to state regulation, in spite of a federal law licensing the sale. While some of these issues seem to include federal/state authority conflicts, those conflicts are not as clearly articulated as in the True_Fed cases. So it is clear how experienced annotators may be able to consistently distinguish the *Federalism* and *Economic Activity* classes. However, we would imagine that inexperienced annotators may have trouble.[11] Similarly, machine learning may require more evidence (more documents) to correctly classify these cases.

The sparsity and imbalanced classes of the dataset presented itself as the most challenging obstacle for training the neural networks. For instance, nearly three fourths of all documents fell under 4 of the 15 legal area categories (*Criminal Procedure*, *Civil Rights*, *Economic Activity*, and *Judicial Power*). Not only was there not an even distribution of documents over each of the labels, many of the classes had little to no training data. Furthermore, our input sequence length was several

[11] We would expect lower inter-annotator agreement on these sort of cases.

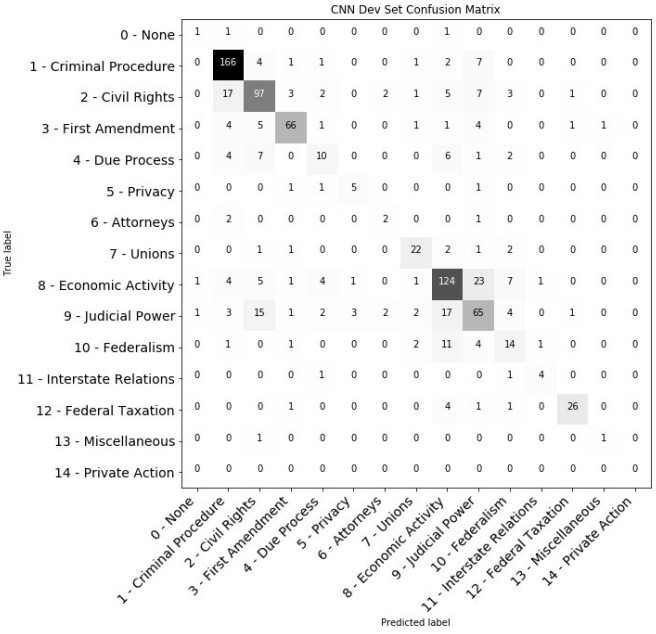


Fig. 3: Confusion Matrix for CNN Development Corpus

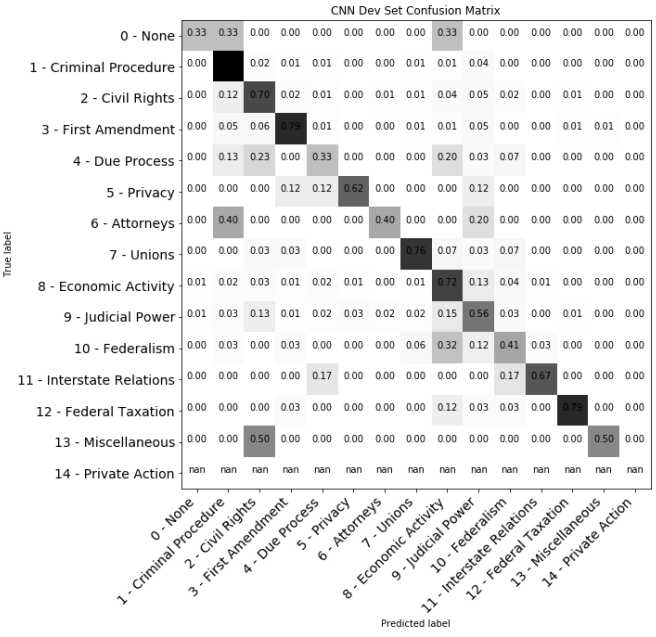


Fig. 4: Normalized Confusion Matrix for CNN Development Corpus

orders of magnitude larger than the inputs in experiments conducted by Kim [20] and Hughes et al. [7].

Due to adjustments we made for the dataset and configuration changes (see Section IV), our CNN performs better than other CNN models (see Table I). The adjusted parameters include dropouts to account for overfitting that occurs earlier in training and the addition of extra layers.

After generating a model for each of our NN-based tests, we use them for our entire corpus and analyze the documents that are misclassified in order to find patterns in the way each of NN architectures complete the classification task. Figure 5 shows a combination of the normalized confusion matrices resulting from classifying our entire corpus with our CNN and the simple RNN models, and each of the true label rows describe the fraction of documents classified per predicted label. As in Figure 5, the CNN performs the best for each of the categories, not just the top four labels (although a small number of errors made by the CNN were with documents from these classes). Unlike our CNN, the classifications from the GRU (our second best system) are more scattered. Figure 5 shows some of labels the GRU most frequently misclassifies. In contrast to the CNN and LSTM, the GRU tends to significantly mis-classify documents from both frequent and infrequent categories, as in the frequent category of *Criminal Procedure* and the uncommon category of *First Amendment*. The GRU also does not classify entire categories correctly, whereas the CNN had high classification accuracy for every category. For example, the GRU mis-classifies every legal opinion in the categories: *Interstate Relations*, *Miscellaneous*, and *Private Action*. Additionally, there was no single label L, such that our GRU system correctly classified more documents in class L than our CNN system.

It seems that the relatively low frequencies of some of the categories is more of a challenge to some of the learning algorithms than others. In particular, CNN and GRU appear to be somewhat more resilient to this effect, than LSTM, as evidence by the merged confusion matrices shown in Figure 5. The LSTM incorrectly classified a majority of the documents to one of the two most frequent labels (*Criminal Procedure* and *Economic Activity*), as shown in Figure 5. Despite categorizing almost all of the documents to only two legal issues, the LSTM did not have a higher accuracy for those two labels. In fact, the LSTM did worse than our LDA baseline system (see Table I). This was somewhat surprising considering that LSTMs often perform similarly to GRUs for a given task. We plan to investigate this further in future research, possibly trying additional models.

As in Chung et al. [11], we test the performance of LSTMs and GRUs. While we apply these models to categorizing legal text, their application was music transcription. Our results show that the structure of the simpler GRU leads to greater accuracy compared to the LSTM when classifying a relatively small number of documents over a large number of labels. The GRU performs better than the LSTM with document-length sequences. The LSTM tends to remember the wrong information needed for the classification because of the small

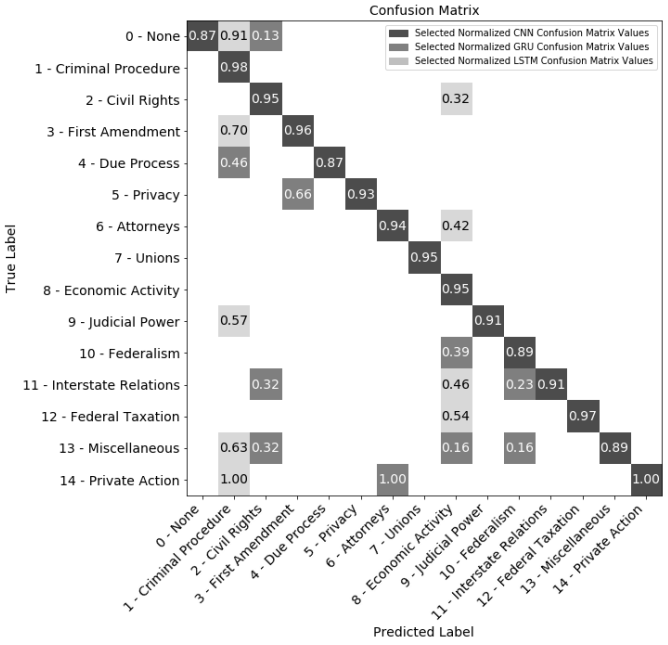


Fig. 5: Merged Confusion Matrix for CNN, GRU, and LSTM

size of the dataset.

Our results also show that the simple BoW+SVM model performs very well for both classification tasks. As Nallapati and Manning [17] mention, "the SVM assigns high weights to many spurious features owing to their strong correlation with the class" [17, p. 442]. In other words, even very infrequent words that have very high correlations with certain classes would help the SVM associate certain words with uncommon classes. This aspect of the SVM seems to explain why the SVM performed well with the 279-label classification task (nearly as well as the best system), in which only a few documents define each category.

In our results, the word2vec model out-performs the simpler bag-of-words model. With more statistical information, the classifiers find common features and patterns to describe categories with higher accuracy. The positive results from our experiment in which we apply an LR to paragraph vectors (doc2vec) show how well the embeddings capture the meaning of the documents (refer to Table I).

The simple GRU network has promising results compared to the CNNs because the GRU is designed to handle long sequences. While a word may carry a large weight in most contexts, the GRU allows for a word's weight to diminish based on specific examples. Yin et al. [19] shows that the accuracy of the CNN decreases as sequence length increases and eventually falls under the accuracy of the GRU. Since our experiments involve inputting entire documents instead of sentences, sequence lengths are orders of magnitude larger than than those used in the experiments conducted by Yin et al. [19], Kim [20], and Hughes et al. [7].

VI. Conclusion and Future Work

In this paper, we find the best method for automated legal document classification is the SCC system that uses a CNN (72.4% accuracy for 15 general categories and 31.9% accuracy for the 279 more specific categories). On the other hand, the GRU architecture shows promising results compared to our tuned CNN (nearly as high for the 15 category task). We believe that a tuned GRU-based network can potentially complete the task with higher accuracy.

The SCC system uses word embeddings from a general domain (Google News). It is possible that embeddings from the legal domain would improve results. Accordingly, we plan to compile a much larger corpus of US legal opinions from appellate and local courts in order to generate domain-specific word embeddings for our model. We will conduct experiments using these embeddings instead of the Google News embeddings used for the results reported here, or possibly in addition.

In future work, we plan to investigate the reasons behind the substantial difference between the performance of the LSTM and GRU. Moreover, we believe that an application of transfer learning as shown in [26] could be used to train classifiers for more specific topics and different subdomains of the legal field.

References

[1] J. Wood, "Source-lda: Enhancing probabilistic topic models using prior knowledge sources," *CoRR*, vol. abs/1606.00577, 2016. [Online]. Available: http://arxiv.org/abs/1606.00577

[2] H. J. Spaeth, L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh, "2017 supreme court database, version 2017 release 01," 2017. [Online]. Available: http://supremecourtdatabase.org

[3] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215–242, 1958.

[4] O. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. van Genabith, "Exploring the use of text classification in the legal domain," *CoRR*, vol. abs/1710.09306, 2017. [Online]. Available: http://arxiv.org/abs/1710.09306

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[6] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[7] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classification using convolutional neural networks," *CoRR*, vol. abs/1704.06841, 2017. [Online]. Available: http://arxiv.org/abs/1704.06841

[8] A. Conneau, H. Schwenk, L. Barrault, and Y. LeCun, "Very deep convolutional networks for natural language processing," *CoRR*, vol. abs/1606.01781, 2016. [Online]. Available: http://arxiv.org/abs/1606.01781

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[10] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014. [Online]. Available: http://arxiv.org/abs/1409.1259

[11] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

[13] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, Mar. 2002. [Online]. Available: http://doi.acm.org/10.1145/505282.505283

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[15] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *In EMNLP*, 2014.

[17] R. Nallapati and C. D. Manning, "Legal docket-entry classification: Where machine learning stumbles," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 438–446. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613771

[18] W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," in *BMC Medical Informatics and Decision Making*, 2017.

[19] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *CoRR*, vol. abs/1702.01923, 2017. [Online]. Available: http://arxiv.org/abs/1702.01923

[20] Y. Kim, "Convolutional neural networks for sentence classification," *CoRR*, vol. abs/1408.5882, 2014. [Online]. Available: http://arxiv.org/abs/1408.5882

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[22] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: http://doi.acm.org/10.1145/2133806.2133826

[23] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014. [Online]. Available: http://arxiv.org/abs/1405.4053

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[26] C. B. Do and A. Y. Ng, "Transfer learning for text classification," in *Proceedings of the 18th International Conference on Neural Information Processing Systems*, ser. NIPS'05. Cambridge, MA, USA: MIT Press, 2005, pp. 299–306. [Online]. Available: http://dl.acm.org/citation.cfm?id=2976248.2976286

11th International Symposium on Multimedia Applications and Processing

SOFTWARE Engineering Department, Faculty of Automation, Computers and Electronics, University of Craiova, Romania "Multimedia Applications Development" Research Centre

Background and Goals

Multimedia information has become ubiquitous on the web, creating new challenges for indexing, access, search and retrieval. Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices—such as laptops, iPods, personal digital assistants (PDA), and cellular telephones—have stimulated the development of intelligent pervasive multimedia applications. These key technologies are creating a multimedia revolution that will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain, especially when it comes to efficiently indexing, mining, querying, searching, retrieving, displaying and interacting with multimedia data.

The Multimedia—Processing and Applications 2018 (MMAP 2018) Symposium addresses several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, entertainment, education) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

As a result the conference goal is to bring together researchers, engineers, developers and practitioners in order to communicate their newest and original contributions. The key objective of the MMAP conference is to gather results from academia and industry partners working in all subfields of multimedia: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications—including middleware—along with the related security issues, such as secure multimedia information sharing. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages offered by multimedia computing techniques, including home-networked entertainment and games. However, innovative contributions that don't exactly fit into these areas will also be considered because they might be of benefit to conference attendees.

Call for Papers

MMAP 2017 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAP 2016 Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP 2016 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference. Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

Topics

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Big Data Science and Multimedia Systems
- Cloud Computing and Multimedia Applications
- Machine Learning, Information Retrieval in Multimedia Applications
- Data Mining, Warehousing and Knowledge Extractio
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Multimedia in Internet and Web Based Systems
- E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications and Computational biology
- Entertainment, Personalized Systems and Games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture and Fuzzy Logic Systems
- Intelligent Multimedia Network Applications
- Future Trends in Computing System Technologies and Applications
- Trends in Processing Multimedia Information
- Multimedia Ontology and Perception for Multimedia Users

BEST PAPER AWARD

A best paper award will be made for work of high quality presented at the MMAP Symposium. The technical committee in conjunction with the organizing/steering committee will decide on the qualifying papers. Award comprises a certificate for the authors and will be announced on time of conference.

STEERING COMMITTEE

- **Amy Neustein, Boston University,** USA, Editor of Speech Technology
- **Lakhmi C. Jain, University of South Australia and University of Canberra,** Australia
- **Zurada, Jacek,** University of Louisville, United States
- **Ioannis Pitas, University of Thessaloniki,** Greece
- **Costin Badica, University of Craiova,** Romania
- **Borko Furht, Florida Atlantic University,** USA
- **Harald Kosch, University of Passau,** Germany
- **Vladimir Uskov, Bradley University,** USA
- **Thomas M. Deserno, Aachen University,** Germany

HONORARY CHAIR

- **Dumitru Dan Burdescu, University of Craiova,** Romania

GENERAL CO-CHAIRS

- **Adriana Schiopoiu Burlea, University of Craiova,** Romania
- **Marius Brezovan, University of Craiova,** Romania

PUBLICITY CHAIR

- **Amelia Badica, University of Craiova,** Romania
- **Milan Simic, RMIT University,** School of Engineering, Australia

ORGANIZING

- **Dumitru Dan Burdescu, University of Craiova,** Romania
- **Costin Badica, University of Craiova,** Romania
- **Marius Brezovan, University of Craiova,** Romania
- **Adriana Schiopoiu Burlea, University of Craiova,** Romania
- **Liana Stanescu, University of Craiova,** Romania
- **Cristian Marian Mihaescu, University of Craiova,** Romania

PROGRAM COMMITTEE

- **Azevedo, Ana,** CEOS.PP-ISCAP/IPP, Portugal
- **Badica, Amelia,** University of Craiova, Romania
- **Burlea Schiopoiu, Adriana,** University of Craiova, Romania
- **Cano, Alberto,** Virginia Commonwealth University
- **Cordeiro, Jose,** EST Setúbal/I.P.S.
- **Cretu, Vladimir,** Politehnica University of Timisoara, Romania
- **Debono, Carl James,** University of Malta, Malta
- **Fabijańska, Anna,** Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland
- **Fomichov, Vladimir,** National Research University Higher School of Economics, Moscow, Russia., Russia
- **Giurca, Adrian,** Brandenburg University of Technology, Germany
- **Grosu, Daniel,** Wayne State University, United States
- **Kabranov, Ognian,** Cisco Systems, United States
- **Keswani, Dr. Bright,** Suresh Gyan Vihar University, Mahal, Jagatpura, Jaipur
- **Korzhik, Valery,** State University of Telecommunications, Russia
- **Kostagiolas, Petros,** School of Information Science and Informatics, Ionian University
- **Kotenko, Igor,** St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia
- **Logofatu, Bogdan,** University of Bucharest, Romania
- **Mangioni, Giuseppe,** DIEEI - University of Catania, Italy
- **Marghitu, Daniela,** Auburn University
- **Mihaescu, Cristian,** University of Craiova, Reunion
- **Mocanu, Mihai,** University of Craiova, Romania
- **Murawski, Krzysztof,** Faculty of Cybernetics, Military University of Technology, Poland
- **MURAWSKI, Krzysztof,** Military University of Technology, Poland
- **Ohzeki, Kazuo,** Professor Emeritus at Shibaura Institute of Technology, Japan
- **Pohl, Daniel,** Intel, Germany
- **Popescu, Dan,** CSIRO, Sydney, Australia, Australia
- **Popescu, Daniela E.,** Integrated IT Management Service, University of Oradea
- **Querini, Marco,** Department of Civil Engineering and Computer Science Engineering
- **Radulescu, Florin,** University "Politehnica" of Bucharest
- **RUTKAUSKIENE, Danguole,** Kaunas University of Technology
- **Salem, Abdel-Badeeh M.,** Ain Shams University, Egypt
- **Sari, Riri Fitri,** University of Indonesia, Indonesia
- **Sousa Pinto, Agostinho,** Instituto Politécnico do Porto
- **Stoicu-Tivadar, Vasile,** University Politehnica Timisoara
- **Trausan-Matu, Stefan,** Politehnica University of Bucharest, Romania
- **Trzcielinski, Stefan,** Poznan University of Technology, Poland
- **Tsihrintzis, George,** University of Piraeus, Greece
- **Tudoroiu, Nicolae,** John Abbott College, Canada
- **Vega-Rodríguez, Miguel A.,** University of Extremadura, Spain
- **Virvou, Maria,** University of Piraeus, Greece
- **Watanabe, Toyohide,** University of Nagoya
- **Woźniak, Marcin,** Institute of Mathematics, Silesian University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 527–534
DOI: 10.15439/2018F139
ISSN 2300-5963 ACSIS, Vol. 15

Immersive Virtual Reality for Earth Sciences

Ilario Gabriele Gerloni, Vincenza Carchiolo
Department of Electrical, Electronic and Computer Engineering, University of Catania, Italy
Email: ilario.gerloni@gmail.com

Fabio Roberto Vitello, Eva Sciacca, Ugo Becciani, Alessandro Costa, Simone Riggi
Astrophysical Observatory of Catania, Italian National Institute for Astrophysics (INAF), Italy
Email: fabio.vitello@inaf.it

Fabio Luca Bonali, Elena Russo, Luca Fallati, Fabio Marchese, Alessandro Tibaldi
Department of Earth and Environmental Sciences, University of Milan Bicocca
Email: alessandro.tibaldi@unimib.it

***Abstract*—This paper presents a novel immersive Virtual Reality platform, named ARGO3D, tailored for improving research and teaching activities in Earth Sciences. The platform facilitates the exploration of geological environments and the assessment of geo-hazards, allowing reaching key sites of interest (some of them impossible to be reached in person) and thus to take measurements and collect data as it can be done in the real field. The target audience of ARGO3D encompasses students, teachers and early career scientists, as well as civil planning organisations and non-academics. The overall workflow for real ambient reconstruction, processing and rendering of the virtual ambient is presented, as well as a detailed description of the VR software tools and hardware devices employed.**

I. Introduction

DUE TO technological advances in instruments and detectors over the last two decades, there has been a true explosion in both the amount and the complexity of scientific data, an exponential trend that will continue at ever increasing rate in the coming years. The concept of Big Data - defined not only as datasets too large to be processed with today's tools and methods but also (and perhaps more importantly) as datasets too complex to be effectively dealt with - permeates our world. Over the last decade, hundreds of thousands measurements involving complex 3D datasets of geological structures of the active seismic zones have been collected. The development of new instruments results in increasingly high resolution images and datasets, e.g. from Unmanned Aerial Vehicle (UAV) surveys. These data offer unique opportunities to build long term capacity for geo-hazard maps and studies never realised so far.

Virtual reality (VR) allows to generate virtual environments which gives the feeling of being elsewhere, within a real place. It can be considered as an advanced form of human-computer interface that allows to interact with a realistic virtual environment generated using interactive software and hardware.

The "Virtual Reality", starting from the 60's, has evolved in different manners becoming more and more similar to the real world. Nowadays, two different kinds of VR can be identified: non-immersive and immersive. The former is a desktop-based environment that can simulate places in the real or imagined worlds [11]; the latter takes the idea even further by giving the perception of being physically present in the non-physical world. The non-immersive VR can be based on a standard computer (using e.g. mouse, keyboard or joysticks), whereas the immersive VR is still evolving as the needed devices are becoming more user friendly and economically accessible.

In the past, there was a major difficulty about using VR equipment such as a helmet with goggles, while now new devices are being developed to make usability better for the user. VR is based on three basic principles: Immersion, Interaction, and User involvement with the environment. It offers a very high potential in education [5], [1], industry [13], research [3] and scientific collaboration [10], [8] by making the learning experience more motivating and engaging, and facilitating the investigation of complex scenarios. Up to now, the use of immersive VR in research, educational games and scientific dissemination has been limited due to the high price of the devices and their non user-friendly usability. At the moment, new virtual reality headsets like the commercial *Oculus Rift*™ or smartphone headsets (e.g. Samsung Gear VR), make it possible to access immersive VR in lots of research, educational and dissemination contexts.

This paper presents the first platform for immersive VR in Earth Sciences, here named ARGO3D (Augmented and virtual reality for geology and geophysics - http://argo3d.unimib.it/). Our immersive VR platform facilitates the survey of geological environments and related geo-hazards assessment; the target audience regards early career scientists, civil planning organisations, as well as academic (e.g. students and teachers) and non academic people. Virtual 3D visualization and navigation allows players to fully explore areas of interest for geo-hazards (e.g. the crater of an active volcano), and, thanks to the data collected (and observations) by virtual exploration, to better define and interpret the geological phenomena affecting the selected site. Thus, players will be able to reach more holistic and comprehensive concepts, as the main specific features

This work is supported by MIUR Accordo di programma ACPR15T4_00098

will be viewed from many different scales and perspectives. Moreover, students exhibit conceptual difficulty in interpreting various 2D diagrams and photos, showing 3D phenomena and geological site. Incorporating 3D visualizations into teaching earth related sciences can help the students in learning activities thanks to a wider spatial perspective and thinking, and may help them in improving their spatial ability.

II. Scientific Background

The use of virtual and augmented reality is a novel and extremely innovative method of 3D immersive approach for research and teaching activities in Earth Sciences. Marine and terrestrial environments will be brought directly into laboratories and classrooms giving the researchers and students the feeling of playing a videogame, instead they are exploring real geological key sites. Geological environments are rendered by means of the Structure from Motion (SfM) Photogrammetry technique. 3D scenarios are derived from UAVs regarding terrestrial environments and from submersibles (ROV) for marine environments, with exceptionally high-detailed images.

The ARGO3D platform is tailored with three main purposes: overcoming problems in research, as well as increasing the quality of training and outreaching activities in Earth and Environmental Sciences.

Regarding the research activity, the best way to study geo-hazards in terrestrial and submarine environments (such as landslides, volcanic eruptions, erosion, floods and recent surface seismic ruptures) is to examine rock outcrops and key sites directly in the field. Due to logistic conditions, this activity is often complicated or even impossible: for example, it is not achievable to directly explore the crater of an active volcano due to toxic gases and high temperatures whereas some peculiar geological spots are too far to be reached on foot or by car. Furthermore, very high vertical outcrops are difficult to be analysed in all their spatial extension.

Virtual reality represents a new way to bypass these problems, giving to the researcher the possibility to easily reach key sites of interest and thus to take measurements and collect data, as it can be done in the field. ARGO3D will allow both students and teachers to analyse environments and processes that would be impossible to observe in person.

In regard to the training and outreach activity, teachers will lead the students into virtually-reconstructed 'hard to reach' environments, explaining them the best ways to enjoy the activity and the geological meaning of the different features that appear on the screen. The development of useful teaching and training tools will allow each student to explore in real time virtually-reconstructed 3D environments, to take pictures of significant geological features, to collect basic geological data, as well as to virtually fly above the area and to export features to a GIS environment for further analysis. At the same time, the other students/people will have a chance to observe, just like in a 3D cinema/movie, how their colleagues navigate and interact with the environment. These tools will improve the students observation and field mapping skills, and their knowledge and understanding of areas subjected to natural hazards.

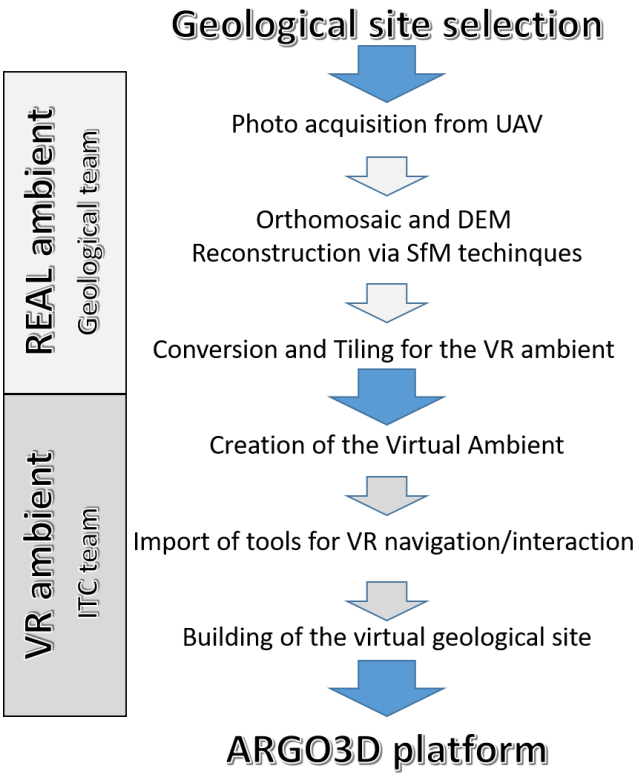


Fig. 1. General workflow describing the main steps involved for performing geological site reconstruction, processing and rendering from the real ambient into the VR ambient.

III. Data Collection and Processing

The overall workflow describing the main steps involved for performing geological site reconstruction, processing and rendering from the real ambient into the VR ambient is shown in Figure 1. Two teams are currently working closely to virtualize the geological sites of interests: geologists and ITC specialists. Geologists are in charge of performing the first three steps that are described in this section.

The input files for the creation of 3D reconstructed environments (in terms of terrain and texture) have been generated thanks to the Structure from Motion algorithms (SfM) applied to Unmanned Aerial Vehicle (UAV or drone)-captured images (e.g. [4]). Such approach involves three main processes (see the first three steps of Figure 1): i) acquisition of a set of images collected in order to cover the selected geological site (with an overlap greater than 60%); ii) processing of these images through a SfM software that leads to the creation of a orthomosaic and a digital elevation model (DEM) as final products; iii) conversion and tiling of the orthomosaic and DEM in a format compatible with Unity.

For the image acquisition we used a quadcopter, DJI Phantom 4 (Fig. 2A), one of the most used and versatile UAVs. They have been collected along a defined flight path (50 m of

altitude) in order to cover the whole selected geological key site with an overlap of 90% (Fig. 2B). In order to accurately georeference the resulting model, in addition to the GPS photo location from the drone, real-world coordinates of, at least, four Ground Control Points (GCPs) have been established within the surveyed area (e.g. [7]; [14]; [15]).

Regarding the SfM processing, UAV-captured photos have been processed with the use of Agisoft PhotoScan [1], a SfM software. The SfM technique allowed to identify matching features in different images, collected along a defined fly path, and combine them to create a sparse and dense cloud (Fig. 2B), a mesh, an orthomosaic and a DEM (Fig. 2C-D) (further details in [12] and [15]). Our model is made of 20722 x 20722 pixels, with a corresponding pixel size of 0.02 m (20 mm), the areal extension in an E-W and N-S direction is equal to 414.44 m. Both the orthomosaic and DEM are in GeoTiff file format.

GeoTIFF file format is widely used for raster imagery and aerial photography since it contains georeferencing information, necessary to establish the exact spatial reference of the file, such as the datum and UTM zone; in our case, we used WGS84 datum and the UTM zone is 28N. The DEM results as a GeoTiff with each pixel associated with altitude value of the real ambient (Fig. 2D), whereas the resulting orthomosaic is composed by pixels each representing the RGB color of the real ambient (Fig. 2C). Both of them have been exported with the same areal extension and pixel size (20 mm/pixel); NULL values in the DEM and orthomosaic have been filled by interpolation (Fig. 2D) and RGB pixel colour from satellite images (50 cm/pixel - image from Worldview-2 sensor, Catalogue ID 1030050051F92D00) (Fig. 2C), respectively.

In order to work as input file for Unity, the DEM must be converted in gray scale 16 bit RAW file (which is a format compatible with most image and landscape editors) with minimum value corresponding to 0 and maximum value to 65536, corresponding to a range of 14.029 m in the real world. Due to the high resolution, in terms of pixel size, of our model, we chose to implement this conversion considering altitude and areal extension of the model in mm. After that, we produced tiles of 512x512 pixels (required to support mobile devices) for the orthomosaic (Fig. 2E) and 513x513 pixels for the RAW file representing the DEM (Fig. 2F). The additional bit along x and y represents an overlap for a better merging of the tiles in Unity environment. In case of the selected key site, it results in 41x41 tiles, already flipped/rotated for the best import in Unity, each one with an extension of 10240x10240 mm.

Each tile is finally associated with a descriptive text file, which adds information about the real environment depicted in the data previously mentioned. The essential information these files must contain are: longitude ("XWorldLimits" keyword contains leftmost one and rightmost one for the current tile), latitude ("YWorldLimits" keyword contains lowest one and highest one for the current tile) and elevation ("MinAlt" keyword is the minimum and "MaxAlt" keyword is the maximum for the current tile). These text files are formatted as JSON strings, used to calculate the position of the player in the real world coordinates.

An example of metadata (Fig. 2D) is provided below:

```
{
"XWorldLimits": [406161.706, 406576.186],
"YWorldLimits": [7316243.990, 7316658.430
    ],
"MaxAlt": 291.108,
"MinAlt": 277.079
}
```

IV. ARGO3D PLATFORM

A. General Description

In order to experience the software properly, the player must wear the VR headset and hold the input peripherals before running the application. At this stage, he also must sit in front of the motion sensor, so that all the startup settings, like default position of the user and peripherals in real world space, are completed.

When the virtual environment is loaded, the user can move and rotate his head to look at the computer-generated objects around himself. As said by Palmer Luckey in [2], there are various infrared lights on the headset, that are used by the sensor to track the position and rotation of the user. When he moves his head, the software rotates the internal camera in order to simulate an immersive well-rounded vision of the environment.

To navigate in the scene, the player can use left hand "Thumbstick" placed on one of the provided input peripherals. These inputs are one for each hand, so multiple functionality are available for various scenarios, such as changing view and velocity through the scene (see subsection IV-B). These must be in the range of sensor placed in front of the peripherals, in order to be localized and rendered in the virtual scene.

B. Use Cases

The player has currently three possible modes to navigate in the scene, called "Walk Mode", "Flight Mode" and "Drone Mode" (Fig. 3). He can switch between these modes in every moment.

When the user is in "Walk Mode" (Fig. 3A) he behaves like a moving rigid body, so he's subject to gravity and can navigate the scene stick to the ground, or doing little jumps, thus experiencing the scene in realistic and immersive way. He can walk in a certain direction and look around thanks to the VR tools and, also, regulate his height with two special keys. This is the default mode, the one the user is in at the beginning.

In "Flight Mode" (Fig. 3B) the user can watch details that can't be observed or reached otherwise, and, at the same time, has a clear view from above of the entire scene. In

[1]Agisoft PhotoScan: http://www.agisoft.com/

[2]https://www.vrfocus.com/2015/06/palmer-luckey-explains-oculus-rifts-constellation-tracking-and-fabric/

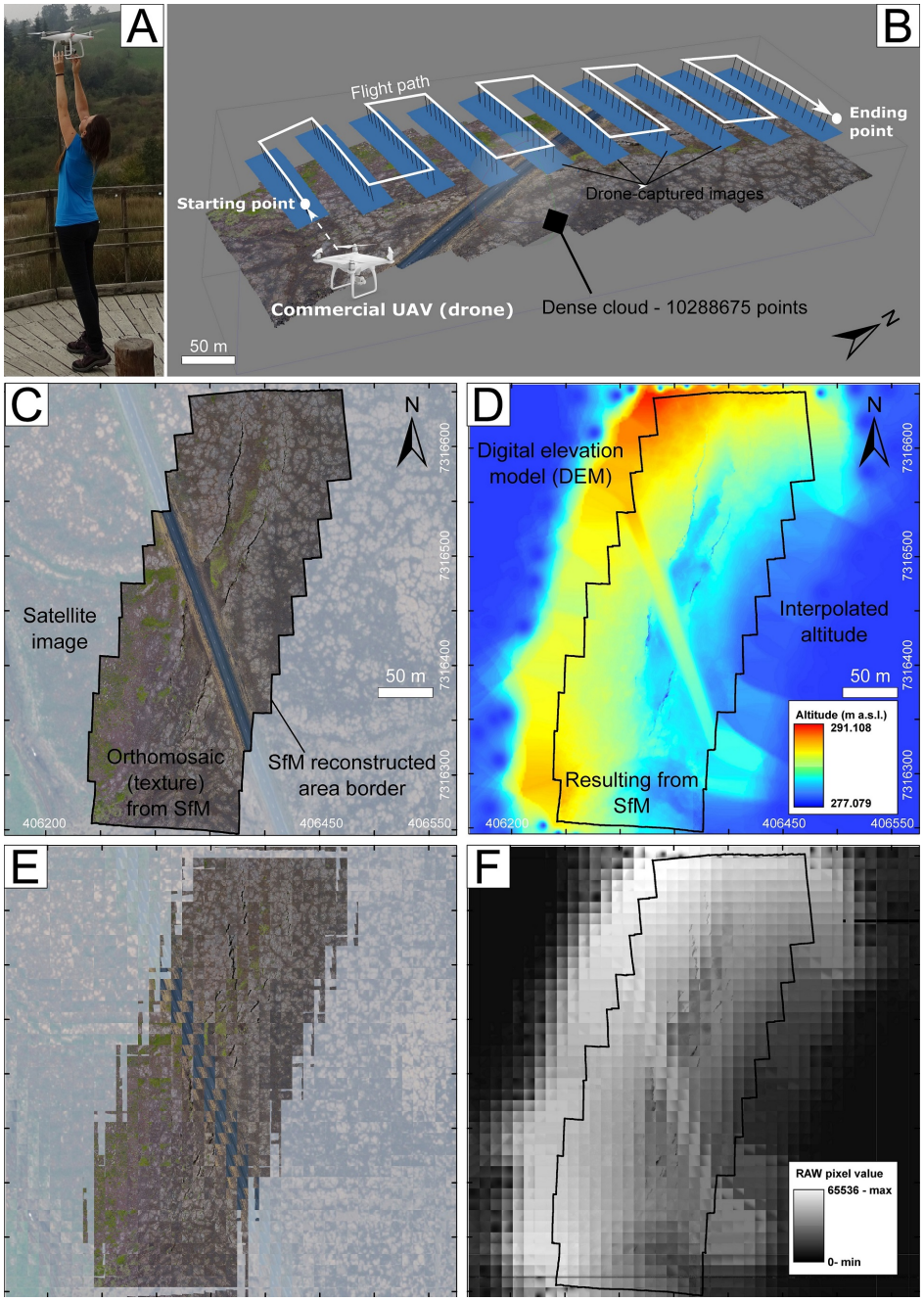


Fig. 2. Main steps involved in 3D reconstruction and processing for the input of the ARGO3D VR platform. (A) UAV (DJI Phantom 4) used for photo collection, person for scale. (B) 3D representation of the flight path, drone-captured photos (with location and orientation) and the dense cloud (resulting from SfM approach). (C) Orthomosaic resulting from the SfM processing (within the black border) characterized by the high resolution of 0.02 m/pixel. Outside the black border, pixels are from high-resolution satellite image (50 cm/pixel - image from Worldview-2 sensor, Catalogue ID 1030050051F92D00). The whole area is scaled at the best pixel resolution (0.02 m/pixel). (D) DEM resulting from SfM processing (within the black border) characterized by the high resolution of 0.02 m/pixel. Outside the black border, DEM values result from IDW interpolation. The whole DEM is scaled at the best pixel resolution (0.02 m/pixel). (E) Tiled texture and DEM (F) ready to be imported in Unity. The tiled texture is in JPG format, DEM values have been converted in Grayscale 16 bit RAW format.

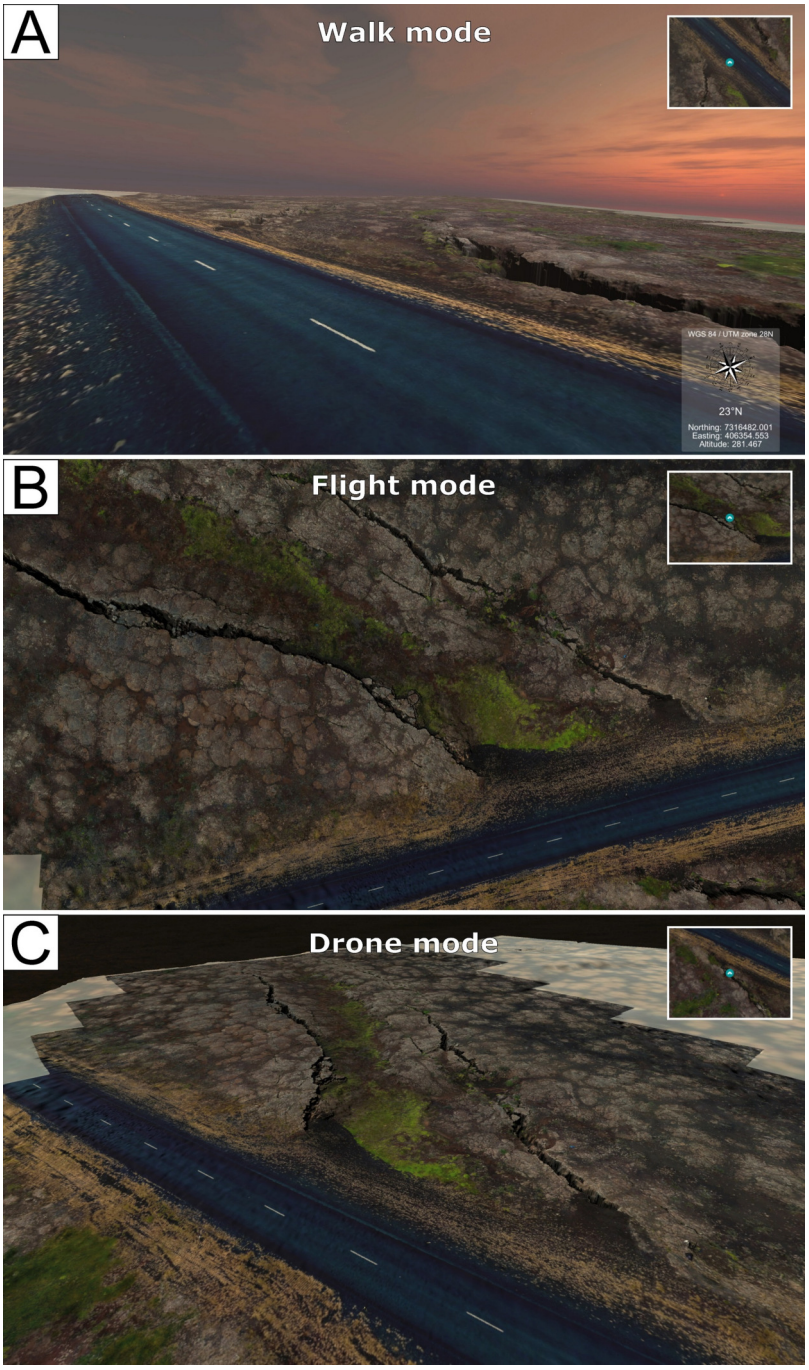


Fig. 3. Three modes of exploration are available in the immersive-VR ARGO3D platform: (A) Walk (B) Flight and (C) Drone mode. Inset in the upper right corner represents location of the player within the VR ambient (blue circle); in the lower right corner, GPS data are reported, they regard easting and northing (in UTM format), altitude (m), as well as the compass and the azimuth of where the player is looking. Road for scale.

this mode, he relocates from the ground position to a certain ground clearance, and moves parallel to the ground as he's flying. The view is oriented orthogonally to the terrain, as the user can look at it from above, experiencing the whole comprehension of the scene. Two special keys allow to get closer or further from the ground, in a range of given values of ground clearance (this to avoid the user to be disoriented being too near to the ground or too far from it). When the user leaves "Flight" mode, it is relocated to the same position he currently was but near to the ground. The subsequent ground clearance depends on the mode the user is switching to: if on "Walk Mode", he finds himself standing on the ground; on the contrary, if on "Drone Mode" he's at the minimum height allowed by this mode.

In "Drone Mode" (Fig. 3C) the user has both the ability to move free in the scene and to not be subjected to gravity. The aim of this mode is to make the user feel like he's a radio-controlled drone, so he can fly on peaks and move to a certain ground clearance, but without the constraint of being forced to look in only one direction. If he tries to get too close to the ground he feels something like an elastic barrier that prevents him from crashing on the ground and relocate him at the correct height. In this mode the user is faster than in the others. A special feature allows him to aim at a point of the scene with tracked controller and rapidly move up to that point by pressing a trigger button. This was meant to be a quick way to reach certain locations far from the user.

In every mode it is visible to the user a GUI that shows useful information: a map in which it is possible to seek user position in the scene (view from above); below the map is located text with real world coordinates (northing and easting) and real world altitude, relative to his position in every moment; below this text there's a compass with azimuth of where the player is looking.

C. Technologies

1) Hardware technologies: The tool on which Virtual Reality is implemented is Oculus Rift (Fig. 4). The Oculus Rift headset uses an OLED panel for each eye, each having a pixel resolution of 1080x1200. The high refresh rate, global refresh and low persistence allow the user to experience none of the motion blurring or judder that is experienced on a regular monitor. It uses lenses with a wide field of view. The separation of the lenses is adjustable by a dial on the bottom of the device, in order to accommodate a wide range of interpupillary distances. The same pair of lenses are used for all users, however there are multiple facial interfaces so that the user's eyes can be positioned at a different distance. This also allows for users wearing glasses to use the Rift, as well as users with widely varying facial shapes.

The Rift has full 6 degrees of freedom rotational and positional tracking. This tracking is performed by Oculus's Constellation used to track the position of the user's head as well as other VR devices, consisting of external infrared tracking sensors that optically track specially designed VR devices.

Fig. 4. Player using the Oculus Rift (VR headset) and the input peripherals.

The constellation sensor comes with a stand of a desk lamp form factor, but has standard screw holes and can be detached from this stand and mounted anywhere appropriate to the user. The Rift, or any other device being tracked by the system, is fitted with a series of precisely positioned infrared LEDs under or above the surface, set to blink in a specific pattern. By knowing the configuration of the LEDs on the objects and their pattern, the system can determine the precise position of the device with sub-millimeter accuracy and near-zero latency.

The Oculus Rift's motion controller system is known as Oculus Touch. It consists of a pair of handheld units, one for each hand, each containing an analog stick, three buttons, and two triggers. The controllers are fully tracked in 3D space by the Constellation system, so they may be represented in the virtual environment, and each controller features a system for detecting finger gestures the user may make while holding them.

2) Software tools: To develop all the scenes and to test the experience it's been chosen Unity3D v. 2017. Unity is composed of a set of windows to control various parts of the scene and development, such as the Scene Window, where it's possible to locate the assets as they'll be seen in the final version of the software, the Game Window, where it's possible to test the runtime behaviour of the software while editing it, and the Hierarchy, where the developer is able to keep track of all the active asset in the scene.

Every asset has a component that describes it. The essential one is the "Transform" component, which describes position, rotation and scale of the object in the scene. If an object is visible he must have a "Renderer" component, which can be a 2D or 3D Renderer, depending on purpose and design. Other important components are "Collider", which calculates where

there's been a collision between objects and collects information about it, "Rigidbody", which simulates the behaviour of an object subject to kinematic and dynamic laws.

When a script, which implements an object-oriented paradigm, is created, it is attached to the object as a component. It's responsible of computation of data, but also of controlling and coordinating components, in order to obtain aimed behaviour in the scene.

V. Related Works

VR has been extensively used in geosciences. In this section we give an overview of some interesting recent use cases exploiting VR technologies.

Kinsland and Borst in [9] have presented the utility of 3D virtual reality systems in the interpretation of various 3D geologic/geophysical datasets. They have more than 10 years experience mainly exploiting large tiled display systems. In their studies they have identified some reasons that can lead to low utilization of the investigated systems for geological studies, i.e.: geologists would like to use their own desks to access other data useful for interpretation, they would require support from computer specialists, they experience motion sickness. Therefore, in the design of our ARGO3D platform we have tried to minimize these issues by improving the software usability and portability to common desktops with head mounted device and mobile devices.

Cerfontaine et al in [2] presented the workflow used to create immersive visualizations and spatial interaction for geophysical data with head mounted devices. They have used Unreal Engine 4 as VR tool for developing the platform to perform simple scene navigation (more interaction mechanisms are foreseen as future works). For the ARGO3D platform it was decided to use a different game engine and editor: Unity 3D 2017. This game engine was best-suited for our development, because the complexity of scripts and the weight of the computation is quite proportional to the amount of asset and functionality of the software. This is different from other game editors, which need more work and computational power to reach initial acceptable performances. Unity supports the object-oriented programming language C#, that is more modular, so it is possible to define different functionalities in different period of development, more simple and structured, so very easy to read and understand; C# also has very good performance compared to his simplicity. Unity is much more documented on the web, so it's been easy to tackle most of the problems.

Finally, the authors in [6] discuss how VR technology could support geohazard research. In our project, we are further expanding the investigation of the application of VR for geo-hazards' assessment, by implementing the ARGO3D platform.

VI. Conclusions

We presented the first results of our innovative research centered on experimenting and implementing a new approach designed to use the immersive VR for geo-hazards' assessment. Such research results in a platform, here named ARGO3D, that contains several ad-hoc tools, and will be soon available on http://argo3d.unimib.it/. Applications comprise both research and teaching activities in Earth Sciences, the target audience spans from academic to non-academic people. ARGO3D is designed to facilitate communicating geological environments and hazards to students, early career scientists, civil planning organizations and citizens.

We described the overall workflow for real ambient reconstruction, processing and rendering of the virtual ambient, the VR tools and hardware devices employed.

Our approach is aimed at giving the possibility of navigating into a 3D geological environment, where it is possible to observe in detail and measure objects of interest for Earth Sciences and geo-hazards in particular. This approach allows to re-create in laboratory real geological settings in order to share data and study key sites of interest at any time. Moreover, this allows to investigate also sites which are logistically difficult to reach in the field, or dangerous as, for example, active volcanic areas.

Acknowledgment

This work benefits from ESA Project Nr. 38829 (PI Fabio L. Bonali) and from fundings from the Italian Ministry of Education, Universities and Research: "Agreement University of Milan Bicocca – Consortium Cometa for the evaluation of leading-edge interactive technologies for improving teaching and popularization of science". This article is also an outcome of Project MIUR - Dipartimenti di Eccellenza 2018-2022.

References

[1] Ajay Karthic B Gopinath Bharathi and Conrad S Tucker. Investigating the impact of interactive immersive virtual reality environments in enhancing task performance in online engineering design activities. In *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages V003T04A004–V003T04A004. American Society of Mechanical Engineers, 2015.

[2] Philippe A Cerfontaine, Anne-Sophie Mreyen, and Hans-Balder Havenith. Immersive visualization of geophysical data. In *3D Imaging (IC3D), 2016 International Conference on*, pages 1–6. IEEE, 2016.

[3] Ciro Donalek, S George Djorgovski, Alex Cioc, Anwell Wang, Jerry Zhang, Elizabeth Lawler, Stacy Yeh, Ashish Mahabal, Matthew Graham, Andrew Drake, et al. Immersive and collaborative data visualization using virtual reality platforms. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 609–614. IEEE, 2014.

[4] Bonali F.L., Tibaldi A., Marchese F., Fallati L., Russo E., Corselli C., and Savini A. Uav survey in volcano-tectonics: Methodology, best practice and application to the iceland rift. *Journal of Structural Geology*, Submitted.

[5] Laura Freina and Michela Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The International Scientific Conference eLearning and Software for Education*, volume 1, page 133. " Carol I" National Defence University, 2015.

[6] Hans-Balder Havenith, Philippe Cerfontaine, and Anne-Sophie Mreyen. How virtual reality can help visualise and assess geohazards. *International Journal of Digital Earth*, pages 1–17, 2017.

[7] MR James and Stuart Robson. Straightforward reconstruction of 3d surfaces and topography with a camera: Accuracy and geoscience application. *Journal of Geophysical Research: Earth Surface*, 117(F3), 2012.

[8] V Juřík, L Herman, P Kubíček, Z Stachoň, and Č Šašinka. Cognitive aspects of collaboration in 3d virtual environments. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:663, 2016.

[9] Gary L Kinsland and Christoph W Borst. Visualization and interpretation of geologic data in 3d virtual reality. *Interpretation*, 3(3):SX13–SX20, 2015.

[10] Danielle Oprean, Mark Simpson, and Alexander Klippel. Collaborating remotely: an evaluation of immersive capabilities on spatial experiences and team membership. *International Journal of Digital Earth*, 11(4):420–436, 2018.

[11] George G Robertson, Stuart K Card, and Jock D Mackinlay. Three views of virtual reality: nonimmersive virtual reality. *Computer*, 26(2):81, 1993.

[12] Cornelis Stal, Jean Bourgeois, Philippe De Maeyer, Guy De Mulder, Alain De Wulf, Rudi Goossens, Marijn Hendrickx, Timothy Nuttens, and Birger Stichelbaut. Test case on the quality analysis of structure from motion in airborne applications. In *32nd EARSeL Symposium: Advances in geosciences*. European Association of Remote Sensing Laboratories (EARSeL), 2012.

[13] JD Tibbett, FT Suorineni, and BK Hebblewhite. The use of virtual reality scientific visualisation for investigation and exploration of block cave mining system data. In *Proceedings of the Virtual Reality and Spatial Information Applications in the Mining Industry Conference, 2015b University of Pretoria, South Africa. The Southern African Institute of Mining and Metallurgy*, pages 1–11, 2015.

[14] Darren Turner, Arko Lucieer, and Christopher Watson. An automated technique for generating georectified mosaics from ultra-high resolution unmanned aerial vehicle (uav) imagery, based on structure from motion (sfm) point clouds. *Remote Sensing*, 4(5):1392–1410, 2012.

[15] MJ Westoby, J Brasington, NF Glasser, MJ Hambrey, and JM Reynolds. 'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 535–544
DOI: 10.15439/2018F206
ISSN 2300-5963 ACSIS, Vol. 15

Image Clustering Method based on Particle Swarm Optimization

Iuliia Kim, Anastasiia Matveeva and Ilya Viksnin
St. Petersburg National Research University of Information Technologies, Mechanics and Optics
Kronverkskiy Prospekt, 49, Saint-Petersburg, 197101, Russian Federation
Email: yulia1344@gmail.com, anastasiamatveevaitmo@gmail.com, wixnin@cit.ifmo.ru

Igor Kotenko
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS)
14-th Liniya, 39, Saint-Petersburg, 199178, Russia
Email: ivkote@comsec.spb.ru

***Abstract*—To implement efficient computer vision mechanisms, efficient image clustering methods are important. The paper elaborates a clustering method based on particle swarm optimization (PSO) which provides automatic establishment of clustering parameters. The developed PSO based clustering method was tested on 860 images for a car vision system and its results and contribution to the pattern recognition quality improvement were assessed in comparison with fuzzy C-means and k-means. The results do not differ significantly, but distinction in average time of work for these methods was noted. The PSO clustering method is faster than k-means and slower than fuzzy C-means. However, fuzzy C-means method does not guarantee correct results during the further analysis, so the PSO clustering method can be more efficient for implementation in computer vision systems.**

I. Introduction

COMPUTER vision is one of the most prospective fields in different spheres of life [1-4]. It is a wide area of theoretical materials and technical methods connected with object detection, object tracking and object classification. It can be embedded in devices in order to automate their working processes and make them perform pattern recognition tasks.

To provide the correct work of pattern recognition, it is necessary to elaborate efficient methods of image processing; and clustering is one of them. However, many existing clustering methods depend on manual tuning and do not guarantee accurate results in case of visual information distortion.

Usage of machine vision is one of the most important challenge of automatization in different spheres of social life. Authors study implementation of machine vision for autonomous vehicles. Main tasks of such research are defining of other road users (autonomous cars), recognition of road infrastructure and road marking [5-7].

Most of the existing solutions are particular – it seems necessary to make some adjustments to use such solutions in other conditions. Some of the approaches for organizing machine vision need a lot of computing power.

The paper is dedicated to the ways of pattern recognition quality improvement. Relevance of the proposed paper to the computer vision area can be explained by the fact that computer vision can be embedded in cars, drones and used in order to automate such process as road traffic and space monitoring. It is vital to provide a correct work of computer vision, because its errors can have fatal consequences: victims and damages.

In order to avoid such consequences, the authors state the purpose to elaborate efficient methods of image processing. Among them it is possible to call segmentation and clustering. However, they do not guarantee accurate results in case of visual information distortion or high noise level. The reason is in the need for manual tuning of these methods.

The authors of the paper tried to develop a clustering method based on the particle swarm optimization (PSO) that sets parameters automatically. This method also includes *k*-means clustering method, which is intended to group pixels into predetermined number of clusters by calculating the minimum value of distance function. Relative to the PSO, the particle group motion (pixel-by-pixel passage through the image) and search for the best solution for the entire swarm (search for pixel with a maximum average intensity value in a certain region) are used.

In comparison with the original *k*-means method, the need of cluster amount predetermining by user is eliminated, and apart from the operation of distance function minimization the operation of color function minimization is added. An object in an image (or a part of an object) is characterized with relative color uniformity, i.e. if the distance function for pixels *a* and *b* that belong to one object tends to minimum, their color function also tends to minimum. In this paper experiments with non-spoiled images are represented in order to prove preparedness of this method for implementing in process of work with distorted visual information.

Further structure of this paper is as follows. Section 2 provides the overview of the relevant works and sets the theoretical background. Section 3 considers the research task statement and the suggested clustering method based on particle swarm optimization. Section 4 demonstrates the experimental results. Conclusions and future research directions are outlined in section 5.

II. Related Work and Theoretical Background

One of the main purposes in the area of computer vision is the error percentage reduction in the pattern recognition. In this paper the pattern recognition error is a situation when the needed object in the image either is not detected or is detected incorrectly. The pattern recognition process consists in three basic stages:

- filtering and image preparation for the analysis; the image preparation includes image compression, selection of needed regions, ridding of noise;
- logical processing of filtering results, which is responsible for object detection in the filtered image;
- decision-making based on the results of the logical processing, which implies classification of the detected objects.

Image segmentation [8] is related to the ways of increasing the accuracy of recognition. It is a process of dividing a digital image into a set of its constituent regions in order to select objects and their boundaries. As there is no general solution for the image segmentation task, different methods and algorithms were created, which are oriented to the determined categories of images. Depending on the category of the image, its priority properties and then the ways of grouping them are chosen.

The methods listed below can serve as examples:

- active contour method [9] (deformation of original image contours to the boundaries of the specified objects);
- topological alignment method (matching two consecutive frames in the image stream);
- watershed method [10] (establishment of boundary watersheds between different segments according to the determined rule).

As for semantic part [11-13] of the visual information, the following kinds of image segmentation can be distinguished:

- semantic segmentation based on the use of Fully Convolutional Networks (pixel-to-pixel mapping without prior allocation of specific areas);
- weakly supervised semantic segmentation (with use of bounding frames and special labels on the images);
- region-based semantic segmentation (region allocation based on predetermined grouping rules).

One of the segmentation methods is clustering. Image clustering is a division of pixels into several non-intersecting groups (clusters) in such a way that pixels from the same group have similar features, meanwhile the features of pixels from different groups vary significantly from each other.

Clustering task statement:

1) There are: X – a set, which consists of N objects; C – a set, which consists of M identifiers, such as number, name or label; the distance function between objects $d(x,x')$, where x and x' are two objects in the image. The distance function is represented in (1):

$$d(x;x') = \sqrt{(x-x')^2} \tag{1}$$

2) It is necessary to divide the set of objects X into M non-intersecting subsets (clusters) in such a way that each cluster was represented as an aggregate of objects from the set X, whose distance function values d are closed to each other. In addition, the following conditions must be fulfilled:
   - each cluster is assigned a cluster identifier C_j; $j \in [1; M]$ (number, name, label);
   - each object x_i; $i \in [1; N]$ can belong to one and only one cluster.

There are plenty of different clustering methods. Some of the most spread examples of these methods are k-means and fuzzy C-means. The idea of the k-means algorithm is to minimize the distance between objects in the clusters. The algorithm stops working when the further minimization becomes impossible.

The main step of the k-means algorithm:

1) At the beginning of the algorithm the quantity of clusters is set and then, according to the determined rule, centroids are allocated (centers of mass of clusters). The minimizing function is represented in (2):

$$J = \sum_{i=1}^{N}\sum_{k=1}^{M} d(x_i, c_k) \tag{2}$$

   where X – a set of clustering objects, $x_i \in X$ a clustering object, $i \in [1; N]$, C – a set of clusters, $c_k \in C$ – centroid, $k \in [1; M]$, M – an amount of clusters, N – an amount of objects, d – a value of distance function between object and centroid.
2) Each object correlates with the determined cluster by calculating the value of the distance function between this object and each center of mass and then selecting the least one among the calculated values. After that the centers of mass of clusters are recalculated, as in (3):

$$c_j = \frac{(\sum_{t=1}^{T}(x_t))}{T} \tag{3}$$

   where $x_t \in C_j$; $t \in [1; T]$; T – an amount of objects in the cluster C_j; j – a cluster number, $j \in [1; M]$; M – an amount of clusters.
3) If $c_j = c_j - 1$, it means that object clustering is completed, otherwise it is necessary to return to the second step and recalculate centroids again.

Fuzzy C-means algorithm is based on the fuzzy logic, i.e. on the assumptions that each clustering object from the set X, which consists of N objects to some extent belongs to a particular cluster from the set of clusters C.

The main step of the fuzzy C-means algorithm:

1) As input values, there are: M – an amount of clusters, $1 < m < \infty$ – a measure of accuracy, $0 < \varepsilon < 1$ – a criterion of the end, $U_0 = u_{ij}(x_i; c_j)$: $x_i \in X$; $c_j \in C$ –

a weighting matrix of belonging of the clustering object $x_i \in X$; $i \in [1; N]$; to the cluster $C_j \in C$; $j \in [1; M]$; $0 < u_{ij} < 1$:

The minimizing function is shown in (4):

$$J = \sum_{i=1}^{N} \sum_{k=1}^{M} u_{ij}^{m} \cdot d(x_i, c_k) \quad (4)$$

2) Then the centroids are calculated, as in (5):

$$c_j = \frac{(\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i)}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (5)$$

3) After this the weights are recalculated, as in (6):

$$u_{ij} = \frac{1}{\sum_{k=1}^{M} \left(\frac{d(c_j, x_i)}{d(c_k, x_i)} \right)^{\frac{2}{m-1}}}$$

4) The next step is to compare $|U_k - U_{k-1}|$ with the value of ε. In case the first value is less than the second one, the algorithm is finished, otherwise it is necessary to return to the step 2 and recalculate the centroids.

III. Clustering Method Based on Particle Swarm Optimization

The algorithms mentioned above have quite substantial drawbacks, such as:

- sensitivity to outliers (values that are allocated from the total sample);
- need for the user to specify the amount of clusters beforehand;
- presence of some uncertainty degree in the threshold parameter definition;
- uncertainty of actions with objects that possess the properties of two different clusters simultaneously;
- need for the user to specify clustering parameters.

Due to shortcomings of clustering algorithms represented by authors, there is such a situation when the clustering task can be solved incorrectly, namely: for the clustering objects $x_i \in X$; $i \in [1; N]$ and the clusters $C_j \in C$; $j \in [1; M]$ with the centroids c_j $\exists t \neq j$ $g_{xi} \neq g_{ct}$; $g_{xi} = g_{cj}$; $x_i \in C_t$; $x_i \notin C_j$; where g_{xi} – a label of belonging of the i^{th} object to the cluster, g_{cj} – a label of belonging of the centroid to the j^{th} cluster.

In this way, authors of this paper have the following research task: it is necessary to find such centroids c_j for the $C_j \in C$ that $\forall x_i \in X$ $\exists! C_j$ $g_{xi} = g_{cj}$; $x_i \in C_j$; $g_{xi} \notin g_{ca}$; $x_i \notin C_a$; $g_{xi} \neq g_{cb}$; $x_i \notin C_b$; $a \in [1; j-1]$; $b \in [j+1; M]$.

To solve this pattern recognition quality improvement task, it is essential to exclude the user's direct participation in the process of specifying the amount of clusters and rule (or set of rules) during the initial centroid allocation.

Methods providing automated image segmentation and clustering were proposed in many scientific works. In the [14] processing of image segments for fruit peel defects detection was proposed. The source image is divided into segments, which are given as an input to Adaptive Artificial Neural Network (AANN). The image processing is automated, because AANN adapts to each input and its features.

Segmentation based on color maps was developed in the article [15]. Its authors in the research rely on four-color labeling theorem. They improved the existed four-color labeling method, which used random initial distribution and developed heuristic four-color labeling. This iterative algorithm generates more reasonable color maps with a global view of the whole image and provides better results in case of images with clutters and complicated structures.

In the paper [16] image clustering is used in order to provide automated retinal screening. In the described method heuristic based clustering is included. Initial centers are allocated according to measures defining statistical distribution of data is incorporated in the proposed methodology.

Another effective clustering method was proposed for heterogeneous disease expression data [17]. Recursive K-means spectral clustering method (ReKS) was developed, which was found to be superior to the hierarchical clustering method and much faster than *k*-means.

In the work [18] a novel data clustering algorithm was elaborated. It is based on heuristic rules, which are built according to *k*-nearest chain, and give an opportunity to get rid of the need in specifying the number of clusters. K-Nearest Neighbors Chain (KNNC) serves as basis for proposing two heuristic rules to find initial clusters and their proper amount. The first heuristic rule reflects the degree of separation of clusters and the second rule determines the inner compactness of a cluster.

In order to resolve the issue of arbitrary choices on clustering parameters, authors decided to use some elements of particle swarm [19] optimization.

Particle swarm optimization [20] method consists in the following steps:

1) There is a swarm of S particles, and each of them is assigned a coordinate $x_i \in \boldsymbol{R_n}$ and a velocity $v_i \in \boldsymbol{R_n}$; f: $\boldsymbol{R_n} \rightarrow$ R is an objective function that needs to be minimized; p_i – the best known position of the i^{th} particle (in the context of solving the given optimization problem); g – the best known state of the entire swarm.
2) For each particle $s_i \in S$; $i \in [1; S]$ it is needed to:
   - generate an initial position using a random vector in the range from r_{min} to r_{max} these values are lower and upper boundaries of the search-space, respectively;
   - assign to the best known position of particle p_i its initial value x_i;

- in case $f(p_i) < f(g)$, there is a necessity to update the value from g to p_i;
- generate velocity value of the particle v_i, which belongs to the interval from $-(r_{max} - r_{min})$ to $(r_{max} - r_{min})$.

3) It is required to repeat the following sequencing for each i^{th} particle until the predetermined stopping criterion is fulfilled:
   - generate random vectors r_p and r_g, which have a range of admissible values in the interval between 0 and 1;
   - update the velocity value of the particle, as in (7):

$$v_i = w \cdot v_i + \varphi_p \cdot r_p \times (p_i - x_i) + \varphi_g \cdot r_g \times (g - x_i) \quad (7)$$

   where × is a vector product operation, w; φ_p; φ_g – are the parameters specified by user;
   - change the particle position according to (8):

$$x_i = x_i + v_i \quad (8)$$

   - compare the values of $f(x_i)$ and $f(p_i)$; if the first value is less than the second one, it is needful to update the best known position of the particle from p_i to x_i and then in case $f(p_i) < f(g)$; it is necessary to change the value of the parameter g to the value of the parameter p_i;
4) As a result of the operations above, the parameter g will contain the best solution.

For pattern recognition quality improvement the authors of the paper developed a clustering method which combines some elements from particle swarm optimization (numerical optimization method) and from k-means algorithm (cluster analysis method). From each method such operations were selected that do not require random parameter settings and do not take into account user's subjective opinion (user has just an observing role). In this clustering method all the parameter calculation will happen automatically, and the user no longer needs to generate manually any input values.

It is necessary to normalize the source image before using the developed algorithm. Normalization allows making an image insensitive to the light changes, ridding it of unnecessary noise.

It is achieved by dividing the RGB vector components of each pixel by the length of this vector, as in (9):

$$(r', g', b') = \left(\frac{r}{\sqrt{r^2 + g^2 + b^2}}, \frac{g}{\sqrt{r^2 + g^2 + b^2}}, \frac{b}{\sqrt{r^2 + g^2 + b^2}} \right) \quad (9)$$

where r, g, b are the initial values of pixel's RGB vector; r', g', b' are the normalized values of pixel's RGB vector.

The algorithm of the developed clustering method consists in the following procedures:

1) Rounding W' pixels horizontally and H' pixels vertically to the nearest values of W and H, respectively, which are multiple of 10.
2) Sequential selection of 10 by 10 regions (clusters) in the image and search for a pixel with a maximum average intensity value in each region – these pixels will be centers of mass c_j; $j \in [1; W \cdot H / 100]$ (in case there are more than one pixel with a maximum average intensity value in the region, it is possible to choose any of them). The formula of pixel's average intensity calculation is represented in (10):

$$I_{av} = \frac{(r' + g' + b')}{3} \quad (10)$$

   where r', g', b' are the normalized values of pixel's RGB vector.
3) Comparison of the rounded average intensity values for elements with maximum average intensity values from neighboring regions. It was found empirically by authors that the most effective rounding is to two decimal places. If the rounded average intensity values are equal to each other, two neighboring clusters are combined into one. In the new cluster the centroid is the pixel with a maximum average intensity value. It is necessary to repeat this step until there are M clusters c_j; $j \in [1; M]$ with the pairwise distinct rounded average intensity values of the centers of mass.
4) Calculation of two parameters for each pixel x_i; $i \in [1; W \cdot H]$ relative to each centroid: distance function d and so-called color function f. The color function is represented in (11):

$$f(x_i, c_j) = \sqrt{(r'_{xi} - r'_{cj})^2 + (g'_{xi} - g'_{cj})^2 + (b'_{xi} - b'_{cj})^2} \quad (11)$$

   where r'_{xi}; g'_{xi}; b'_{xi} are the normalized [19] values of RGB vector of the pixel x_i; c_j is the centroid of the cluster C_j; r'_{cj}; g'_{cj}; b'_{cj} are the normalized values of RGB vector of the centroid c_j.
5) Then for the pixel x_i it is necessary to find a centroid c_a, $a \in [1; M]$ relative to which the square root of distance function value will be minimum and a centroid c_b; $b \in [1; M]$ relative to which the value of color function value will be minimum. After that the following differences need to be calculated, as it is represented in (12) and (13):

$$d_{diff} = | d(x_i, c_a) - d(x_i, c_b) | \quad (12)$$

$$f_{diff} = | f(x_i, c_a) - f(x_i, c_b) | \quad (13)$$

6) The function, whose difference was less, is chosen as a priority function (in case d_{diff} equals f_{diff}, the distance function obtains a priority, because pixels that are closer to each other more likely belong to the same object than the ones that have similar colors). The allocation of pixels to clusters is realized according to the priority function, i.e. the pixel x_i will be assigned to a cluster, if the priority function value

between this pixel and this cluster's centroid is minimal.

7) Ridding the image of noise:
For this purpose the authors chose non-local means method. It is illustrated in (14):

$$u(p)=\frac{1}{C(p)}\int_{\Omega} v(q), f(p,q)dq \quad (14)$$

where *u(p)* is the filtered intensity value of pixel color component at point *p*, *v(q)* is the unfiltered intensity value of pixel color component at point *q*; *f(p; q)* – weighting function, *C(p)* – normalizing factor.
As the weighting function Gaussian function is used, it is shown in (15):

$$f(p,q)=e^{\frac{-|B(q)-B(p)|^2}{h^2}} \quad (15)$$

where *h* is the filter parameter (in general, for RGB color images *h* = 3), *B(p)* is the local average intensity value of color components of pixels around the point *p*, *B(q)* is the local average intensity value of color components of pixels around the point *q*.
Normalizing factor *C(p)* is calculated, as in (16):

$$C(p)=\int_{\Omega} f(p,q)dq \quad (16)$$

The developed clustering method uses the following elements borrowed from the particle swarm optimization method: particle group motion (pixel-by-pixel passage through the image), search for the best solution for the entire swarm (search for pixel with a maximum average intensity value in a certain region).

At the same time the main differences from the original algorithm are the next points: each particle has a fixed velocity value which excludes the necessity of its manual recalculating by user, initial particle parameters are not specified randomly.

As for the *k*-means cluster analysis method [21], its next aspects were improved: the need of cluster amount predetermining by the user was eliminated, apart from distance function minimization an operation of color function minimization was added, which gave an opportunity to increase the probability that pixels will be assigned to clusters correctly.

IV. Experiments

To check the effectiveness of the clustering method based on particle swarm optimization 860 test images of cars and 860 images of road signs were picked and normalized [22].

For this purpose the mixture of manual photos, images provided by Stanford University laboratory and images from Russian Traffic Sign Dataset (both datasets are publicly available) was used. Authors have analyzed existing pictures with some samples with ideal results of clustering and figure recognition.

Fig. 1 and Fig. 2 outline examples from this set of test images.

The normalization [23] results of these source images are showed in Fig. 3 and Fig. 4.

Fig. 5 and Fig. 6 depict the work results of the clustering method based on particle swarm optimization.

One of the shortcomings of the developed method is that optimal rounding of the average intensity value for cluster combining empirically established by the authors is not universal.

It may provoke two opposite situations:

1) there are extra clusters in the output image, especially in the places of glare;
2) in the output image several different objects are merged into one cluster.

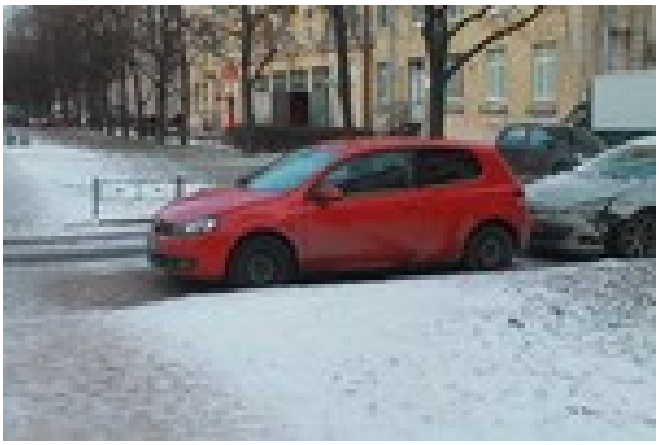


Fig. 1. Source image of a car

Fig 2. Source image of a road sign

Fig. 3. Normalized image of the car

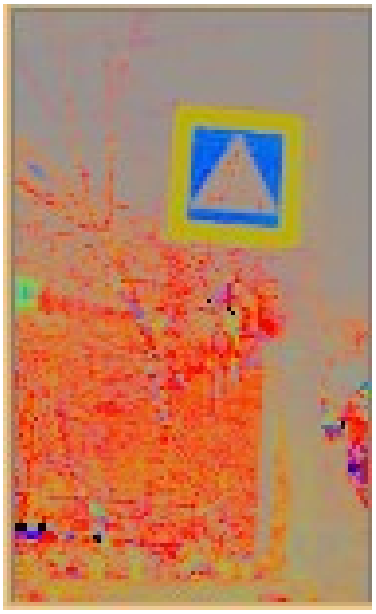


Fig. 4. Normalized image of the road sign

Fig. 5. Clustered image of the car (PSO clustering method)

The examples of work of the clustering method based on the PSO with reduced number of decimal places (1 decimal place) in the rounded average intensity value are represented in Fig. 7 and Fig. 8. All the objects were combined into one.

Thus, the image in Fig.7 seems to be one-colored – the proposed approach can find only one color, so, it seems impossible to recognize any figure on the picture.

Fig. 6. Clustered image of the road sign (PSO clustering method)

Fig. 7. Example of clustering quality worsening because of reduced number of decimal places in the rounded average intensity value (car image)

The examples of work of the clustering method based on the PSO with increased number of decimal places (3 decimal places) in the rounded average intensity value are represented in Fig. 9 and Fig. 10. In the given image there are extra detected regions: piece of land, parts of sky, building.

Also, this method depends on the input image size: the more the number of the analyzed pixels, the greater is risk that more clusters will be detected, and due to this the time of work will be enlarged. This causes the necessity of proportional reducing the source image size or increasing the initial cluster size.

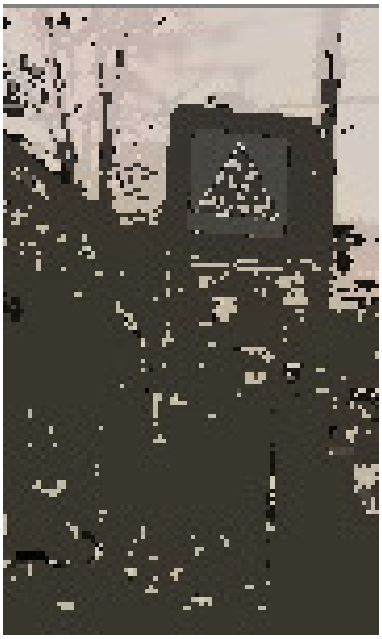


Fig. 8. Example of clustering quality worsening
because of reduced number of decimal places
in the rounded average intensity value (road sign image)

Fig. 9. Example of clustering quality worsening
because of increased number
of decimal places in the rounded average intensity value (car image)

The example of work of the clustering method based on the PSO with increased number of resultant cluster quantity due to absence of preliminary image size reduction is represented in Fig. 6. The size of the source image of the road sign is 500 × 250. It is notable that in the resultant image many extra clusters were detected that during further analysis can provoke different issues.

To perform clustering, the authors reduced proportionally the image size. Currently the size of reduced image needs to be established empirically. Particularly, for the images represented in this paper the maximum size of 150 pixels wide and tall was chosen.

Thus, the developed clustering method currently is suitable for small images, which are not overloaded with details and excessive amount of glares.

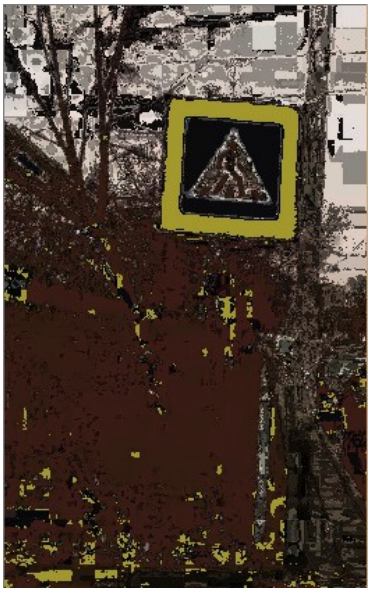


Fig. 10. Example of clustering quality worsening
because of increased number of decimal places
in the rounded average intensity value (road sign image)

To improve the effectiveness of this method, it is planned to disclose dependencies between image features (pixel quantity, histogram of gradients) and such aspects as: number of decimal places in rounded average intensity value, initial cluster size.

To perform comparison, apart from the clustering method based on the PSO, the test images were clustered by the fuzzy *C*-means and the *k*-means (as an input value, the number of clusters obtained in the PSO clustering method was used; maximum iteration quantity – 3). The average working time of mentioned clustering methods are represented in Table 1.

TABLE I. COMPARISON TABLE OF TIME OF WORK
FOR PSO, FUZZY C-MEANS AND K-MEANS CLUSTERING METHODS

PSO	Fuzzy *C*-means	*k*-means
1.69	0.0013	35.98

Fig. 13 and Fig. 14 represent the worked results of *k*-means clustering method. Fig. 11 and Fig. 12 show the working results of the fuzzy *C*-means clustering method.

With *k*-means and fuzzy C-means clustering methods it became possible to separate objects from background. However, the algorithm proposed by the authors marks details more legibly.

To illuminate influence of the represented clustering methods on the pattern recognition quality the authors implemented Haar [23] cascade classifier on the given example to detect the car. It is possible to see the results in Fig. 15-20.

PSO and *k*-means clustering methods contributed to obtaining correct results: region with the car was entirely detected. Fuzzy *C*-means algorithm despite its high speed does not guarantee accurate recognition.

Fig. 11. Clustered image of the car
by fuzzy *C*-means clustering method (car image)

Fig. 12. Clustered image of the car
by fuzzy *C*-means clustering method (road sign image)

Fig. 13. Clustered image of the car by *k*-means clustering method
(car image)

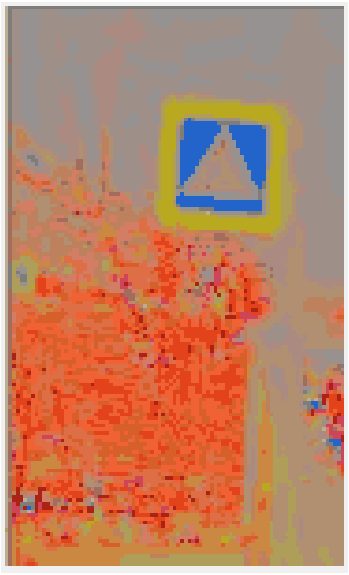


Fig. 14. Clustered image of the car
by *k*-means clustering method (road sign image)

Fig. 15. The result of classifying (PSO clustered image of car)

Fig. 16. The result of classifying (PSO clustered image of road sign)

Fig. 17. The result of classifying (fuzzy *C*-means clustered image of car)

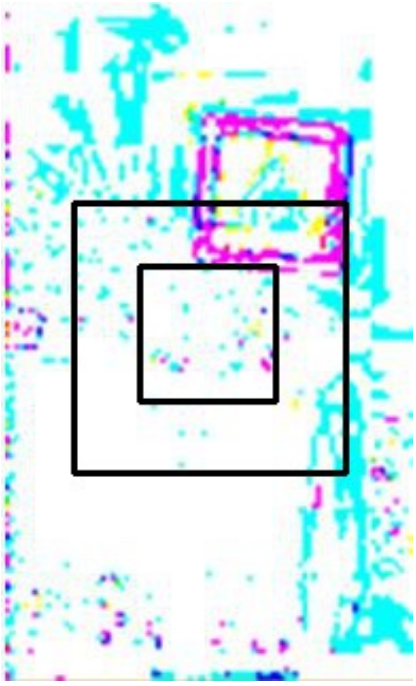


Fig. 18. The result of classifying
(fuzzy *C*-means clustered image of road sign)

Fig. 19. The result of classifying (*k*-means clustered image of car)

Comparing PSO and k-means, the second one is slower because of using several iterations.

Reducing iteration number can have a negative impact on clustering quality. PSO determines clusters during the first two steps without the risk of worsening results.

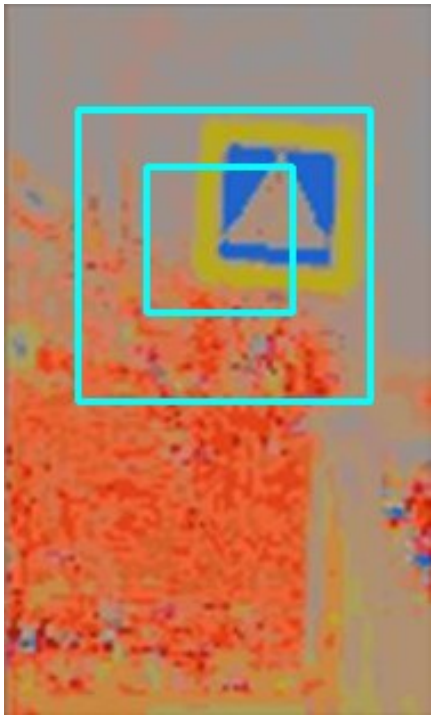


Fig. 20. The result of classifying
(*k*-means clustered image of road sign)

V. Conclusion

The paper proposed and investigated the clustering method based on particle swarm optimization. The developed method was tested for a car vision system and its results and contribution to the pattern recognition quality improvement were evaluated.

The experiments demonstrated that the suggested clustering method is faster than the *k*-means, but it cedes in speed to the fuzzy *C*-means. However, it provides more accurate results. The advantages of the proposed method are as follows:

(1) use of particle swarm optimization helped to eliminate the necessity for user of cluster amount predetermining - one of the main reasons of incorrect clustering in the majority of cluster analysis methods;

(2) the authors liquidated the need for user of calculating and specifying threshold parameters; this need often led to a situation when one pixel could be assigned to several clusters at the same time which contradicts with the clustering task;

(3) there are rules that regulate grouping of objects, which possess the features of the different clusters; thanks to this the uncertainty in correlation of objects and clusters is minimized and, as a result, the pattern recognition error probability is reduced;

(4) clustering parameters are predetermined automatically and do not require user's intervention.

However, currently, this method has the following disadvantages: absence of clear dependencies between way of rounding the average intensity values and image features; absence of clear dependencies between size of initial cluster and image features.

In further research it is planned to liquidate the mentioned drawbacks of the developed clustering method, propose the ways of its improvement and to finalize the embedded vision prototype. Implementation of the proposed approach may be used for different cyber-physical systems. One of the most interesting sphere is autonomous cars.

Further research will be also aimed on modification of the proposed method and development of software for video streaming analysis for real-time systems. There are such systems for different physical models of autonomous vehicles [24]. Some of them are particular, it is necessary to rebuild algorithm for different conditions of environment, some of them work only with certain functions, required for correct movement of the cars [25-27].

In this case, authors will prepare the method not only for image recognition in usual cases, but also for fully automatic recognition in case of violations. Current results seem to be useful as the main approach for machine vision organization, but this approach should be improved with usage of different existing machine learning methods.

ACKNOWLEDGMENT

This research is being supported by grants of Russian Foundation for Basic Research (projects No. 16-29-09482 and 18-07-01488), by the budget (project No. AAAA-A16-116033110102-5), and by Government of the Russian Federation, Grant 08-08.

REFERENCES

[1] D. A. Forsyth and J. Ponce. Computer Vision: A Modern Approach. Saddle, River, Nj, U, A: Prentice Hall, 2003.

[2] A. Kochan, "Machine vision guides the automotive industry," Sensor Review. J., vol.22, 2002, pp. 119-124.

[3] D. Vernon. Vernon. Machine Vision in the Electronics and PCB Inspec-tion Industry. The Current Position and Future Directions Maynooth College Ireland, 2004.

[4] S.V.Kozlov, E.Yu.Neretin and V.V. Kukolkina, "Machine vision applica-tion in digital dermoscopy for suspected melanoma of the skin," Saratov Journal of Medical Scientific Research. J., vol.10, 2014, pp. 281-285.

[5] B. Ranft, C. Stiller, "The role of machine vision for intelligent vehicles," IEEE Transactions on Intelligent Vehicles, 2016 Mar;1(1):8-19.

[6] R. Baran, A. Glowacz, A.Matiolanski, "The efficient real-and non-real-time make and model recognition of cars," Multimedia Tools and Applications, 2015, 74(12), pp.4269-88.

[7] J. Janai, F. Güney, A. Behl, A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," arXiv preprint arXiv:1704.05519. 2017.

[8] A. A. Aly, S. B. Deris, N. Zaki. Research review for digital image segmentation techniques. 2011.

[9] Y. Xiang, A. C. S. Chung, J. Ye, "A new active contour method based on elastic interaction," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol.1, 2005, pp. 452-457.

[10] N. Li, M. Liu, Y. Li, "Image Segmentation Algorithm using Watershed Transform and Level Set Method," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu J., 2007.

[11] R. G. Mathieu, R. L. Woodard, "Data integrity and the Internet: implications for management," Internet Research J., vol.3, 1995, pp. 3-7.

[12] H. Ibrahim, "A strategy for semantic integrity checking in distributed databases," Ninth International Conference on Parallel and Distributed Systems. J., 2002, pp. 139-144.

[13] K. P. Udagepola, L. Xiang, A.W. Wijeratne, Y. Xiaozong, "Semantic integrity constraint violations check for spatial database updating," Journal of Applied Mathematics and Computer Sciences J., vol.4, 2007.

[14] M. Woźniak, D. Połap, "Adaptive neuro-heuristic hybrid model for fruit peel defects detection," Neural Networks, vol. 98, 2018, pp. 16–33.

[15] K. Li, W. Tao, X. Liu, L. Liu, "Iterative image segmentation with feature driven heuristic four-color labeling," Pattern Recognition, vol. 76, 2018, pp. 69-79.

[16] R. Geetha Ramani, "Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening," Computer methods and programs in biomedicine, vol. 160, 2018, pp. 153-163.

[17] G. T. Huang, K. I. Cunningham, P. V. Benos, C. S. Chennubhotla, "Spectral clustering strategies for heterogeneous decease expression data," Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2013, pp. 212-223.

[18] J. Lu, Q. Zhu, Q. Wu, "A novel data clustering algorithm using heuristic rules based on k-nearest neighbors chain," Engineering Applications of Artificial Intelligence, vol. 72, 2018, pp. 213-227.

[19] A. P. Karpenko, E. Y. Seliverstov, "Overview of the particle swarm methods for the global optimization problem," Science and Education: a scientific edition of the Bauman Moscow State Technical University J., vol.3, 2009, pp. 1-26.

[20] J. Wang, and D. Wang, "Particle swarm optimization with a leader and followers," Progress in Natural Science. J., vol.18, 2008, pp.1437-1443.

[21] J. Qi, Y. Yu, L. Wang, J. Liu, "K*-Means: An Effective and Efficient K-Means Clustering Algorithm," 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), 2016, pp.242-249.

[22] L. Xie, Q. Tian, B. Zhang, 'Feature normalization for part-based image classification," 2013 IEEE International Conference on Image Processing., Melbourne, 2013, pp. 2607-2611.

[23] S. Choudhury, S. P. Chattopadhyay, T. K. Hazra, "Vehicle detection and counting using haar feature-based classifier," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017, pp. 106-109.

[24] B.A. LaPenta, The Ducklingbot: a self-driving robot on a Pi Zero (Doctoral dissertation, Massachusetts Institute of Technology).

[25] V. Kastrinaki, M. Zervakis, K. Kalaitzakis, "A survey of video processing techniques for traffic applications," Image and vision computing, 2003, 21(4), pp.359-81.

[26] J.C. McCall, M.M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," IEEE transactions on intelligent transportation systems, 2006, 7(1), pp.20-37.

[27] M.Y. Fu, Y.S. Huang, "A survey of traffic sign recognition," International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), 2010, IEEE, pp. 119-124.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 545–551
DOI: 10.15439/2018F71
ISSN 2300-5963 ACSIS, Vol. 15

The impact of parallel programming on faster image filtering

Kamil Książek[1], Zbigniew Marszałek[1,2], Giacomo Capizzi[2,1],
Christian Napoli[3], Dawid Połap[1], Marcin Woźniak[1,2]
[1]Institute of Mathematics, Silesian University of Technology,
Kaszubska 23, 44-100 Gliwice, Poland
[2] Department Electrical, Electronics and Informatics Engineering,
University of Catania, V. Andrea Doria 6, 95125 Catania, Italy
[3] Department of Mathematics and Computer Science, University of Catania,
V. Andrea Doria 6, 95125 Catania, Italy
Email: KamilKsiazek95@gmail.com, Zbigniew.Marszalek@polsl.pl, Gcapizzi@diees.unict.it,
Napoli@dmi.unict.it, Dawid.Polap@polsl.pl, Marcin.Wozniak@polsl.pl

***Abstract*—Parallel programming is a field of science with a great potential nowadays due to the development of advanced computers architectures. Appropriate usage of this tool can be therefore highly beneficial in multimedia applications and significantly decreases the time of calculations.**

In this article, we analyze how the speed of calculations is influenced by the usage of parallel algorithms in image filtering processes. We present a method based on multithreading and the division of the image for rectangles. The filter is applied parallel on each part of the image. Results show that in some cases our proposition can bring over 90% benefit when compared to the classical approach.

***Index Terms*—parallel programming, image filtering, Laplacian, multithreading**

I. Introduction

PARALLEL programming is currently a dynamically growing field of computer science. Modern multicore processors enable a significant reduction of computation time. Proper use of computing power is an enormous challenge for programmers. A skillful preparation of parallel instructions that are to be carried out causes a lot of problems. However, benefits of parallel programming are obvious. Some multimedia applications require a huge amount of time - it is visible for instance in image processing. Filtering the images containing the several thousand pixels can take a lot of time, especially, when there is a substantial number of images to be filtered. Therefore an intelligent methods that improve image processing are very important.

It is clear that image filtering has a lot of applications. It is possible to improve the quality of photos with blur, noise or other undesirable effects. Moreover, sharpening the edges can be helpful in the objects detection. Parallel methods can be very useful in graphics processing and cloud computing directed to multithreading. High-quality images are also crucial in medicine in diagnosis of diseases (for instance in X-ray pictures). Therefore, capturing the details is very important. In this article we present multithreading in image filtering, and its impact on the whole process. Our approach is designed to equally distribute work among all the threads. The input image is divided into equal rectangles and each thread filters only the designated area. Therefore by the proposed algorithm we construct a method which uses all available cores. Depending on the number of CPU threads in the computer we can significantly decrease time of processing, reaching even 90% of improvement. It has a significant importance for HD multimedia systems where all the images and multimedia streams are very complex structures. Therefore our proposed method may reduce time and improve the efficiency of processing. For the research we have used an architecture with 32 CPU threads and 320 GB of memory.

The main part of this article is following: Section II describes some related works, Section III presents a theoretical background of the image filtering and three Laplacian filters applicated in the research. In Section IV it is shown a detailed description of the tested parallel method. Section V gives a results of measurements and Section VI contains conclusions and remarks after studies.

II. Related works

Very important application of multitasking is connected with data processing. In [1] it was shown how to parallelize fast sorting algorithm, while in [2] it was proposed a new more efficient parallel merge sort algorithm. These decrease computation time in large databases for instance in case of data analysis. In [3] an overview on intelligent systems for data retrieval was discussed. Graphics processing is frequent and very important topic of many publications. In [4] it was proposed a method how to more efficiently analyze the information from images for detection, and in [5] a segmentation of images based on graph analysis of the semantic image structure was presented. The image decomposition method which combines information from the infrared and visible images is presented in [6]. This method can be helpful for instance in target recognition. In [7] and [8] was presented a system for image data classification by the use of fast selection methods based on shapes comparisons. Authors in [9] propose a Weighted Guided Image Filtering algorithm (WGIF) which prevents so

called "halo artefacts" effect. In many cases combination of different data ensures more efficient analysis. Such approach in medical images is shown in [10]. The method intended for filter identification is presented in [11]. An interesting problem connected with underwater imaging is introduced in [12]. Authors in [13] show the Core algorithm designed for document's identification from images. It is compared with classical detectors like ORB, SIFT and SURF-BRISK. In our paper it will be shown a method which can speed up the calculations on images. A properly and quickly processed image streamlines further work. The presented method is intended for initial processing of the images.

III. Image filtering

A very common situation is that the analysis of the original image is difficult due to noise, blur or other factors. Furthermore, if the number of details is too large, a detection of the crucial parts of the image (for instance, the contours of presented objects) is impossible. Therefore it is necessary to pre-process the image. Further analysis will be easier thanks to such tools as the image filtering.

A. Theoretical background

One of the most popular color models is RGB model [14]. Each pixel consists of three components: R (red), G (green) and B (blue). They can be integers from the range $\{0, 1, ..., 255\}$. For instance $[0, 0, 0]$ represents black, $[255, 255, 255]$ determines white, $[255, 255, 0]$ yellow, etc.

In this paper we assume that calculations will be performed by using the RGB model. Let $R_{n\times m}$ be a two-dimensional array with the values of pixels of a given image (n is the width of an image and m is the height of an image, $R[i, j]$ represents the position (i, j) on the image, $i \in \{1, ..., n\}$, $j \in \{1, ..., m\}$). The idea relies on modification of the image by moving the convolution mask over the pixels. The new values of three components R, G, B depend on the pixels in the nearest neighborhood of the calculated one. This process is illustrated in Fig. 1, by the example of 3×3 mask. The new value of the position (x, y) depends on 9 pixels. Of course, other sizes are also allowed.

The greater the mask, the larger number of pixels is taken into account during processing. The impact of each pixel is defined by the table of weights, called a filter. In our research, there were applied three types of convolution masks: 3×3, 5×5 and 9×9. The larger the filter is, the more details are lost [15]. During calculations on larger masks it is necessary to create an auxiliary image with borders filled with black pixels (the convolution masks exceeds the original one). This operation has an nonsignificant influence on the final image but it enables the filtering. The pattern for the new value of each component of the pixel located at position (i, j) in the case of 5×5 mask is as follows: [16]:

$$R'(i,j) = \frac{1}{M} \sum_{k=-2}^{2} \sum_{l=-2}^{2} w(k,l) \cdot R(i+k, j+l), \quad (1)$$

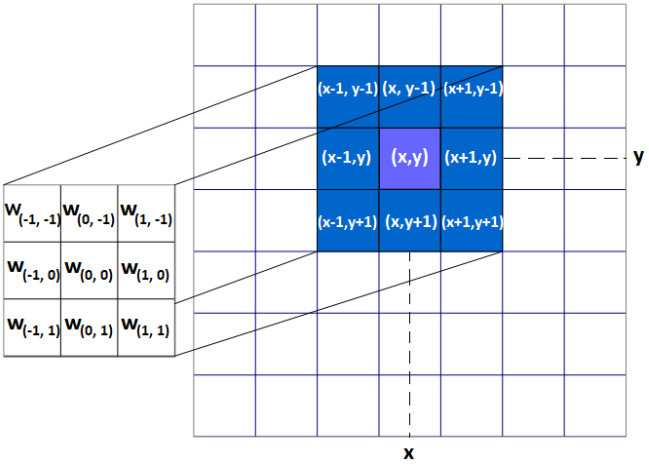


Fig. 1: The sample of using a 3×3 convolutional mask.

where M is the sum of all values in the array (the convolution mask), $w(k, l)$ is the weight of pixel located at position $(i + k, j + l)$ and $R(i + k, j + l)$ is the previous value of the pixel at given position. Sometimes M may be equal to 0. In that situation, the factor $\frac{1}{M}$ is omitted. Patterns for other convolution masks are created similarly.

Fig. 2: Illustration of the image split into 4 rectangles. (The original photo was taken by Jonathan Andreo, and is available at unsplash.com)

B. The applied filters

During further calculations, three Laplacian filters will be used (Fig. 3). [17]. Their main task is sharpening the edges of the objects and hence, loosing of irrelevant details. The picture which has been filtered, is presented in Fig. 4. It is possible to see how the Laplacian filters influence the original image. The edges are therefore definitely more visible than the rest of the image. This kind of filters facilitates the detection of shapes.

IV. Parallelization

Image filtering involves a lot of calculations. The larger the photo is, the greater the time of filtering is. In case of analyzing a large number of pictures, minimizing the computation time is

$$\begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

(a) 3×3 Laplacian filter

$$\begin{pmatrix} -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 24 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

(b) 5×5 Laplacian filter

$$\begin{pmatrix} 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 2 & 5 & 0 & -24 & -40 & -24 & 0 & 5 & 2 \\ 2 & 5 & 3 & -12 & -24 & -12 & 3 & 5 & 2 \\ 1 & 4 & 5 & 3 & 0 & 3 & 5 & 4 & 1 \\ 1 & 2 & 4 & 5 & 5 & 5 & 4 & 2 & 1 \\ 0 & 1 & 1 & 2 & 2 & 2 & 1 & 1 & 0 \end{pmatrix}$$

(c) 9×9 Laplacian filter

Fig. 3: The presentation of Laplacian filters applied in the paper.

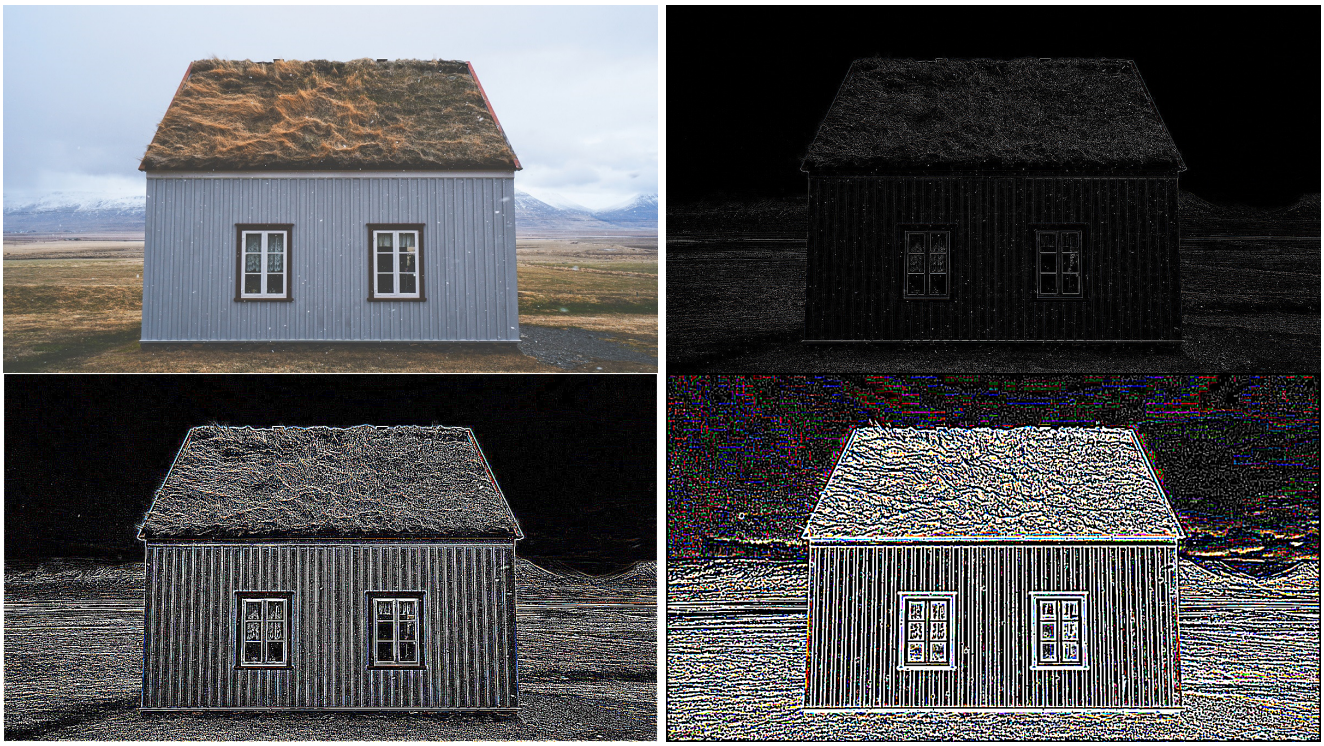


Fig. 4: The figure presents as follows: the original image and the image after 3×3 Laplacian filter (the first row) and images after 5×5 and 9×9 Laplacian filter (the second row). (The original photo was taken by Jonathan Andreo, and is available at unsplash.com)

crucial. The idea presented in this paper relies on moving the mask parallely over the image. The photo can be divided into t rectangles, where t is the number of threads (an exemplary division into 4 rectangles is presented in Fig. 2). Then each thread applies the filter to the allocated area and does not influence any other parts. Finally, all filtered fragments are combined into one image.

The only issue is to determine the way of assigning suitable areas for t threads. Each thread after running receives its own index (the numbering starts from zero). Let w be the width of the processed image (in pixels). Then the number of pixels per one thread is equal to $\lfloor \frac{w}{t} \rfloor$. The ceiling function is necessary because $\frac{w}{t}$ may not be an integer. In such a situation the last thread can have slightly more pixels to calculate. The Algorithm 1 presents the pseudocode of the proposed method.

V. Experimental results

As was said before, application of three Laplacian filters was tested. In all cases, 100 measurements were performed and the results were averaged. The investigated image is 1200 pixels wide and 676 pixels high. For each convolution mask the image was filtered by using 1, 2, 4, 8, 16 and 32 threads. The algorithm was implemented in C# language. The testing parallel architecture was Quad-Core AMD Opteron 8356 8p (32 CPU threads). Detailed results of experiments are presented in Tab. I and shown in Fig. 5 - 7. On all graphs the horizontal axis represents the number of threads and the vertical axis represents time.

TABLE I: Results of image filtering by using Laplacian filters (100 averaged measurements).

threads	average time (seconds)	percentage	standard deviation	average time (CPU ticks)	standard deviation
3×3 convolution mask					
1	17.87600	100%	0.23354	40153258	524569
2	10.18082	56.95%	0.30065	22868252	675319
4	6.03999	33.79%	0.28318	13567088	636091
8	3.84961	21.54%	0.25843	8647022	580477
16	2.93947	16.44%	0.21808	6602667	489848
32	2.42189	13.55%	0.08085	5440081	181596
5×5 convolution mask					
1	46.05641	100%	1.08065	103452367	2427362
2	24.74928	53.74%	0.26110	55592085	586472
4	14.43128	31.33%	0.26109	32415682	1912613
8	8.82084	19.15%	0.97053	19813459	2180018
16	6.05791	13.15%	0.49399	13607337	1109609
32	4.32978	9.40%	0.15660	9725596	351746
9×9 convolution mask					
1	141.72927	100%	1.55923	318353710	3502354
2	75.59142	53.34%	1.27017	169794198	2853071
4	43.08413	30.40%	4.14568	96776003	9312067
8	25.00103	17.64%	2.58287	56157567	5801661
16	17.28079	12.19%	1.61762	38816286	3633518
32	10.67084	7.53%	0.51039	23968946	1146444

Algorithm 1 Pseudocode of the parallel method of the image filtering.

Input: the image for filtering, the size of the image: width *w*, height *h*, the convolution mask, number of threads t
Calculate the number of pixels per one thread: $\lfloor \frac{w}{t} \rfloor$.
Create t threads.
for $i = 0$ to $t - 1$ **do**
  Set the range for the thread from $i \cdot \lfloor \frac{w}{t} \rfloor + 1$ to $i \cdot \lfloor \frac{w}{t} \rfloor + \lfloor \frac{w}{t} \rfloor$.
  if $i = t - 1$ **then**
    Set the range for $(t-1)$-th thread from $(t-1) \cdot \lfloor \frac{w}{t} \rfloor + 1$ to w.
  end if
  Filter the determined area according to the convolution mask.
end for
Merge all the parts into the one image.

It is possible to observe that even the division of the image into 2 rectangles speeds up significantly the calculations (53%-57% of the calculation time for one thread). In the case of the 9×9 convolution mask, the time was decreased from 141 to 75 seconds. Fig. 5 - 7 show a hyperbolic decline of the computing time. The differences between consecutive cases are getting smaller but still the most beneficial application is while using 32 threads (the maximum number of available CPU threads). The larger the convolution mask is, the more time can be saved thanks to multithreading (7-10 seconds in the case of the 3×3 mask, 22-42 seconds in the matter of the 5×5 mask and 66-131 seconds regarding to the 9×9 mask). As we have proposed multithreading method can be a very useful tool in speeding up the calculations.

VI. Conclusions and final remarks

The research has shown that proposed parallelization significantly decreases the time of calculations. The profit is best visible when all CPU threads are used. The larger the picture and the computation mask is, the more important reduction in time spent for calculations is visible. It is worth to mention that the time necessary to filter the image of a size 1200×676 pixels by using only one thread and the 9×9 convolution mask exceeds two minutes. It can be concluded that filtering larger image (for instance of a size 6000×6000 pixels or even higher) involves several minutes, especially in the case of large convolution masks, also bigger than 9×9. Striving for a significant reduction of the calculation time is essential.

The method can be further developed by examining other parallel algorithms or trying to process many images at the same time. In our future research we will also investigate this methodology in movie processing, since this application can be the most important for HD multimedia systems.

Acknowledgment

Authors acknowledge contribution to this project of the "Diamond Grant 2016" No. 0080/DIA/2016/45, and from the program "Best of the Best 3.0" both from the Polish Ministry of Science and Higher Education.

References

[1] Z. Marszałek, "Parallel fast sort algorithm for secure multiparty computation," *J. UCS*, vol. 24, no. 4, pp. 488–514, 2018.
[2] Z. Marszalek, "Parallelization of modified merge sort algorithm," *Symmetry*, vol. 9, no. 9, p. 176, 2017, DOI: 10.3390/sym9090176.
[3] J. Protasiewicz, "Inventorum: A platform for open innovation," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 10–15.

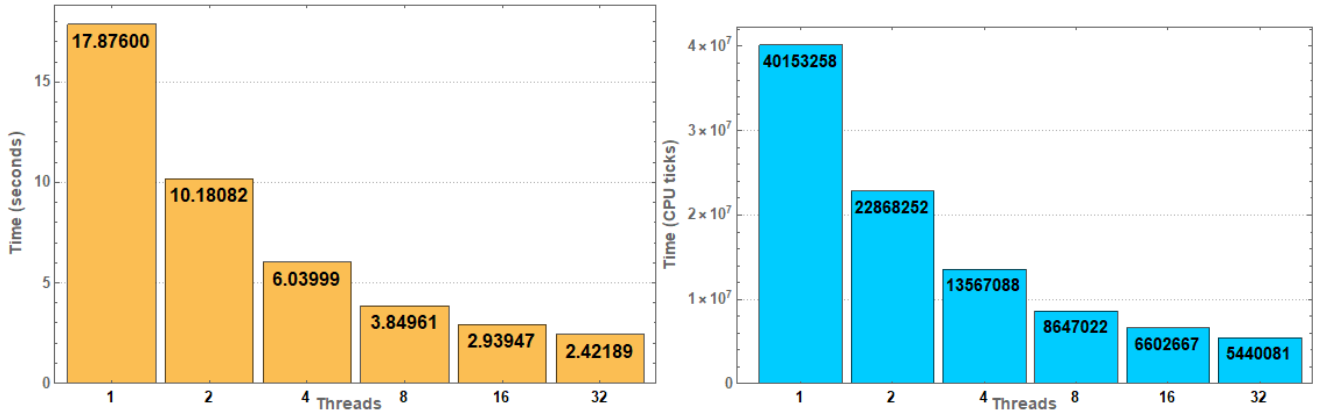


Fig. 5: Results for 3×3 convolution mask.

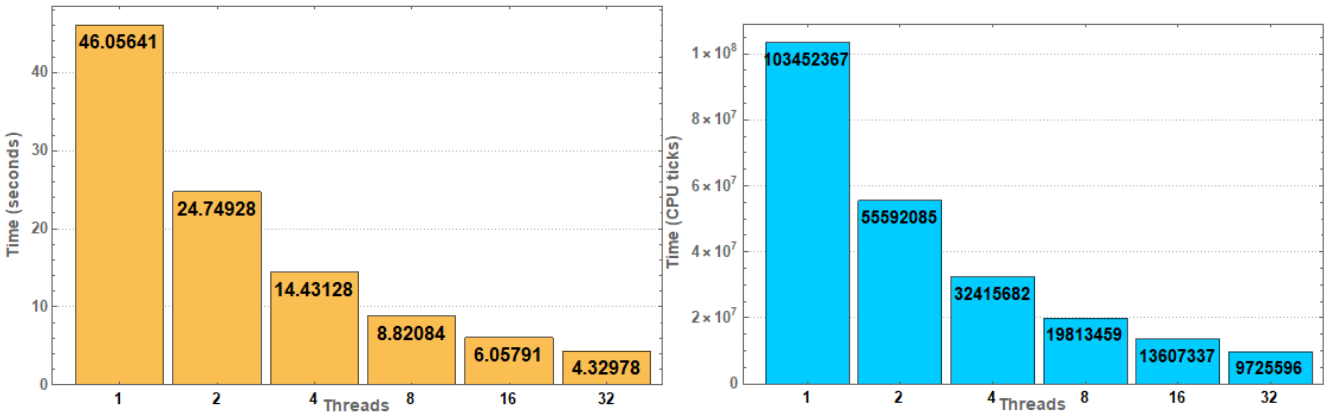


Fig. 6: Results for 5×5 convolution mask.

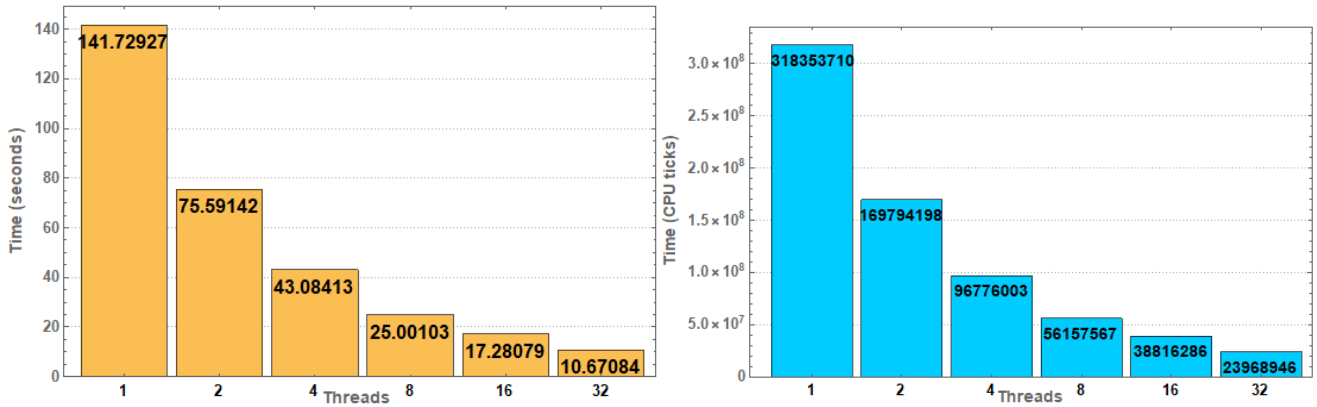


Fig. 7: Results for 9×9 convolution mask.

[4] D. D. Burdescu, L. Stanescu, M. Brezovan, F. Slabu, and D. Ebanca, "Multimedia data for efficient detection of visual objects," in *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, ser. IMCOM '17. New York, NY, USA: ACM, 2017, pp. 61:1–61:8, DOI: 10.1145/3022227.3022287.

[5] D. D. Burdescu, M. Brezovan, L. Stănescu, C. S. Spahiu, and D. C. Ebâncă, "Graph-based semantic segmentation for 3d digital images," in *2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2017, pp. 114–119, DOI: 10.1109/WAINA.2017.69.

[6] Y. Jia, C. Rong, C. Wu, and Y. Yang, "Research on the decomposition and fusion method for the infrared and visible images based on the guided image filtering and gaussian filter," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp.

1797–1802, DOI: 10.1109/CompComm.2017.8322849.

[7] S. Deniziak and T. Michno, "New content based image retrieval database structure using query by approximate shapes," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, Prague, Czech Republic, September 3-6, 2017.*, 2017, pp. 613–621, DOI: 10.15439/2017F457.

[8] ——, "Query-by-shape interface for content based image retrieval," in *8th International Conference on Human System Interaction, HSI 2015, Warsaw, Poland, June 25-27, 2015*, 2015, pp. 108–114, DOI: 10.1109/HSI.2015.7170652.

[9] R. Karumuri and S. A. Kumari, "Weighted guided image filtering for image enhancement," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 545–548, DOI: 10.1109/CESYS.2017.8321137.

[10] N. Dhengre, K. P. Upla, H. Patel, and V. M. Chudasama, "Biomedical image fusion based on phase-congruency and guided filter," in *2017 Fourth International Conference on Image Information Processing (ICIIP)*, 2017, pp. 1–5, DOI: 10.1109/ICIIP.2017.8313792.

[11] C. Chen and M. C. Stamm, "Image filter identification using demosaicing residual features," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4103–4107, DOI: 10.1109/ICIP.2017.8297054.

[12] S. K. Dewangan, "Visual quality restoration enhancement of underwater images using hsv filter analysis," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 2017, pp. 766–772, DOI: 10.1109/ICOEI.2017.8300807.

[13] E. Royer, J. Chazalon, M. Rusiñol, and F. Bouchara, "Benchmarking keypoint filtering approaches for document image matching," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 01, 2017, pp. 343–348, DOI: 10.1109/ICDAR.2017.64.

[14] R. Jain, R. Kasturi, and B. G. Schunck, *Machine vision*. McGraw-Hill New York, 1995, vol. 5.

[15] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, NJ: Prentice Hall, 2012.

[16] Z. Czech, *Wprowadzenie do obliczeń równoległych*. Wydawnictwo Naukowe PWN, 2013.

[17] T. Zieliński, *Cyfrowe przetwarzanie sygnałów: od teorii do zastosowań*. Wydawnictwa Komunikacji i Łączności, 2007.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 551–554
DOI: 10.15439/2018F141
ISSN 2300-5963 ACSIS, Vol. 15

A Multimedia Signal Processing Cloud Concept for Low Delay Audio and Video Streaming via the Public Internet

Christoph Kuhr*, Alexander Carôt†

Department of Computer Sciences and Languages,
Anhalt University of Applied Sciences, Köthen
Email: *christoph.kuhr@hs-anhalt.de, †alexander.carot@hs-anhalt.de

***Abstract*—A rehearsal environment for conducted orchestras via the public Internet requires a specialized server infrastructure, in order to provide minimal latencies between the musicians involved. In this document, we present a cloud computing concept for digital signal processing of audio and video data in realtime. Since 60 musicians and one conductor shall connect to the cloud, it is most important to distribute the signal processing and machine learning algorithms over multiple processing servers. The server infrastructure under investigation is built on top of an AVB network segment to be scalable and to maintain low latencies and jitter under heavy load. Latency and jitter are the most important properties of the realtime streams that are connected to the cloud, and are analyzed and discussed. The results have proven the proper design of the concept, but revealed the need for further optimization.**

I. Introduction

SOUNDJACK [1] is a realtime communication software that establishes up to five peer to peer connections via the public Internet. This software was designed from a musical point of view and first published in 2006 [2]. Playing live music via the public Internet is very sensitve to round trip as well as one-way latencies. Thus, the main goal of this application is the minimization of latencies and jitter, while limited by the speed of light. Participating musicians require some soft skills to tolerate the latencies none the less.

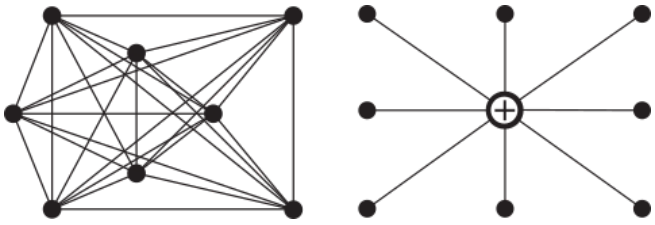


Figure 1. *Left: Peer to Peer Network Topology, Right: Star Network Topology*

A. Soundjack and fast-music

The goal of the research project fast-music is to develop a rehearsal environment for conducted orchesteras via the public Internet. Up to 60 musicians and one conductor, who are randomly distributed throughout Germany, shall be able to play together live. The central node represents the multimedia signal processing cloud under investigation, which ideally is located in Frankfurt on the Main. First investigations of round trip times suggest Frankfurt on the Main as the logical center of Germany.

Further fields of research are the transmission of Wavelet based low delay live video streams and motion capturing of the conductor. The latter shall be displayed on a holographic LED cube [3] that was developed by our project partner Symonics GmbH [4].

B. Motivation

The most important aspect for any further network and software design decision for the cloud concept is the application of digital signal processing algorithms to the audio and video streams. Examples for digital signal processing applications are audio error concealment due to UDP packetloss in the public Internet, based on a machine learning approach, or virtual room acoustics in the form of individual binaural rendered Ambisonics soundfields that simulate the musicians location inside an orchestra for a better immersion.

The second important requirement is the service time of Ethernet frames that are arriving on a serial network interface at the wide area network (WAN) side of the server cloud. In this document the network under investigation is the campus network of the university. Thus, Ethernet is also considered a WAN technology for the scope of this paper.

During the service time, no datagrams of any concurrent UDP streams can be received. Consequently any stream arriving on such a serial network interface experiences a latency, equal to accumulated latency of all streams arriving at this interface.

A scalable and extendable concept with multiple proxy servers that are connected to the same signal processing network segment is chosen over a single processing server. The Soundjack processing cloud has to be segmented accordingly. The approach that we chose is shown in fig. 2.

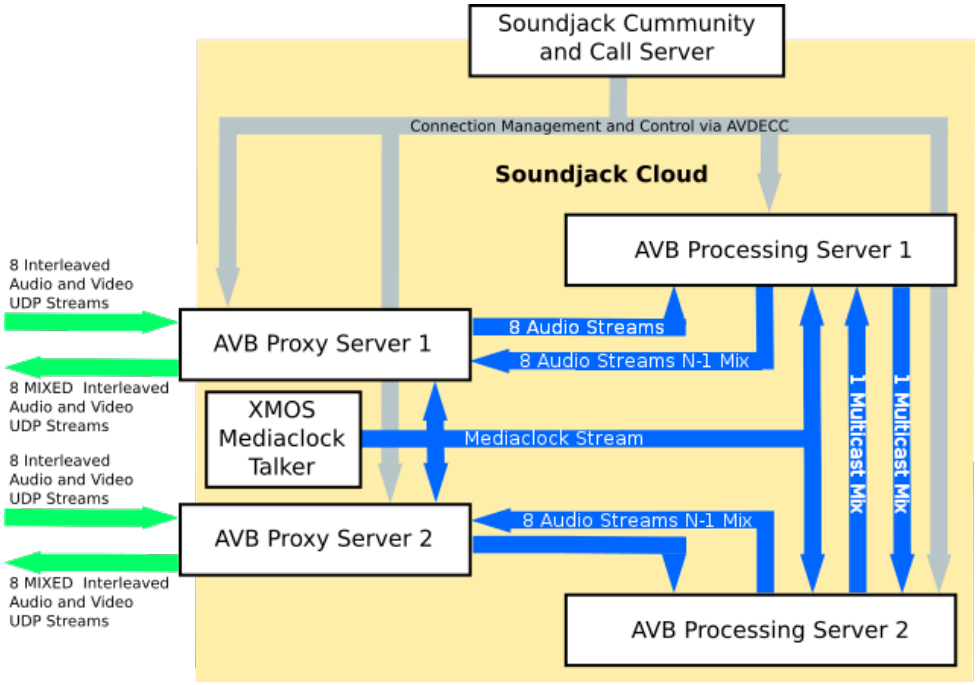


Figure 2. *Soundjack Realtime Processing Cloud Concept*

II. Soundjack Realtime Processing Cloud Concept

Audio Video Bridging / Time-Sensitive Networking (AVB/TSN) is a technology with the focus on audio and video streams in computer networks, that require realtime responsiveness. This technology is a set of IEEE 802.1 industry standards, which operate on OSI-Layer 2 [5]. These standards are:

- IEEE 802.1AS [6] - Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks
- IEEE 802.1Qat [7] - Virtual Bridged Local Area Networks - Amendment 14: Stream Reservation Protocol (SRP)
- IEEE 802.1Qav [8] - Virtual Bridged Local Area Networks - Amendment 12: Forwarding and Queueing Enhancements for Time-Sensitive Streams (FQTSS)
- IEEE 1722 [9] - IEEE Standard for Layer 2 Transport Protocol for Time-Sensitive Applications in Bridged Local Area Networks (AVTP)
- IEEE 1722.1 [10] - IEEE Standard for Device Discovery, Connection Management, and Control Protocol for IEEE 1722 Based Devices (AVDECC)

These AVB standards extend the IEEE 802.1 standard family with precise synchronization, resource reservation and bandwidth shaping. Lower latencies and jitter, the avoidance of packet bursts and bandwitdh shortage are addressed by these extensions, providing realtime responsiveness to a generic Ethernet computer network. These properties are used to ensure a constant bandwidth streaming with low latency and jitter inside the Soundjack realtime processing cloud. Thus, the Soundjack client streams can be processed inside the cloud, without interfering with each other.

AVB networks require special hardware for timestamping Ethernet frames with seperate transmission queues for each traffic class, i.e. AVB traffic with Stream Reservation (SR) classes A/B and generic Ethernet traffic. IEEE 802.1-2014 [11] defines the two distinct stream reservation (SR) classes A and B to differentiates audio and video traffic from other Ethernet traffic. SRP declares and registers resources on all switch ports along the path from the AVB talker to the AVB listener. This way, the resources to maintain an inter packet gap (IPG) of 125 μs (250 μs for SR class B) are reserved.

The two AVB server types required for the Soundjack cloud are an AVB proxy server and an AVB processing server. The AVB proxy and processing servers are each connected to the same AVB LAN segment. The media transport in this design should not exceed a round trip time of 4 ms, since AVB networks are constrained to a bounded one way latency of 2 ms [9, p. 14]. Additionally, each server is connected to a non-AVB LAN segment. The Soundjack community and call server is also connected to this generic network segment and handles the session management of the different user sessions on the Internet and of the respective AVB server sessions. Since IEEE 1722.1 AVDECC traffic is not necessarily time-sensitive, it is sufficient to use a non-AVB LAN segment for command and control purposes. The Soundjack community and call server also provides community services of Soundjack.

A. Mediaclock Server

All AVB servers subscribe to a mediaclock stream, which is supplied by an XMOS/Atterotech development board [12].

The mediaclock concept is based on the idea of the recovering from a clock stream, as it is proposed in [13].

B. Common AVB Server Software Architecture

To meet the requirements of an AVB server software, it is not sufficiant to write a multiprocessing and multithreading application in C. The hardware support for AVB requires a properly configured and optimized OS. The Linux kernel can be patched to operate in realtime mode [14], a Linux mainline kernel 4.8.6 was configured, patched with the corresponding realtime patch 4.8.6-rt5 and compiled.

The AVB talkers running on the system need to use hardware queues of the network interface to utilize the FQTSS mechanism for enqueueing AVTP packets. A detailed description of the software architecture of the two different server configurations can be found in [15].

C. AVB Proxy Server

In contrast to the public Internet where IP packets are forwarded on best effort, the Soundjack processing cloud provides a fully managed and controlled Ethernet network with AVB support. The traffic shaping property of AVB prevents the Soundjack cloud network from bursty traffic by means of a credit-based bandwidth shaper. The proxy server is used as a wave trap to break down large and erratic UDP datagrams into more and smaller, but constant AVTP packets. Thus, the stream packets can travel inside the Soundjack cloud network in a deterministic fashion, managed by the credit-based bandwidth shapers.

The AVB proxy server receives and transmits UDP streams from and to Soundjack Clients, that were assigned by the Soundjack session server. A maximum of eight streams is assigned to one AVB proxy server to keep the latency introduced by the service times low. The UDP streams received on the WAN interface are fragmented, because they need to be transmitted in the AVB LAN segment at another bitrate with a different payload. A UDP datagram of such a stream contains 256, 512 or 1024 Bytes of compressed or raw audio data. AVB implements its traffic shaping based on the idea of constant bitrate streaming. The resulting AVTP stream is sent from the AVB proxy server to the AVB processing server, which processes the eight streams and sends them back as AVTP with the same, but processed payload. Inside the Soundjack cloud a constant link capacity utilization per stream is of paramount importance to maintain the deterministic behavior of the AVB LAN segment. To achieve constant link capacity utilization, the payload of each UDP datagram needs to be properly fragmented into multiple AVTP packets. IEEE 802.1Qav [8, p. 44] requires an AVB talker to send an AVTP packet every $125\ \mu s$ for a class A stream and $250\ \mu s$ for a class B stream. This inter packet gap (IPG) is necessary to maintain a constant sample flow at 48kHz sample rate: $6/48\ kHz = 125\ \mu s$, and AVTP packet rate of $48\ kHz/6 = 8\ kHz$. A single stream may also contain multiple audio channles, thus an AVTP packet contains six samples per audio channel. Inside the Soundjack processing cloud however, all AVTP streams contain two channels - a stereo stream. The proxy server talker instances mix the possible eight, four and two channels down to (or one channel up to) two channels per received UDP datagram, which reduces the payload size for the AVTP stream. The return stream of the Soundjack cloud has always 64 samples per UDP datagram for two audio channels - a stereo mix. The stereo mix from the samples of the UDP datagrams needs to be distributed over multiple AVTP packets. Soundjack uses a sample rate of $48\ kHz$ exclusively and all channels are mixed down (or up) to two channels by the proxy server. The stereo mix channel count, the sample rate and the sample encoding determine the actual AVTP payload size of $48\ bytes$. Since AVTP packets arrive with the IPG of $\Delta t = 125\ \mu s$ at the proxy listener buffer, a constant latency is introduced by waiting for the correct amount of AVTP packets required to properly fill the UDP datagram.

Apart from the audio samples, the UDP stream also contains interleaved video data, which is described in [16].

D. AVB Processing Server

The AVB processing server receives the audio and video streams, originating at the clients as AVTP streams with a constant packet rate of $8\ kHz$. It provides signal processing facilities for audio and video processing.

The JACK [17] audio server is deployed as infrastructure for the audio signal processing stage. It is a professional and open source audio server to share sample accurate audio data between different applications. A large number of signal processing applications and algorithms are available for JACK.

As off now, an audio multicast mixing application, to mix all streams that are connected to the Soundjack cloud, is deployed. A detailed description can be found in [18]. Further signal processing application, as mentioned in the introduction, are still under development.

III. Evaluation

A first evaluation was done with a single UDP audio stream that enters the Soundjack cloud via the WAN network interface of the AVB proxy server in the campus network. The proxy server forwads the stream as AVTP stream to the AVB processing server. After processing the audio signals, the processing server returns the AVTP stream to the AVB proxy server, which in turn constructs an UDP stream to return to the Soundjack client. The latency was measured with a scope connected to the digital-analog and analog-digital converters of the Soundjack clients audio interface.

Furthermore, the arrival timestamps of the UDP send and return stream and the different AVTP streams have been captured with the packet analyzer Wireshark. The probability density function of the difference to the previous timestamp of the AVTP streams is the bounded IPG of $125\ \mu s$ defined by AVB. In the context of the UDP streams, this probability density function gives a measure for the audio quality of the stream - whether it has high or low jitter.

IV. Discussion

A generated sine wave with a frequency of 1 kHz was transported through the entire cloud and is late by 16 ms round trip time. At some point in time however, some buffer seems to underrun which could not be located yet. In this case, the latency of the return stream is stable at 316 ms round trip time.

Both, the send and the return stream, show the expected IPG for the tested payload sizes. The send stream maintains an IPG of 2.666 ms for 256 bytes (128mono samples/48 kHz) and the return stream maintains an IPG of 1.333 ms for 256 bytes (64stereo samples/48 kHz), respectivly.

An round trip time of 16 ms (8 ms end-to-end) for the Soundjack client streams, still violates the end-to-end delay limit of 2 ms formulated in [9, p. 14]. Additional latencies that are introduced by the networks and not compensated yet are for example some minimal portion of undeterministic network behavior by the campus network, buffer synchronization latencies, packet de- and fragmentation inside the cloud. On the audio processing side we have to consider drift of the JACK audio server, because it is not phase locked to the incoming mediaclock stream. When the mediaclock is very early and the audio interface hardware interrupt is very late in relation to each other, the drift between those two interfaces can lead to a worst case latency of $\Delta t_{max} = 1/48\ kHz = 20.833\ \mu s$. This latency is much less significant than the latencies introduced by the signal processing algorithms that shall be implemented. A measurement of the latency introduced by the multicast mixing application has still to be done.

V. Conclusions

The Soundjack cloud prototype is not fully tested yet, but the evaluations in this paper show the proper operation of the presented concept. Analysis has shown that the AVB requirements could be mostly fullfilled. Further sources for the remaining latency, besides the actual algorithmical latency of the digital signal processing involved, have to be exposed.

VI. Future Work

The software has to be further optimized to reliably meet the AVB constrains.
Furthermore, the AVB server software behavior requires evaluation under heavy load with up to eight possible streams. In parallel to the creation of this paper further signal processing applications have been developed which will be integrated into the streaming process. The jitter and latency behavior under heavy load with signal processing applications in place will be evaluated in the future. Those evaluations will mainly focus on the task scheduling precision in tems of meeting the calculated task deadlines with the EBF-CBS scheduler.

Signal processing and machine learning application are under development an might be ready to be tested prior to a deployment in the public internet.

As soon as the proper networking operations of the cloud are verified, measurements will be performed in the public Internet. We will then deploy the Soundjack cloud in Frankfurt on the Main and test in a real world environment. A comparison between an IPv4 and an IPv6 deployment will be done as well.

VII. Acknowledgements

fast-music is part of the fast-project cluster (fast actuators sensors & transceivers), which is funded by the BMBF (Bundesministerium für Bildung und Forschung).

References

[1] (2018, Apr. 23) Soundjack - a realtime communication solution. [Online]. Available: http://http://www.soundjack.eu

[2] A. Carôt, U. Krämer, and G. Schuller, "Network music performance (nmp) in narrow band networks," in *in Proceedings of the 120th AES convention, Paris, France*. Audio Engineering Society, May 20–23, 2006.

[3] A. Carôt, S. Ebeling, C. Hoene, P. Platz, and H. Loridan, "Glass panel displays with addressable leds," in *Mensch und Computer 2018 (MUC2018)*. Dresden, Germany: Gesellschaft für Informatik, Technische Universität Dresden, Sep. 2–5, 2018.

[4] (2018, Apr. 23) Symonics gmbh. 72144 Dusslingen, Germany. [Online]. Available: http://symonics.de

[5] H. Zimmermann, "Osi reference model -the iso model of architecture for open systems interconnection," in *IEEE Transactions on Communications, Vol. 28, No. 4*, Apr. 1980, pp. 425–432.

[6] *Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks*, IEEE Std. 802.1AS, Mar. 2011.

[7] *Virtual Bridged Local Area Networks - Amendment 14: Stream Reservation Protocol (SRP)*, IEEE Std. 802.1Qat-2010, Sep. 2010.

[8] *Virtual Bridged Local Area Networks - Amendment 12: Forwarding and Queuing Enhancements for Time-Sensitive Streams*, IEEE Std. 802.1Qav-2009, Jan. 2010.

[9] *Layer 2 Transport Protocol for Time-Sensitive Applications in Bridged Local Area Networks*, IEEE Std. 1722, May 2011.

[10] *Device Discovery, Connection Management, and Control Protocol for IEEE 1722 Based Devices*, IEEE Std. 17 221, Aug. 2013.

[11] *(Revision of IEEE Std 802.1Q-2011) - IEEE Standard for Local and metropolitan area networks–Bridges and Bridged Networks*, IEEE Std. Std 802.1Q-2014, Dec. 2014.

[12] (2018, Apr. 23) Xmos ltd. / attero tech inc. [Online]. Available: http://www.atterodesign.com/cobranet-oem-products/xmos-avb-module/

[13] H. Weibel and S. Heinzmann, "Media clock synchronization based on ptp," in *Audio Engineering Society Conference: 44th International Conference: Audio Networking*, Nov 2011. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16146

[14] J. Kacur, "Realtime kernel for audio and visual applications," in *Proceedings of the Linux Audio Conference 2010*. Wittenburg, DE: Red Hat, Apr. 2010.

[15] C. Kuhr and A. Carôt, "Software architecture for a multiple avb listener and talker scenario," in *Proceedings of the Linux Audio Conference 2018*. Berlin, Germany: Linuxaudio.org, Jun. 7–10, 2018.

[16] A. Carôt and G. Schuller, "Towards a telematic visual-conducting system," in *AES 44th International Conference, San Diego, USA*. Audio Engineering Society, Nov. 18–20, 2011.

[17] (2018, Apr. 23) Jack audio connection kit. [Online]. Available: https://jackaudio.org

[18] C. Kuhr, T. Hofmann, and A. Carôt, "Use case: Integration of a faust signal processing application in a livestream webservice," in *Proceedings of the 1st International Faust Conference 2018*. Mainz, Germany: Johannes Gutenberg-Universität Mainz, Jul. 17–18, 2018.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 555–559
DOI: 10.15439/2018F279
ISSN 2300-5963 ACSIS, Vol. 15

Query by Approximate Shapes Image Retrieval with improved object sketch extraction algorithm

Stanisław Deniziak
Kielce University of Technology
al. Tysiaclecia Panstwa Polskiego 7, 25-314 Kielce, Poland
Email: s.deniziak@tu.kielce.pl

Tomasz Michno
Kielce University of Technology
al. Tysiaclecia Panstwa Polskiego 7, 25-314 Kielce, Poland
Email: t.michno@tu.kielce.pl

***Abstract*—In this paper a new Content Based Image Retrieval based on a sketch method was proposed. The main idea of the algorithm is based on decomposing an object into predefined set of shapes (primitives): line segments, polylines, polygons, arches, polyarches and arc-sided polygons. All primitives are stored as a graph in order to store the mutual relations between them. Graphs are stored in a tree-based structure which allows fast querying. As an improvement to the algorithm, a conversion to the HSL color space was proposed in order to detect primitives more accurately. Moreover, computing all line slopes in relation to the object oriented bounding box was also proposed. Additionally, in order to better detect objects present in images, the usage of Edge Boxes algorithm was proposed.**

I. Introduction

MULTIMEDIA databases are becoming more and more popular. Most often they store huge number of images which causes the need of effective storage, processing and query methods. This is becoming an important problem because they are used more often in everyday life, from searching for information in the internet to authorising, recognizing and monitoring systems. Moreover, most of social media portals stores images as the part of provided content or even are oriented only on providing images, thus they also need effective methods to manage and provide data to users.

There are many different methods in the area of image retrieval from multimedia databases. Most of the algorithms may be grouped into three categories: the Keywords Based Image Retrieval (KBIR) algorithms, the Content Based Image Retrieval (CBIR) algorithms and the Semantic Based Image Retrieval (SBIR) algorithms.

The KBIR algorithms use textual annotations in order to describe objects present in the image. Then, during the query they are compared with keywords typed by the user [1]. The CBIR algorithms use content present in the image in order to perform queries. Most often global statistical features are used (e.g. contrast, entropy [2] or a normalized histogram of colors [3]) or grouping similar pixels into regions in order to construct graphs which are then compared using e.g. Maximum Likelihood [4]. There are also methods which use sketches instead of images as a query(e.g. [5]). The SBIR algorithms tries to minimize the difference between the data present in the image and the information which can be deduced by a human [6], [7]. The text may be classified e.g. by one of the methods described int [8]. There are methods which use textual queries and graphical queries.

This paper presents a new Content Based Image Retrieval method from multimedia database which is based on querying by a sketch. The main idea of the algorithm is based on representation of objects using a set of predefined shapes, called primitives: line segments, arches, polylines, polygons, polyarches (a chain of connected arches) and arc sided polygons (a looped chain of connected arches). For each type of a primitive, there are defined attributes which describes them, e.g. a line slope for a line segment or an angle for an arc. The primitives may be extracted from images using proposed approach based on HSL color space segmentation. All primitives are connected into a graph which stores all relations between them. Graphs are stored in a tree-based database structure which gathers similar objects graphs in the same subtrees, but without fast tree height grow. In comparison to our previous works, in this paper we focused on improving the primitives extraction algorithm and the precision of the results. The proposed approach provides easy graphical queries for users, both using sketches drawn by themselves (without need of high drawing skills) and example images. All similar graphs are stored in congruent nodes, thus the querying is faster than in e.g. linear data structure. Also the structure allows performing queries in parallel which also improves the computation time.

The paper is organised as follows: the first section is an introduction. Next section presents the proposed object representation. The third section describes the shapes extraction algorithm with improvements and the fourth possible usages of Edge Boxes algorithm. The fifth section presents the experimental results. The next section is the summary and future research section. The last one section is a list of references used in this paper.

II. The object representation

During previous stages of our research, we found that most objects and images can be described using sketches. This method is very efficient, because there is no need of example image, but if existent, it can be processed in order to extract a sketch. Moreover, sketch representation of objects may be used when there is no full knowledge about searched matter, e.g. when only a front part of a car is known.

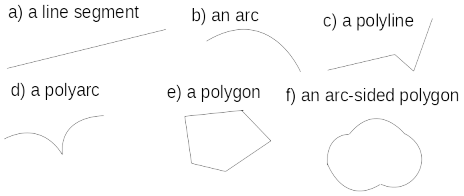


Fig. 1. The proposed primitives: a) a line segment, b) an arc, c) a polyline, d) a polyarc, e) a polygon, f) an arc-sided polygon

The sketch may be represented by an image (e.g. using black pixels as sketch edges [9]), an object outline (e.g. as in [5]) or, as proposed in our research, as a set of predefined shapes. The advantage of our approach is the simplicity to draw a sketch by users, ability to store additional information for each shape (e.g. information about material or color). The predefined, simple shapes are called primitives. During our research we noticed, that line segments (Fig. 1 a)) and arches (Fig. 1 b)) are suitable for describing a sketch of most objects. Moreover, during experiments, we realised that storing the information about connected segments and arches improves the efficiency of the algorithm. Thus, additional primitives are proposed:

- based on line segments: polylines (Fig. 1 c)) and polygons (Fig. 1 e))
- based on arches: polyarches (a chain of connected arches, Fig. 1 d)) and arc-sided polygons (a looped chain of connected arches, Fig. 1 f))

Since all primitives are based on line segments or arches, they are called base primitives, afterwards all primitives created using them are called complex primitives.

Each primitive is defined by its type and an attribute or set of attributes, which are defined as follows: a line segment attribute is its line slope, an arc attribute is its angle, a polyline and a polygon attributes are number of segments and their line slopes, a polyarc and an arc-sided polygon attributes are number of segments and their angles.

During our research we realised that not only the information about primitives is needed but also an information which shapes are connected. Thus, such an information is stored using a graph where nodes are used to store primitives and edges used to store connections between them. This approach is similar to graphs of regions in region-based CBIR algorithms. Moreover storing the information about mutual positions between connected primitives also improves the efficiency of the image retrieval from the multimedia database [10]. Therefore, for each connection, there is stored the information about positions using the geographical windrose (N, S, E and W directions). The example of the graph is shown in the Fig. 2.

Since an image may contain many objects, in order to clearly emphasize that they are separated in a graph, a structure called complex shape was introduced [10]. It may be used optionally or mandatory based on types of stored images in the multimedia database.

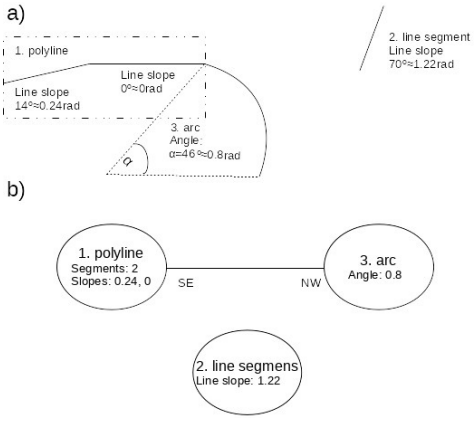


Fig. 2. The example of the graph which contains 3 primitives: a) primitives and their attributes (a polyline, an arc and a line segment), b) the graph: a polyline (1) and an arc (3) are connected, no. 3 lies on the right bottom (SE) of no. 1 and consequently no. 1 lies on the top left (NW) of no. 3; the node no. 2 does not have any connection

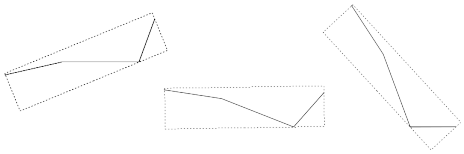


Fig. 3. The example polyline object with its Object Oriented Bounding Box (marked by dotted lines) in three different orientations.

In order to improve the algorithm resistance to different objects orientations, as an addition to the previous works, we propose to store primitives attributes in axis independent manner. During the stage of attributes values computing, line slopes may be computed in relation to the longest side of the Object Oriented Bounding Box built on top of the object. The example bounding box of a polyline is shown in the Fig. 3. Such an approach allows comparisons of differently oriented primitives which are strictly the same or very similar with high precision. Without such solution, there may be some problems with proper image retrieval from the database when users draw a sketch which is e.g. rotated in comparison with sketches stored in the database.

During previous experiments we found that storing the information about mutual positions of connected nodes highly improves the precision of the image retrieval algorithm results. Due to that fact, we propose to store the same information also for each pair of nodes (not only connected), similarly to coincidence matrix used in graphs. In this paper we test if such an approach improves the precision of the results.

III. The Image Shapes Extraction Algorithm Based on HSL Image Representation

In this paper we propose to enhance the shapes extraction algorithm in comparison with the previous version used in [1]. The main idea of the improvement is based on converting an image into a HSL color space. Due to the conversion three images are created: first with only hues present in the image,

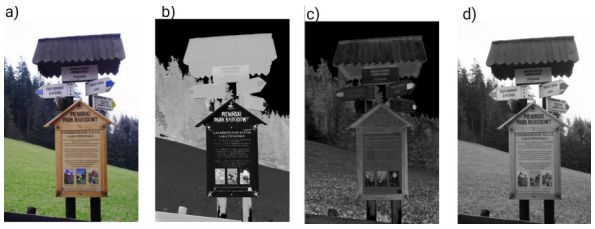


Fig. 4. The representation of image using HSL color space: a) initial image, b) hue channel, c) saturation channel, d) lightness channel which is most often the same as grayscale image representation.

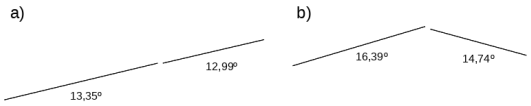


Fig. 5. The example of line segments merging with close endings: a) two segments with line slopes close to 13° which can be merged, b) two segments with completely different line slopes.

the second one with lightness information and the last one which stores saturations of different areas (the example image representation is shown in the Fig. 4). Thanks to that, the algorithm is able to detect line segments and arches with the same hue, the same lightness level and the same saturation.

Similarly to other objects detection algorithms, morphological operations should be performed on the image in order to remove unnecessary details or noise.

The first algorithm step is the detection of base primitives - line segments and arches. Firstly all line segments are detected in hue channel using Line Segment Detector algorithm. Next, for each channel thresholding with different values is performed in order to create uniform areas of similar pixels and reduce the slight differences on the same shapes. After that, line segment detection is performed again for each channel. Thanks to that more line segments are detected. Next all detected line segments are merged if possible, checking if they have the same line slope and connection of very near endings (Fig. 5). As a next step, all segments shorter than defined threshold are removed from the list. The algorithm is shown in the Alg. 1.

When all line segments are found, then the list is searched in order to find the candidates of arches. In order to find line segments which may construct an arc, we check if in the list there is a chain of connected line segments with length equal or greater than 3. Then, if a chain is found, angles between each segments are checked if they are equal or very similar. Moreover a test if their values are in the range $(0°, 180°)$ (Fig. 6) is performed. Additionally, lengths of segments are checked in order to detect if they are very similar. The arc detection algorithm is shown in the Alg. 2. In the practical implementation, the algorithm is assisted with Circular Hough Transform in order to improve the detection of circles.

The last shapes extraction algorithm step is the creation of more complex primitives defined as follows: polylines, polygons, polyarches and arc-sided polygons. Firstly all detected

Algorithm 1 Line segment detection algorithm

Ensure: img - the image which has to be processed, $lsList$ - list which stores line segments
$img_H \leftarrow$ hue channel of img;
detect all line segments in img_h and add them to the $lsList$
for each $img_x \in$ `H,L,S channel of` img **do**
  `process` img_x `as follows:`
  for each $pixel$ of img_x **do**
    round $pixel\ value$ to the nearest $threshold$ value
  end for
  detect all line segments in img_x and add them to the $lsList$
end for
for each $ls_1 \in lsList$ **do**
  for each $ls_2 \in \{\{lsList\} \setminus \{ls_1\}\}$ **do**
    if ls_1 and ls_2 endings are close enough **then**
      $angle_{LS1} \leftarrow ls_1\ line\ slope$
      $angle_{LS2} \leftarrow ls_2\ line\ slope$
      if $angle_{LS1}$ and $angle_{LS2}$ are similar enough **then**
        merge ls_1 and ls_2
      end if
    end if
  end for
end for
for each $ls \in lsList$ **do**
  $len \leftarrow length(ls)$
end for

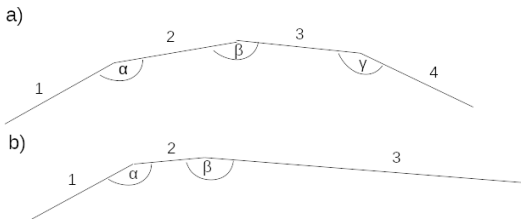


Fig. 6. Different chains of line segments a) a chain from which an arc may be created: segments lengths are very similar (1=146, 2=145, 3=146,8, 4=145,8) and angles α, β, γ are equal to 160° b) the chain from which an arc cannot be created: segments lengths are different (1=138,7, 2=85,4, 3=371) and angles are not equal or very close ($\alpha = 157°$, $\beta = 170°$)

line segments are processed and checked for chains. After chains detection, they are tested in order to detect looped chains from which polygons are created. Then, all others are transformed into polylines. Similarly, founded arches list is processed: firstly chains of arches are detected and then looped ones are transformed into arc-sided polygons and all others into polyarches.

IV. Shapes Extraction Improvement using Edge Boxes

The shapes extraction algorithm performance could be improved by a usage of object proposal algorithms. One of the

Algorithm 2 Arches detection algorithm

Ensure: $lsList$ - list which stores line segments; $chains$ - list of chains of line segments; $archesList$ - list which stores arches
find all connected line segments from $lsList$ longer than 3 segments and add them to $chains$
compare lengths of all segments in each $chain \in chains$
if lengths are similar enough **then**
  for each $ls \in chain$ **do**
    compute an angle between ls and next segment from the same chain
  end for
  if angles are similar enough **then**
    create new arc based on chain
    compute arc angle and center point
    add arc to $archesList$ and remove all $chain$'s line segments from $lsList$
  end if
end if

most suitable algorithm for our approach is the Edge Boxes [11]. It is based on the idea that when there are many contours enclosed by bounding box, there is a high likelihood that it contains an object. As an edge extractor a Structure Edge detector is used. All contours are examined using a sliding window approach. In order to improve the computation time, efficient data structures are used. The result of the algorithm is a set of bounding box objects proposals.

In this paper we propose two approaches using Edge Boxes:

- first - where shapes extraction is only performed for edges in the found bounding boxes
- second - where information about object bounding boxes is used as an assistance to the shapes extraction algorithm

The first approach highly decreases the number of shapes detection areas which should highly improve the computation time. Moreover, the primitives extracted from each bounding box should be stored in a one complex shape which should improve the query comparison time and precision.

The second approach detects primitives in the whole image area and then uses Edge Boxes proposals in order to strength the links between shapes in the same bounding box. This is performed by adding connections between unconnected nodes in a graph. Due to that fact, all shapes which are meant to be in a one object should be connected. Such an approach should also improve the precision of comparisons with graphs in the database but may decrease the computation time.

V. The database structure

The database structure is based on a tree which stores similar object's graphs in the same part of a subtree. Two types of tree nodes are defined:

- common graph nodes
- data nodes

The common graph nodes are used in order to store the common parts of graphs which are stored in its children. For example, if there are two children with a car and a bicycle graphs, as a common graph, wheels are used. The common graph nodes does not store any image information and are only used to check if a whole subtree should be tested or abandoned during the query.

The data nodes are used to store graphs of objects, images and metadatas connected with them. In order to reduce the height of the tree, all similar graphs are stored in the same node in a structure called a slice. A slice is a vector of very similar graphs where the first element is the most similar to the parent common node graph and the last one the least. There may be more than one slices in the data node in order to provide the ability to process queries in parallel.

The tree database structure and operations which may be performed (adding, deleting and querying graphs) were described more detailed in [1].

TABLE I
THE COMPARISON OF PRECISION AND RECALL RESULTS FOR BICYCLE IMAGES (NORMAL, ALIGNED TO THE X-AXIS AND SLIGHTLY ROTATED) FOR COMPUTING LINE SLOPES VALUES IN RELATION TO THE X-AXIS AND TO THE LONGEST SIDE OF THE OBJECT ORIENTED BOUNDING BOX (OOB).

object	TOP 10 precision		Recall	
	X-axis	OOB	X-axis	OOB
normal	0.5	0.9	0.26	0.63
rotated	0.3	0.7	0.16	0.42

VI. Experimental results

In order to examine the proposed approach, an experimental application was written in C++ language for database implementation and python language with Django, HTML and JavaScript for GUI. The database stores images of three types of objects: cars, bicycles and motorbikes.

Firstly the computation of primitives attributes in relation to the object oriented bounding boxes was examined. In order to compare the results, two commonly used coefficients were used:

$$precision = \frac{number\ of\ relevant\ results\ images}{total\ number\ of\ results\ images} \quad (1)$$

$$recall = \frac{number\ of\ relevant\ results\ images}{total\ number\ of\ relevant\ images\ in\ the\ database} \quad (2)$$

The experimental results are presented in the Table I. As can be seen, when line slopes values are computed in relation to the longest side of the object oriented bounding box, both precision and recall reach much higher values in comparison to the previous version where slopes are computed in relation to the X axis. It can be noted that even if there are used the same images but with rotations, there are still some differences in precision and recall values for them. This problem is caused i.e. by inaccuracies during shapes extraction stage and will be taken into consideration in the future research.

Another tests were performed in order to evaluate the usage of the additional information about mutual positions of all nodes during queries. The results are presented in the Table II. The experiments does not prove that adding such an

TABLE II
THE COMPARISON OF PRECISION AND RECALL RESULTS FOR BICYCLE IMAGES (NORMAL AND ROTATED) AND A CAR IMAGE FOR PREVIOUS ALGORITHM VERSION AND A VERSION WITH INFORMATION ABOUT MUTUAL POSITIONS OF ALL NODES IN THE GRAPH.

object	TOP 10 precision		Recall	
	previous	mutual positions	previous	mutual positions
bike normal	0.5	0.6	0.26	0.79
bike rotated	0.3	0.4	0.16	0.95
car	0.8	0.6	1	1

TABLE III
THE COMPARISON OF NUMBER OF DETECTED PRIMITIVES FOR PREVIOUS VERSION (RGB) AND VERSION WHICH USES HSL COLOR SPACE

primitive	bike		car		motor	
	RGB	HSL	RGB	HSL	RGB	HSL
line segments	24	36	73	54	64	71
arches	2	3	2	2	2	3
polylines	0	3	8	11	2	10

information improves the precision of the query results. For the bicycles images the precision values are only slightly higher, but for the car it is lower than in previous version. Moreover, the computation time was much longer. Contrary, for bicycle images, the recall values are much higher than in previous version. Due to that fact, the usage of additional information about mutual positions of all nodes in a graph is doubtful and more tests should be performed.

Another types of experiments were performed in order to evaluate the shapes detection improvement by a usage of HSL color space. The results are presented in the Table III. There can be noted that for all objects the usage of HSL color space increased the number of detected primitives. The recognition of the third arc in a bicycle is not a mistake, because not only wheels were detected, but also a gearwheel. In the motorbike image also other parts were detected as circles correctly. It can be seen that the number of polylines is much higher when using HSL color space. This is caused by higher number of detected base line segments and more precise detection which allowed to create more complex primitives.

Due to the limited time, practical experiments for usage of the Edge Boxes algorithm will be performed as a future research.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper a new Content Based Image Retrieval algorithm based on a sketch method was proposed. The main idea of the algorithm is based on decomposing an object into predefined set of shapes (primitives): line segments, polylines, polygons, arches, polyarches and arc-sided polygons. All primitives are stored as a graph in order to store the mutual relations between them. As an improvement to the algorithm, a conversion to the HSL color space was proposed in order to detect primitives more accurately. Moreover, computing all line slopes in relation to the object oriented bounding box was proposed which also increased the precision of the results. The experiments which were performed proved that such propositions increased both precision and recall values. Additionally, in order to better detect objects present in images, the usage of Edge Boxes algorithm was proposed which will be tested as a future research. The proposed results improvement by the information about mutual positions of all nodes is doubtful and more tests should be performed.

The future research includes testing the usage of Edge Boxes and deciding which proposed approach would be more suitable for most situations. Moreover, more tests should be performed for all proposed improvements in this paper. Additionally, the computation time should be also tested for different algorithm versions. Another tests may be performed in order to examine different database structure implementations (e.g. using SD2DS data structures and relational databases). Moreover, the primitives detection algorithm could be improved, e.g. using approaches proposed in [12]. Also other graph comparison algorithms may be evaluated, e.g. using the optimization methods with constraints [13].

REFERENCES

[1] S. Deniziak and T. Michno, "New content based image retrieval database structure using query by approximate shapes," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2017. doi: 10.15439/2017F457 pp. 613–621.

[2] H. P. Kriegel, P. Kroger, P. Kunath, and A. Pryakhin, "Effective similarity search in multimedia databases using multiple representations," in *2006 12th International Multi-Media Modelling Conference*, 2006. doi: 10.1109/MMMC.2006.1651355. ISSN 1550-5502 pp. 4 pp.–.

[3] M. Mocofan, I. Ermalai, M. Bucos, M. Onita, and B. Dragulescu, "Supervised tree content based search algorithm for multimedia image databases," in *2011 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, May 2011. doi: 10.1109/SACI.2011.5873049 pp. 469–472.

[4] C. Y. Li and C. T. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 447–456, April 2008. doi: 10.1109/TMM.2008.917421

[5] S. Parui and A. Mittal, "Sketch-based image retrieval from millions of images under rotation, translation and scale variations," *CoRR*, vol. abs/1511.00099, 2015. [Online]. Available: http://arxiv.org/abs/1511.00099

[6] H. H. Wang, D. Mohamad, and N. A. Ismail, "Approaches, challenges and future direction of image retrieval," *CoRR*, vol. abs/1006.4568, 2010.

[7] A. Singh, S. Shekhar, and A. Jalal, "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images," in *Information and Communication Technologies (WICT), 2012 World Congress on*, Oct 2012. doi: 10.1109/WICT.2012.6409187 pp. 817–821.

[8] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Systems with Applications*, 2018.

[9] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database-query by visual example," in *11th IAPR International Conference on Pattern Recognition, Vol.I. Conference A: Computer Vision and Applications*, Aug 1992, pp. 530–533.

[10] S. Deniziak and T. Michno, "Content based image retrieval using query by approximate shape," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sept 2016, pp. 807–816.

[11] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014. ISBN 978-3-319-10602-1 pp. 391–405.

[12] M. Woźniak and D. Połap, "Adaptive neuro-heuristic hybrid model for fruit peel defects detection," *Neural Networks*, vol. 98, pp. 16 – 33, 2018. doi: https://doi.org/10.1016/j.neunet.2017.10.009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608017302526

[13] P. Sitek and J. Wikarek, "A hybrid programming framework for modeling and solving constraint satisfaction and optimization problems," *Scientific Programming*, vol. 2016, 2016. doi: 10.1155/2016/5102616

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 561–568
DOI: 10.15439/2018F73
ISSN 2300-5963 ACSIS, Vol. 15

Study of the Influence of a Light Source on the Result of the Reconstruction of the Flaccid Membrane of an Artificial Heart

Krzysztof Murawski
Military University of Technology
Urbanowicza Str. 2,
00-908 Warsaw, Poland,
IEEE Member # 92707852
Email: krzysztof.murawski@wat.edu.pl

Wojciech Sulej
Military University of Technology
Urbanowicza Str. 2,
00-908 Warsaw, Poland,
Email: wojciech.sulej@wat.edu.pl

***Abstract*— The article presents the influence of the used light sources (illuminators) on the result of the 3D reconstruction of the flaccid membrane surface shape. The research was carried out using a flaccid membrane developed by the authors for the pulsating, extracorporeal, pneumatic heart support pump model. Subsequently, the tested illuminators are an integral part of the sensor for measuring the stroke volume (SV) and the cardiac output (CO). Due to the operation principle of the computer video measurement method SV and CO developed by authors, the consistency of the obtained reconstruction of the flaccid membrane surface shape with the actual (reference) shape of the membrane is fundamentally important. For this reason, the influence of the developed lighting constructions was experimentally verified. For this purpose, a laboratory station and a rigid equivalent of the flaccid membrane were created, thus ensuring the reproducibility of the experiments. For LED-based illuminators, printed circuit boards were also designed, which eliminated the impact of changes in the position of light sources. Thanks to this, it was possible to determine the actual impact of the illuminator's construction on the obtained reconstruction result. The tests were performed for several selected light sources. The following were taken into account: LED diodes operating in the visible band, near infrared LED diodes, optical fibers and electroluminescent film.**

I. Introduction

The publication is the result of studying the effect of lighting on the result of the 3D reconstruction of the flaccid membrane surface shape. Reconstruction of the surface shape was carried out with a proprietary video sensor used to measure the instantaneous stroke volume (SV) and the cardiac output (CO) of the pulsating, extracorporeal, pneumatic myocardial support pump (ventricular assist devices - VAD), described in works [1 - 5]. One of the known VAD constructions is a heart pump made as part of the *"Polish Artificial Heart"* program, discussed in [6] and shown in Fig. 4. Based on its description, the first proprietary VAD model shown in Fig. 1 (currently no longer used) has been produced. The research discussed in this paper was carried out using the new pump models shown in Fig. 2 and Fig. 3. They represent modification made by the authors of the original VAD design. The model shown in Fig. 2 was made using a spatial printing technique. The pump shown in Fig. 3 was made using a casting technique. The presented models, except for the blood chamber, have a pneumatic chamber, which view is shown in Fig. 6. The chambers are separated by a non-elastic, flaccid membrane composite visible in Fig. 5 and Fig. 7. The up/down movement of membrane causes the blood to be pushed or sucked to/from the blood chamber. The membrane, along with the markers on it, is illuminated by the illuminator. At the same time, it is observed by the sensor camera. Reconstruction of the surface shape of the membrane, Fig. 2 and Fig. 3, necessary to determine the SV and CO is obtained by the visual technique discussed in [5, 7 - 9]. This technique uses the distance measurement method described in items [10 - 14].

In works [15 - 18], it was shown that by stabilizing the level of lighting the markers, it is possible to determine the distance from the markers to the image sensor of the camera and to determine the flaccid membrane surface shape.

II. Motivation

The motivation to undertake works aimed at determining the effect of lighting on the result of the reconstruction of the VAD flaccid membrane surface shape was to obtain, as a result of the 3D reconstruction, the shape of the membrane surface in accordance with the actual state. The consistency with the actual state of the determined shape of the flaccid membrane surface is extremely important due to the

Fig. 1 The first of the pump heart assist models, developed by the authors of this work, based on commercial descriptions of pumps

This work was not supported by any organization

Fig. 2 Currently used proprietary ventricular assist pump model created using 3D printing (2017)

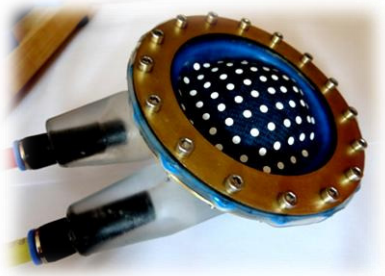


Fig. 3 Currently used proprietary ventricular assist pump model created using a casting technique (2018)

Fig. 4 The extracorporeal pneumatic heart assist pump developed in the framework of the Polish Artificial Heart

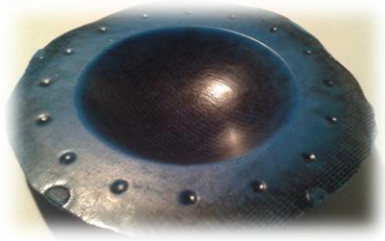


Fig. 5 Proprietary composite flaccid membrane created for pneumatic heart assist pumps shown on Fig. 2 and Fig. 3 (2017)

Fig. 6 Blood chamber presented in Fig. 3 including the red pneumatic chamber (2017)

precision of determining the VAD's instantaneous stroke volume and the cardiac output. Therefore, it was advisable to examine the influence of the lighting parameters on the accuracy of the determined reconstruction of the flaccid membrane surface shape.

III. Problem Definition

The fundamental problem that is being solved in this work is to compensate for the effect of the membrane lighting on the result of 3D reconstruction of the shape of its surface. This impact is the result of constant changes in the position of the surface of the flaccid membrane in relation to the elements that illuminate it.

The magnitude of the influence of lighting on the result of the reconstruction of the flaccid membrane surface shape depends on: a) the shape of the pneumatic chamber; b) the type of light source and its parameters; c) the mode of light distribution in the pneumatic chamber.

Despite many difficulties, this problem was effectively eliminated in the case of the pump shown in Fig. 1. The following was used: a) an illuminator built out of VSMY 1850X0 diodes arranged in a specially selected pattern; b) PCA 9622DRT controller; c) a visible light blocking filter mounted on the camera lens; d) PMMA material for making a pneumatic chamber. This ensured almost uniform lighting in the entire membrane operation range. Unfortunately, this method in the pumps currently used cannot be implemented, due to their construction, and the small distance of the membrane to the walls of the pneumatic chamber, Fig. 7. For this reason, work was undertaken to develop and study the new lighting constructions.

IV. Fitness Function

An important element of the conducted research was the creation of reference shapes of the flaccid membrane surface, consistent with the actual state. The compatibility study was carried out using the membrane surface shapes defined mathematically, Fig. 8a. These include: a) extremely convex – this is the condition of the membrane, in which the membrane is lifted upwards at its maximum; b) extremely concave – this is the state of the membrane in which the downward membrane is at its maximum; c) flat membrane –

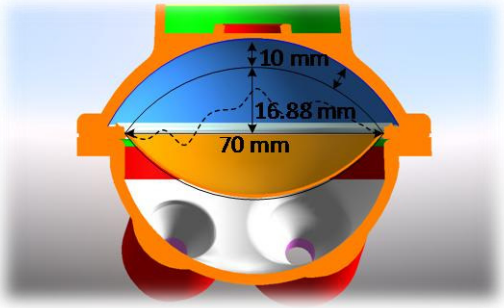


Fig. 7 Cross section of pump model presented in Fig. 6 along with selected dimensions

Fig. 8 Designed (a) and produced (b) model membrane shapes in 3D printing technology

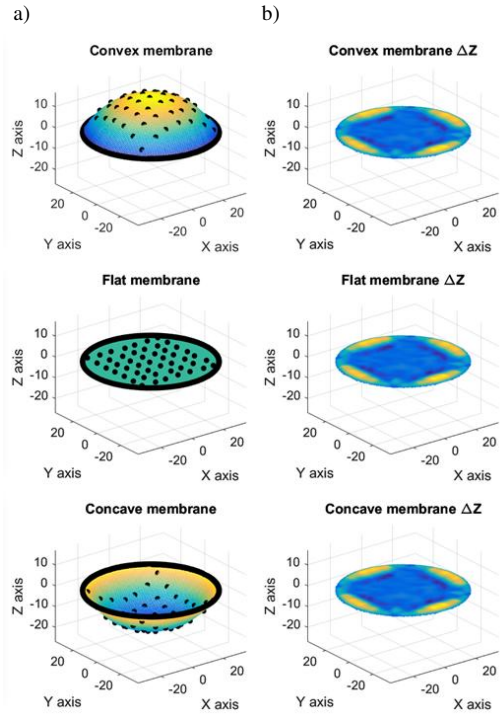


Fig. 9 Model grids determined for membranes from Fig. 8b (a), exemplary charts of membrane reconstruction errors (b)

this is the state in which the entire surface of the membrane is parallel to the plane of the camera image sensor. Due to "additional" material, this condition is not achievable by the membrane shown in Fig. 5. Nevertheless, this membrane is helpful when determining the error of the shape of the membrane surface when it is at the connection level of the blood chamber with the pneumatic chamber.

The reference membranes were modeled in CAD software. As a result, the three-dimensional models shown in Fig. 8a were obtained. These models were then saved in STL format and printed on a 3D printer. The membranes obtained in the manner presented are shown in Fig. 8b. The print was made with a layer height in the Z axis equal to 0.09 mm. The next step was to create reference grids describing the shape of the tested membranes. These grids were determined from known mathematical descriptions. The obtained reference grids are shown in Fig. 9a. Figure 9b shows sample error diagrams of reconstructing the shape of the membrane surface. They were obtained by determining the position difference in the Z axis for all the corresponding points on the reference grid and the measurement grid, representing the currently obtained reconstruction result of the membrane surface shape. The differences determined in this way were used to calculate the mean square error (MSE), which for the best solutions was only 0.0041 mm^2.

V. LIGHT SOURCE DESIGN

The tested illuminator constructions, (A) Fig. 10 and (B), (C) Fig. 11, were made based on the previously tested PCA 9622 DRT system, the application scheme of which is given in Fig. 16 in the paper [20]. 16 PWM channels were used in the illuminators. Each channel controls a LED, which is protected by a resistor limiting the current. The FYLS-0805UWC diode was used in the illuminator (A), working in the visible band at an angle of illumination $\varphi = 130$ deg and a current $I_F = 20$ mA. In systems (B) and (C), infrared diodes were soldered: VSMY 1850X01 ($\varphi = 120$ deg, $I_F = 100$ mA) and VSMG 10850 ($\varphi = 150$ deg, $I_F = 65$ mA). A high intensity LED OF-LED1G4W is used in the fiber optic illuminant (D) shown in Fig. 10, which introduces light into the fiber optic light shining on the side FOSS-3 [21]. The last tested illuminator (E) was made of EL

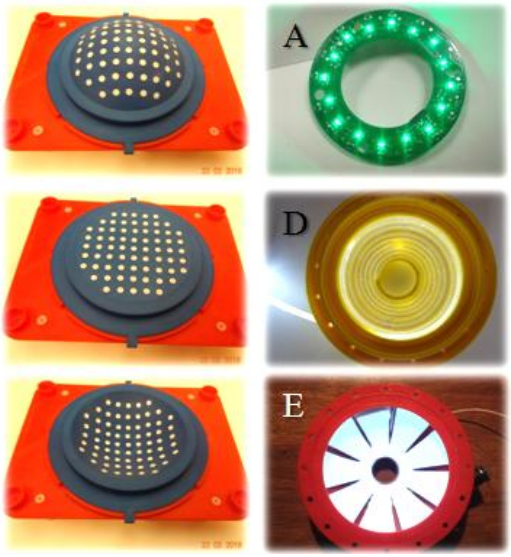


Fig. 10 View of stations and membranes for testing the illuminators (left), view of tested illuminators (right)

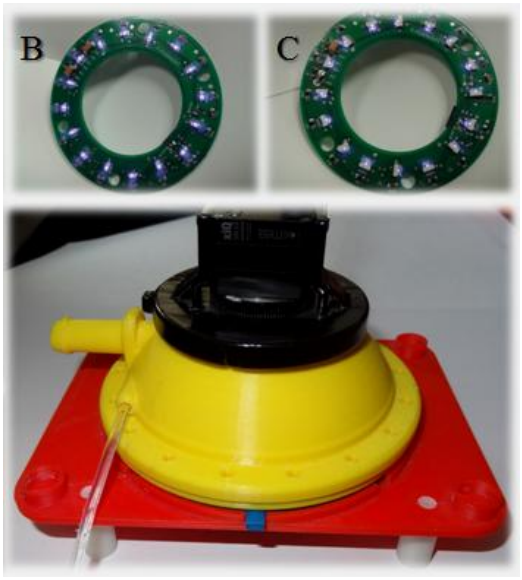


Fig. 11 View of tested illuminators (top), measuring station with a mounted fiber optic illuminator (bottom)

(electroluminescent) film. The EL film was cut into the shape shown in Fig. 10 (E). The film was powered by the converter, which converts 12V DC into AC with an amplitude of 330V and a frequency of 1.7 kHz. A white EL film was used in the study. Application schematics of specialized converters for powering electroluminescent film can be found in [22].

VI. Description of the Experiment

The experiments were carried out on a proprietary measuring station, made using 3D printing technology, Fig. 10 (left). Fig. 11 (bottom) presents the pneumatic chamber during testing with the optical fiber illuminator (D) shown in Fig. 10. The testing of illuminators (A) – (C) shown in Fig. 10 and Fig. 11 started by determining the diode current, which was unchanged for all tested membranes. The experiment involving system (A) gave a negative result – the flat and concave membrane (Fig. 12-2a and Fig. 12-3a) were underexposed. Similar effects were obtained for system (B). Although the flat membrane was properly illuminated, the convex membrane was clearly overexposed and the concave membrane underexposed Fig. 12-1b and Fig. 12-3b. Incorrect results were also obtained for system (C). This system uses infrared LEDs shining to the sides. The applied diodes reduced the spots observed on the surface of the convex membrane, but this adversely affected the illumination of the other two membranes, where most of the markers were below the limit of being distinguishable, Fig. 12-2c and Fig. 12-3c. System (D) turned out to be a promising solution. The fiber optic side lighting illuminated the flat and concave membranes well, Fig. 12-2d and Fig. 12-3d, however, the convex membrane remained underexposed, Fig. 12-1d. The reason for this was seen in the significant (approx. 16 mm) width of the camera lens, and consequently in the lack of the ability

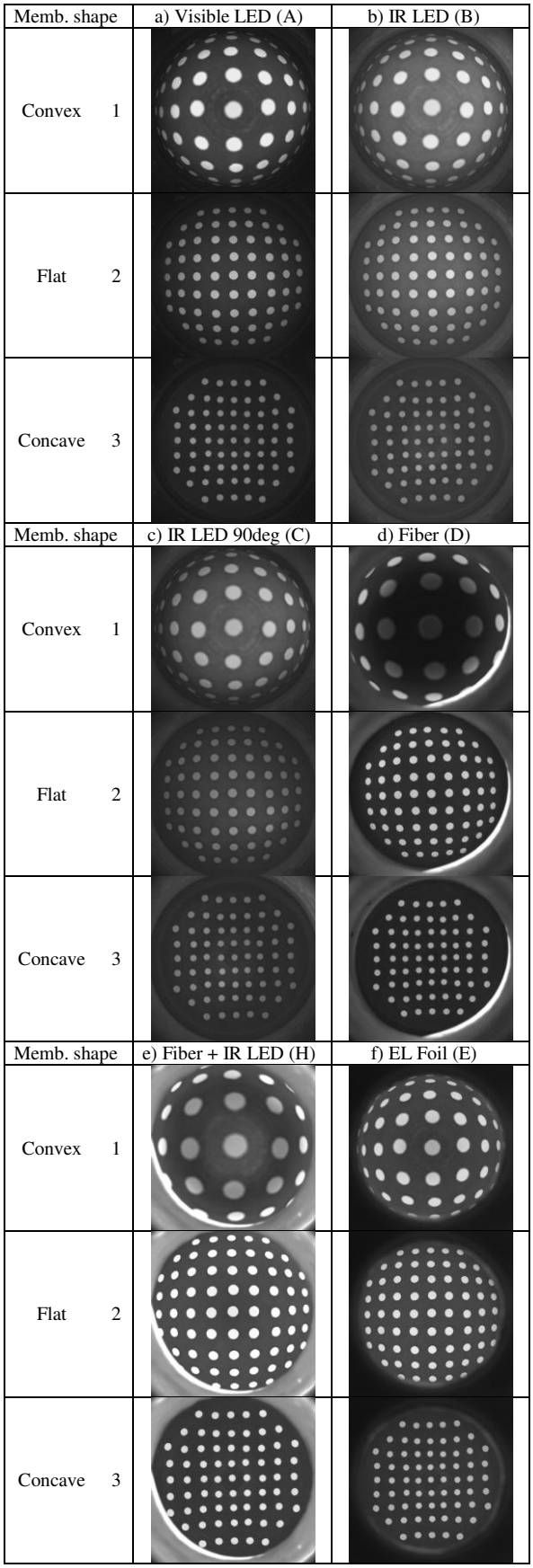


Fig. 12. Views of the studied model membranes, achieved at different lighting settings

to arrange the optical fiber on the entire surface of the dome of the pneumatic chamber. This inconvenience was solved by creating a hybrid (H) illuminator from systems (C) and (D). In the produced system, the optical fiber illuminated the concave membrane and flat membrane, Figs. 12-3e and Fig. 12-2e, and illuminator (C) illuminated the convex membrane, Fig. 12-1e. Moreover, the fiber optic liner arranged on the top of the pneumatic dome dispersed the infrared light, and thus eliminated the spots previously observed on the membrane surface in system (C). Better results were obtained only when testing the illuminator based on electroluminescent film (E). In this case, all the membranes tested were illuminated properly, Fig. 12f.

The evaluation of lighting was done using the technique presented in point IV, and discussed in detail in [5, 19]. Illuminators were examined on a group of three reference membranes shown in Fig. 10 (left). For all tested combinations of illuminator–membrane, measurement marker positions detection was conducted, the results of which are shown in Fig. 13. On this basis, the nodal points of polygons that reflect the surfaces of the tested membranes were determined using the technique presented in [5, 19]. In the following step, by using the two-dimensional data interpolation method, the obtained nets were supplemented with nodes (compacted) to the required measurement resolution. Next, the degree of marker identification, Fig. 13, and the correctness of the shape of the generated grids were analyzed, Fig. 14.

VII. IMAGE PRE & POST PROCESSING

The tests were carried out using a fixed camera equipped with a wide angle lens ($\varphi = 170$ deg) with a fixed focal length ($F = 1.2$). The acquired images were subjected to a distortion compensation procedure developed by the authors [8, 9, 23]. As a result, the membrane views shown in Fig. 12 were obtained. These views show measuring markers with a diameter of 3 mm distributed on the surface of the reference membranes. The observed variations in the size of the spots in the images result from the marker approaching or moving away to/from the plane of the camera image sensor. This effect is further enhanced by the marker zooming out or approaching to/from the focus point. As a result, the view of the marker is blurred. This fact was used in [10] to measure the distance, and in [5] and [23] to determine the shape of the membrane surface in virtual and real space. The condition for determining the shape of the membrane surface is to perform image defuzzification. It was carried out using a thresholding operation that is always carried out with a threshold preset. As a result, a binary image is obtained, ready for labeling. During labeling, (x, y) coordinates of the center of gravity for each spot, Fig. 13, as well as its area are determined. Knowledge of the determined area after the calibration of the measurement system is equivalent with knowing the distance from the marker to the camera image sensor [10]. During further analysis, only nodal points of the grid are processed

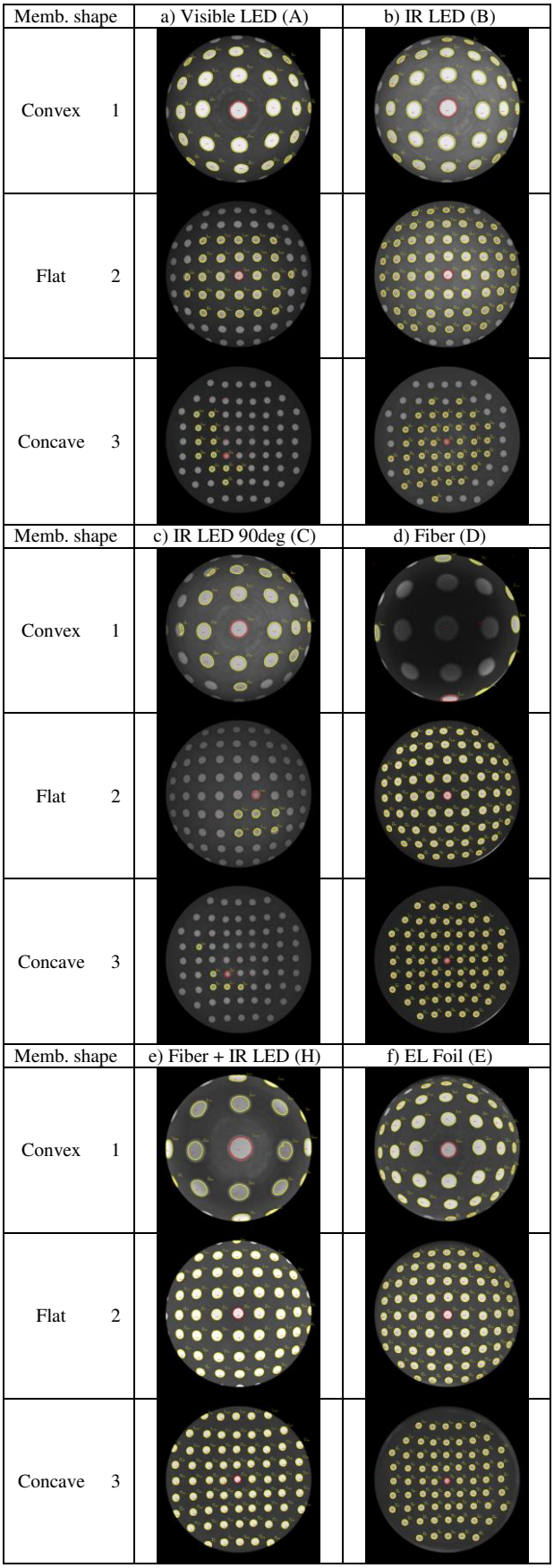


Fig. 13 Results of detecting the position of measurement markers achieved at different lighting settings

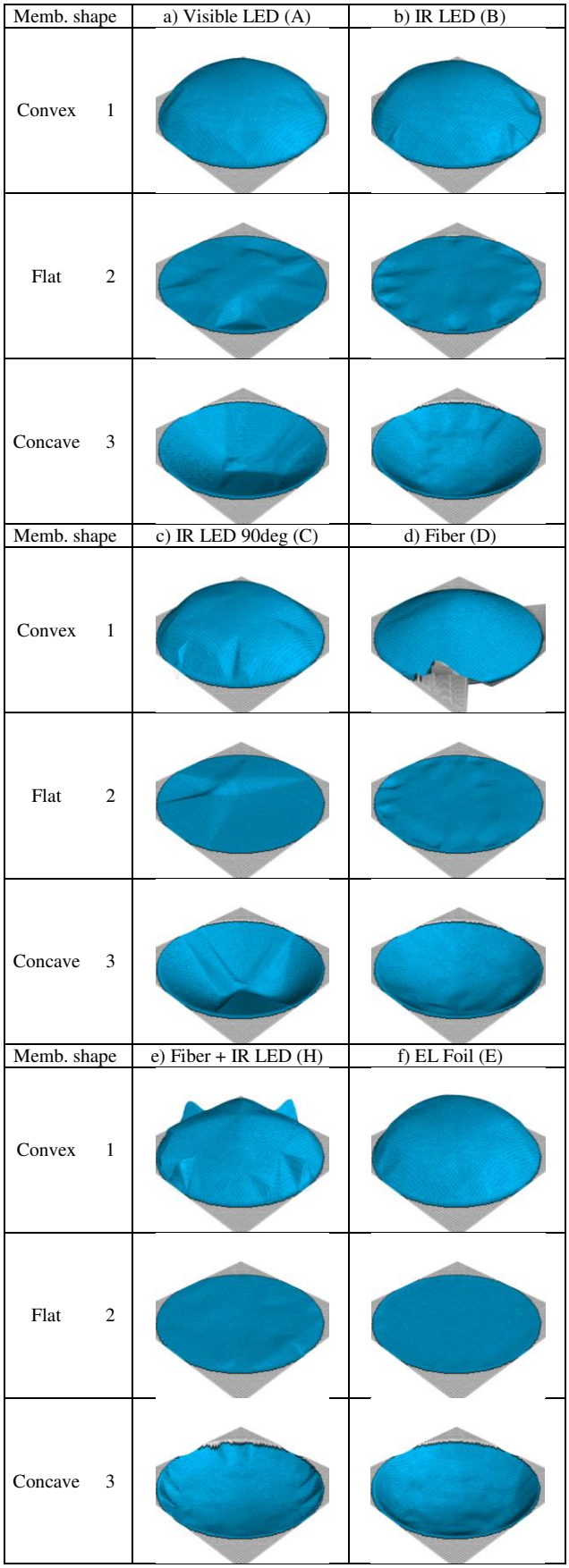


Fig. 14 Results of reconstructing the surface shape of the studied membranes achieved at different light settings

with coordinates (x_i, y_i, z_i) where i = 1,...,69. These points, using the interpolation method, determine a measuring grid of 300 x 300 points.

A graphical representation of the determined grids for all the studied membranes is presented in Fig. 14.

VIII. Results analysis

The analysis of the obtained reconstructions of convex membrane shapes shows that the best shape mapping was obtained for illuminator (E). Next were solutions (A), (B), (C), (H) and (D). In solution (A), some overexposure is visible, which led to local flattening of the determined membrane surface shape. The case of solution (B) is similar with the difference that the entire membrane was illuminated except for the four markers that remained underexposed. In effect the distortion correction function and their values have been "improved", which is visible in the form of characteristic inequalities. When using illuminator (C), it was noticed that the power of light was not fully sufficient. As a result, some of the markers did not work at all, and several markers were underexposed. A result similar to solution (B) is observed for illuminator (H). However, in this case the markers distant from the center of the membrane were overexposed and the markers located closer to the center were illuminated, but their brightness was noticeably lower. In addition, four markers under the given lighting conditions had too small a surface and were "improved" by the correction function. The worst illuminator in the study of the convex membranes was illuminator (D). Analyzing the picture shown in Fig. 13-1d it is easy to see that this state results from the underexposure of the middle portion of the membrane. No markers have been identified in this area. The obtained result is therefore the result of interpolation made on the basis of knowing the values of the characteristics determined for markers located only on the edge of the membrane.

The flat membrane, just like the convex membrane, was best visualized when illuminated with illuminator (E). A very good result was also obtained for illuminators (H), (D) and (B). When using the illuminators (H) and (D), all the markers on the membrane were working, Fig. 13-2e and Fig. 13-2d. In solution (B) only seven markers less worked, Fig. 13-2b. In this group, the smallest distortion of the shape of the membrane was observed for illuminator (H). Next came illuminator (D), and finally (B). When illuminating the membrane with system (D) the membrane was gently raised in the center. The use of the construction (B) allowed illuminating the central part of the membrane evenly, but the markers at the edge of the membrane were less illuminated. This is visible in the form of fluctuation observed near the edge of the membrane, Fig. 14-2b. The remaining solutions (A) and (C) did not work at all. When illuminator (A) was used only 22 markers placed in the center of the membrane were correctly marked. Along with them were still six markers operating, but they were not well-lit and, as a result, they were the cause of errors. In the case of illuminator (C),

only five markers worked correctly, which could not lead to a correct membrane shape reconstruction.

The last test was to determine the shape of the concave membrane. The best results were obtained for illuminators (E), (D), (H) and (B). Illuminators (A) and (C) due to bad lighting of the membrane in its lower position, did not allow obtaining correct reconstruction results.

According to the authors, the shape of the concave membrane was best determined using illuminator (E). Then the sides of the membrane are reproduced correctly, and only minor remarks can be made to the shape of the central part of the membrane. In the second position, it was decided to place illuminator (D), which properly illuminated the measurement markers, and yet in the central part of the membrane reconstruction errors were noticeable. An equally good result was obtained for illuminator (H). Just like in the previous case all markers on the obtained membrane shape "worked". As a result, the shape of the central part of the membrane is reproduced correctly (according to the authors even better than for illuminator (E)). However, disturbances were observed on the sides of the membrane, which forced illuminator (H) to be classified in third position. The last discussed illuminator is construction (B). This system achieved a surprisingly good reconstruction result, taking into account that only 41 out of 69 markers worked properly. Their lack, however, explains the reconstruction errors observed on the side surfaces of the concave membrane.

IX. Conclusions

The work discusses new illuminator constructions made for a vision sensor to measure the instantaneous stroke volume and the cardiac output of the heart pump. The experiments were carried out on a group of three reference membranes with a known mathematical description. Five variants of illuminators were subject to examination. Three tested constructions worked in the visible spectrum. The other two were implemented in the near-infrared band.

The experiments carried out have shown that it is possible to construct an illuminator that will effectively eliminate the problem of uneven surface illumination of a flaccid membrane. Among the illuminators, which design is partly based on LED diodes, the highest potential was demonstrated by the hybrid illuminator (H). It combines the advantages of a fiber optic illuminator operating in the visible band (D) and illuminator (C), made of IR LEDs illuminating the sides. Although this variant requires two supply circuits, it allowed obtaining the correct shape of a flat and a concave membrane. Initially, the incorrectly determined shape of the surface of the convex membrane was finally corrected, increasing the control current of VSMG 10850 diodes. This operation was carried out programmatically by changing the setting of the PWM circuits.

The best of the examined systems was an illuminator built with the use of an electroluminescent film. The membrane surface shapes obtained were the closest to model shapes. This solution, however, due to the significant voltage supply in a commercially produced a heart-assisting pump, requires a special approach for the production of a pneumatic chamber.

The final effect of reconstructing the shape of the membrane surface also depends on the number and position of the markers. The tests were carried out taking 49 markers, Figs. 8 and 69 markers, Fig. 10 - 13. A larger number of markers, when properly illuminated and unfolded, more precisely determines the shape of the membrane surface sought. It is worth noting that, in principle, the material from which the diaphragm is made may be any. The test membrane was made of a graphite-silicone composite. It was used in the VAD, for which the SV was determined. The obtained values of the volume were affected by an error smaller than 4.8 ml, which constituted about 4% of the volume of the blood chamber VAD [23].

X. Author's statement

The work presented in this publication was financed only by the authors' own funds.

References

[1] Grad, L., Murawski, K., Pustelny, T., “Measuring the Stroke Volume of the Pneumatic Heart Prosthesis using an Artificial Neural Network”, *Proc. SPIE* 10034, (2016), doi: 10.1117/12.2243952.

[2] Grad, L., Murawski, K., Sulej, W., “Research to Improve the Accuracy of Determining the Stroke Volume of an Artificial Ventricle Using the Wavelet Transform”, *Proc. SPIE* 10455, (2017), doi: 10.1117/12.2280804.

[3] Murawski, K., Pustelny, T., Grad, L., Murawska, M., “Estimation of the Blood Volume in Pneumatically Controlled Ventricular Assist Device by Vision Sensor and Image Processing Technique”, *Proc. MMAR*, 100 – 106, (2016), doi: 10.1109/MMAR.2016.7575115.

[4] Sulej, W., Murawski, K., “Determining the Stroke Volume of the Artificial Ventricle Using the Numerical Integration Method”, *Proc. SPA*, 207 – 2012, (2017), doi: 10.23919/SPA.2017.8166865.

[5] Murawski, K. Murawska, M., Pustelny, T., “The System and Method of Determining the Shape of the Membrane Pneumatic Pump of Extracorporeal Heart Assist Device”, *Patent Application*: nr P.414104, 2015. (in Polish).

[6] Kustosz, R., Jarosz, A., Gawlikowski, M., Kapis, A., Gonsior, M., “The Role and Perspectives of Development of the Polish Air Pump Heart Assist on the Market of Heart Prosthetic”, Polish artificial heart, the development of design, qualification tests, preclinical and clinical, ISBN 978-83-63310-12-7, 2013, (in Polish).

[7] Sulej, W., Murawski, K., “The Membrane Shape Mapping of the Artificial Ventricle in the Actual Dimensions”, *Proc. FEDCSIS*, 675 – 680, (2017), doi: 10.15439/2017F269.

[8] Sulej, W., Murawski, K., Pustelny, T., “Improvement of Accuracy of the Membrane Shape Mapping of the Artificial Ventricle by Eliminating Optical Distortion”, *Proc. MMAR*, 799 – 804, (2017), doi: 10.1109/MMAR.2017.8046934.

[9] Sulej, W.; Murawski, K.; Pustelny, T. “Optical Distortion Compensation in Visual Measurement with a New Depth From Defocus Method”, *Photonics Letters of Poland*, 9, 4, 122 – 124, (2017), doi: 10.4302/plpv9i4.775.

[10] Murawski, K., “Method of Measuring the Distance using One Camera”, *Patent Application*: P.408076, 2014. (in Polish).

[11] Murawski, K., “Method of Measuring the Distance to an Object Based on One Shot Obtained from a Motionless Camera with a Fixed-Focus

Lens", Acta Physica Polonica A 127 (6), 1591-1595 (2015). DOI: 10.12693/APhysPolA.127.1591
[12] Murawski, K., Arciuch, A., Pustelny, T., "A New Innovative Depth From Defocus Method - Identification of the Impact of the Marker Size on the Distance Measurement Result", Proc. SPIE 10034, (2016), doi: 10.1117/12.2244130.
[13] Murawski, K., Arciuch, A., Pustelny, T., "Studying the Influence of Object Size on the Range of Distance Measurement in the new Depth From Defocus Method", Proc. FEDCSIS, 817 – 822, (2016), doi:10.15439/2016F136.
[14] Murawski, K.; Arciuch, A.; Pustelny, T. "Study of Distance Range Visual Measurement in the new Depth From Defocus Method",Photonics Letters of Poland, 8, 2, 48 – 50, (2016), doi:10.4302/plp.2016.2.07.
[15] Murawski, K., "Measurement of Membrane Displacement with a Motionless Camera Equipped with a Fixed Focus Lens", Metrology And Measurement Systems, 22, 1, 69 – 78, (2015), doi: 10.1515/mms-2015-0011.
[16] Murawski, K., "Measurement of Membrane Displacement Using a Motionless Camera", Acta Physica Polonica A 128 (1), 10–14, (2015). doi: 10.12693/APhysPolA.128.10.
[17] Murawski, K., "New Vision Sensor to Measure and Monitor Gas Pressure", Acta Physica Polonica A 128 (1), 6 – 9, (2015). doi: 10.12693/APhysPolA.128.6.
[18] Murawski, K., "New Vision Sensor to Measure Gas Pressure", Measurement Science Review, 15, 3, 132 – 138, (2015), doi: 10.1515/msr-2015-0020.
[19] Sulej, W., Grad, L., Murawski, K., "The Technique of Accuracy Measurement of Membrane Shape Mapping of an Artificial Ventricle", Proc. SPIE 10455, (2017), doi: 10.1117/12.2280806.
[20] https://www.nxp.com/docs/en/data-sheet/PCA9622.pdf.
[21] http://www.soled.pl/oswietlenie-swiatlowodowe/swiatlowody-swiecace-bokiem.html
[22] http://www.microchip.com
[23] Sulej, W., "Measurement of the stroke volume of the extracorporeal pneumatic heart assist pump using image processing and analysis techniques", PhD. Dissertation, Military University of Technology, Warsaw, (2018) – in Polish.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 569–576
DOI: 10.15439/2018F66
ISSN 2300-5963 ACSIS, Vol. 15

Secret Key Sharing Protocol between Units Connected by Wireless MIMO Fading Channels

Valery Korzhik, Aleksandr Gerasimovich,
Cuong Nguyen, Vladimir Starostin,
Victor Yakovlev, Muaed Kabardov

The Bonch-Bruevich Saint-Petersburg
State University of Telecommunication
Saint-Petersburg, Russia
Email: val-korzhik@yandex.ru, star_vs47@mail.ru

Guillermo Morales-Luna
Computer Science
CINVESTAV-IPN
Mexico City, Mexico
Email: gmorales@cs.cinvestav.mx

***Abstract*—The method of secret key sharing between units that did not possess any secret keys in advance is considered. It is assumed that between these units there are duplex wireless MIMO fading channels. In a recent paper published by D. Qin and Z. Dingh a new key sharing protocol has been proposed between legitimate users based on eigenvalues which are invariant under permutation of two matrices in their product. We extend this statement to a characteristic polynomial and by the way to matrix trace. Methods of key bits extraction are optimized both theoretically and experimentally. On the contrary to a statement of D. Qin and Z. Ding we prove that their key sharing protocol occurs insecure if eavesdroppers have the same channels as legitimate users. In order to provide reliability and security of the shared keys both error correction codes and privacy amplification methods can be used.**

***Index Terms*—Physical layer security, key sharing protocol, MIMO transmission system, characteristic polynomial, privacy amplification, error correction codes**

I. Introduction

THE pioneered paper devoted to key sharing protocol for users that did not have any secret keys in advance belongs to Diffie and Hellman [1]. It is well known that security of this protocol rests on the intractability of the *Diffie-Hellman Problem* or simply the related *discrete logarithm computing problem* [2].

There is also a class of keyless cryptography (KC), where encryption of messages can be provided secure even without any prior secret key sharing. One of such KC can be implemented by some protocol if we have encryption algorithm satisfied to the following relation for any different keys K_A, K_B and any plaintext M:

$$f_{K_A}(f_{K_B}(M)) = f_{K_B}(f_{K_A}(M)) \quad (1)$$

where f_K is the encryption algorithm for plaintexts given a key K. Then the encryption/decryption protocol between users A and B can be performed as shown in Table I. But unfortunately the condition (1) is not valid for strong symmetric block ciphers.

Alpern and Schneier [3] proposed a cryptographic technique in which the security lies in hiding the identify of the message ordinator.

TABLE I
Encryption/decryption protocol.

M: $A(K_A)$		$B(K_B)$
1.	$C_A = f_{K_A}(M)$	
2.	$\rightarrow C_B = f_{K_B}(C_A)$	
3.	$\leftarrow C'_A = f_{K_A}^{-1}(C_B); f_{K_B}^{-1}(C'_A) = M$	

In [4] some extensions to the previous scheme was suggested that was called as *semi-anonymous channel*. Although the last scheme seems to be more realistic than the previous one but both scenarios require serious restrictions regarding communication network between users that want to share secret keys.

On the other hand it was developed in recent years a new domain known as *physical layer security* (PHY) *in multiuser wireless networks*. In this setting it is assumed that users are connected by some communication (mostly continuous) channels and the properties of these channels allow either implement directly secure information transmission between users or to share secret keys for their further usage with conventional encryption/decryption. It is worth to note that such keyless cryptosystem was based firstly on Wyner's *wire-tap channel concept* proposed in 1975 [5]. This approach has been developed later in fundamental papers [6]–[8].

But we should emphasize that in order to provide information theoretic security in wireless networks it is necessary to have in any case some advantages in legitimate communication channels against eavesdropper's channels. Such advantages are presented in Table II jointly with list of references where they were used in order to provide information security of messages or key string sharing in frames of given conditions.

Summarizing the content of Table II, we can conclude that no one of the keyless cryptosystems satisfy the natural requirements: to be secure independently on eavesdropper channel or equipment states. In fact, legal user cannot provide that SNR in eavesdropper channel is not larger than some

TABLE II
POSSIBLE ADVANTAGES OF THE LEGITIMATE CHANNELS AGAINST EAVESDROPPER CHANNELS.

Nr.	Advantages of the legitimate channels	Defect of such setting	References
1.	SNR in legitimate channels is superior to SNR in eavesdropper channel	SNR as a rule is unknown in eavesdropper channel	[5], [6], [8], [9]
2.	Not all symbols of legally transmitted blocks can be intercepted by eavesdropper	It is very specific and rare case	[10], [11]
3.	Legal users have authenticated channel for public discussion	Even so authenticated channel is provided by additional measures it is unknown SNR in the eavesdropper channel in order to optimize parameters of legal transmission	[7], [12], [13]
4.	Legal channels are sensitive to any adversary intervention. (Quantum cryptography)	Special legal channels and devices are required	[14], [15]
5.	Legal users are mobile and communication channels have multipath wave propagation. (MIMO technology can be used also for security enhancing)	Mobile units can stop sometimes. Eavesdropping is still possible on very short distance from legitimate units. Reciprocity theorem of radio wave propagation can be invalid in some cases.	[16], [17], [18]
6.	Smart antennais excited randomly by electronic means and a presence of multipath communication channels is requested. (It is not required that units can be nonstop; and eavesdropper channel can be even noiseless)	Eavesdropping is possible on very short distance from legitimate units. Reciprocity theorem of radio wave propagation can be invalid in some cases.	[19], [20]
7.	The number of antennas in legitimate MIMO system is not less than the number of eavesdropper antennas	Cryptosystem can be broken if the number of eavesdropper antennas is larger than the number of legitimate antennas	[21], [22], [23]

given value, that the number of antennas in eavesdropper MIMO system is not larger than the number of legitimate antennas and finally that reciprocity of channels is always valid.

But fortunately, it has been published recently the paper [24] in that some of mentioned above problems can be removed.

In Section II we describe one of key sharing schemes presented in [24] that is on our opinion very interesting from a practical point of view. Later we extend the protocol in [24] and examine theoretically how to optimize its parameters. In Section III we present experimental results obtained by simulation. Section IV devoted to error correction and privacy amplification of key string shared by legitimate units after performing of protocol. Section V concludes the paper and proposes some open problems for further investigations.

II. EXTENSION OF EVSKEY SCHEME

Let us remind the key sharing protocol proposed in [24] and called there *EVSkey scheme*. The scenario corresponding to this scheme is presented in Figure 1.

For simplicity reasons we restricted our consideration by the condition of equality for the numbers of antennas of the legitimate users Alice (A) and Bob (B), both at the transmitter and at the receiver are n.

Before transmission, Alice and Bob generate their own reference matrices $X_A, X_B \in \mathbb{C}^{n\times n}$, as well as randomly generated unitary matrices $G_A, G_B \in \mathbb{C}^{n\times n}$. In line with our previous assumption all matrices are square of order $n \times n$.

Let the noise matrices $N_{B1}, N_{A1}N_{B2}, N_{A2}$ have *additive white Gaussian numbers* (AWGN) as random values. After the postmultiplication of the channel matrices H_{AB} and H_{BA} by G_B and G_A, respectively and sending the resulting matrices back, users Alice and Bob get the following matrices:

$$\text{Alice:} \quad Y_A = PQX_A + PN_{B1} + N_{A2} \quad (2)$$

$$\text{Bob:} \quad Y_B = QPX_B + QN_{A1} + N_{B2} \quad (3)$$

$$\text{with } P = H_{BA}G_B \quad , \quad Q = H_{AB}G_A \quad (4)$$

For small enough noises $N_{B1}, N_{A2}, N_{A1}, N_{B2}$ we get a good estimation for the matrices PQ and QP respectively as

$$PQ \approx Y_A X_A^{-1} \quad , \quad QP \approx Y_B X_B^{-1}.$$

Since Alice knows Y_A, X_A and Bob knows Y_B, X_B, they are able to compute the matrices PQ and QP although with some errors due to the presence of noises.

In [24] it is suggested to extract a common key as the quantized complex eigenvalues of matrices PQ and QP since those eigenvalues coincide one to another although these matrices may be completely different. We extend their statement and prove the following:

Lemma 1: Given two non-singular complex matrices $P, Q \in \mathbb{C}^{n\times n}$, the matrices PQ and QP have the same *characteristic polynomials*.

Proof. By definition, the characteristic polynomial of PQ is $\pi(\lambda) = \det(PQ - \lambda I)$, where I is the identity matrix.

Then the roots λ of the characteristic polynomials satisfy the equation

$$\det(PQ - \lambda I) = 0. \quad (5)$$

It follows from (5), being Q a unitary matrix,

$$\begin{aligned} 0 &= \det(PQ - \lambda I) \\ &= \det Q \det(PQ - \lambda I) \\ &= \det(QPQ - \lambda Q) \\ &= \det(QPQ - \lambda Q) \det Q^{-1} \\ &= \det(QP - \lambda I) \end{aligned}$$

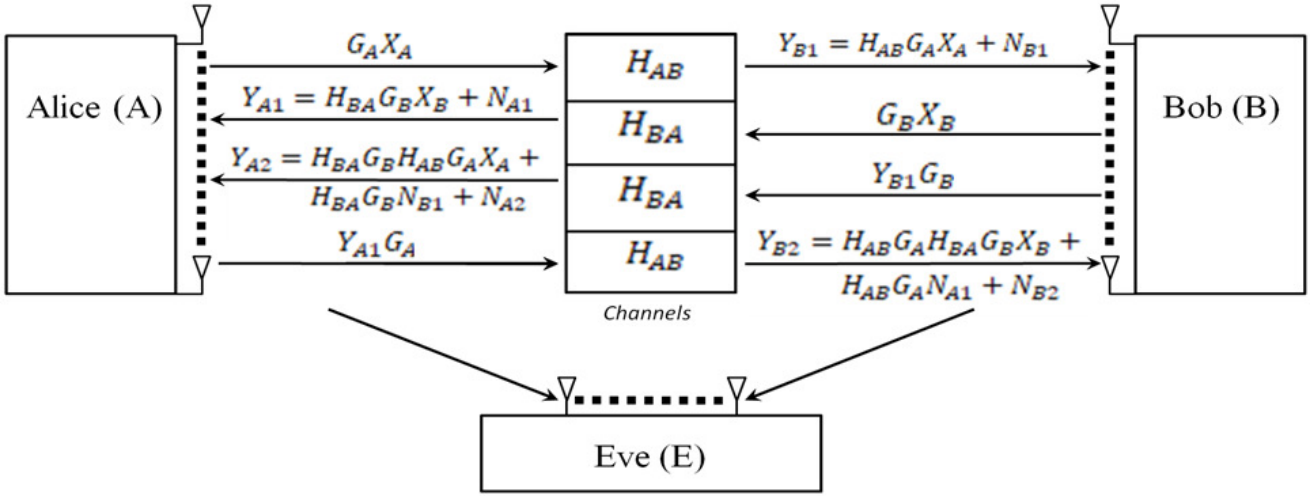


Fig. 1. The scenario corresponding to EVSKey scheme.

Then the roots of the characteristic polynomials of matrices PQ and QP coincide one to another and hence these matrices have the same characteristic polynomials. □

Thus we can calculate for the key bit generation not only the eigenvalues but also all coefficients of the characteristic polynomial and in particular case the traces of matrices PQ and QP or their determinants. Let us investigate, at first theoretically, which of the main invariants-eigenvalues or traces are less sensitive to channel noises, more closer to uniform distribution and give the most number of reliable key bits for legitimate users.

A. Using quantized matrix traces as shared key bits

Since the traces of matrices are complex values they can be quantized both on amplitude and on phase. It is proved in the Appendix that the quantization intervals on amplitude of the traces in order to provide equal probabilities of their occurrence should be chosen as follows:

$$r_{k-1} \le |Z| \le r_k \quad , \quad k = 1, 2, \ldots, N \tag{6}$$

where Z is the trace of the matrices, $r_k = \sigma\sqrt{-\ln\left(1 - \frac{k}{N}\right)}$ and N is the number of intervals.

Then the probabilities that quantized trace amplitudes coincide for users Alice and Bob will be determined by

$$p' = \sum_{k=1}^{N} \left((1 - (k-1)p)^{\frac{1}{\gamma^2}} - (1 - kp) \right) \tag{7}$$

where $\gamma = \frac{1}{1+\alpha}$, $\alpha = \sigma^2\left(1 + \frac{1}{N}\right)$, $p = \frac{1}{N}$.

In Table III there are presented the results of calculations by (7) for some parameters. We see from this table that the probability of errors are still acceptable for $N = 16$ if $\sigma^2 \le 0.001$ and for $N = 32$ if $\sigma^2 \le 0.0001$.

TABLE III
THE PROBABILITIES OF KEY COINCIDING BY (7) AFTER A PERFORMANCE OF KEY SHARING PROTOCOL BASED ON QUANTIZATION BY (6) THE MATRIX TRACES ON AMPLITUDE.

$N \setminus \sigma^2$	0.01	0.001	0.0001
4	0.98	0.998	0.9998
8	0.96	0.996	0.9996
16	0.92	0.992	0.9992
32	0.84	0.984	0.998
64	0.68	0.968	0.9968

σ^2: SNR N: Number of quantization intervals

B. Using quantized matrix eigenvalues as the shared key bits

Then every eigenvalue can be quantized on phase and amplitude intervals. Unfortunately there appears one problem: how to compare the numbering of eigenvalues adopted by the users?

Let us denote by N_P, N_A the numbers of quantization intervals on phase and amplitude, respectively. Then total number of quantization intervals is $N = N_A \cdot N_P$. We will fix the number of eigenvalues that hit in each of the N intervals (cells). After a completion of eigenvalues extraction, we get a string of integers $g_1, g_2, \ldots, g_i$, where g_i is the number of the i-th cell containing at least one eigenvalue. If several eigenvalues occur in the same cell, then the cell number is repeated as $g_i, \ldots g_i$. Next each number g_i is presented as a string of bits and such strings are connected in a consecutive binary manner. The final binary string forms a part of the shared key. It is easy to see that the total number of bits for each session of protocol can be, if $N \gg n$, approximately computed [25] as:

$$\log_2 \binom{N+n-1}{n} = \log_2 \left[\frac{1}{n!} \prod_{i=N}^{N+n-1} i \right] \tag{8}$$

C. Security of the proposed key sharing protocol

As it is shown in Figure 1, the eavesdropper Eve is able to receive only the matrices $G_A X_A, G_B X_B, Y_{A1}, Y_{B1}, Y_{A2}, Y_{B2}$ even for the ideal case when eavesdropping channels are noiseless. It is claimed in [24] that even in the very unrealistic case when Eve's receivers are located very close to the locations of Alice and Bob, and hence she is able to estimate correctly the channel matrices H_{AB}, H_{BA} of legitimate users, she is unable to compute the matrices P and Q (see eq (4)) because they are "randomized" by the unitary matrices G_B and G_A. The last matrices cannot in turn be estimated by Eve because they are "randomized" by the reference matrices X_A and X_B.

In [24] it is concluded that such key sharing system is *ideal secure* and its security is regardless of the state of the channels and the SNR in the eavesdropper channel, in contrast to all key distribution protocols described actually in Table II. *Unfortunately this statement is wrong.* In fact, following the steps below, Eve for sure is able to receive the key shared by the legitimate users:

1. $H_{BA}G_B = H_{BA}G_B H_{AB}G_A X_A \left(H_{AB}G_A X_A\right)^{-1}$
2. $X_B = \left(H_{BA}G_B\right)^{-1} H_{BA}G_B X_B$
3. $QP = Y_B X_B^{-1}$
4. $QP \rightarrow$ characteristic polynomial (equivalent to the shared key)

The key bit string should have good statistical properties as it is common for all secret cryptographic keys. (Such property is verified in the next Section using the NIST tests on pseudorandomness.)

On the other hand in order to provide a good key bit agreement between legitimate users it is very important a strong correlation between channel matrices in the first and in the second steps of the key sharing protocol.

In fact, if they would be different, say H_{AB}, H_{BA} at the first step and H'_{AB}, H'_{BA} at the second step, we would get (even in noiseless channels) instead of relations (2-4) the following ones:

$$\begin{array}{rcccl} Y'_A & = & Y'_{A2} & = & H'_{BA}G_B H_{AB}G_A X_A \\ Y'_B & = & Y'_{B2} & = & H'_{AB}G_A H_{BA}G_B X_B \end{array} \quad (9)$$

From the second equation in (9), there is no a matrix permutaition of the first one and hence the matrices Y'_A and Y'_B have not necessarily equal characteristic polynomials.

In order to provide a strong correlation between channel matrices in the first and in the second steps of the key sharing protocol (channel coherence property – in other words) it is necessary to agree physical channel properties with the rate of communication.

Typical data rates for Wi-Fi network or celular communication (LTE, 56) lies in a range of several hundreds ms. Coherence time for channels used in mobile unit communication is in range (1-10 ms) [26] and then during coherence time a number between 103 and 106 of bits can be transmitted which is sufficient to provide practical coincidence of Y_A, Y_B with matrices Y'_A, Y'_B.

TABLE IV
SIMULATION RESULTS OF THE BIT ERROR PROBABILITIES (IN PERCENT) FOR EXTRACTION THEM FROM EIGENVALUES. BOTH NUMBERS OF PHASE QUANTIZATION INTERVALS AND AMPLITUDE ONE ARE 8.

SNR $\frac{1}{\alpha}$ (dB) $\setminus$ n	4	8	16
20	21.6	22	24
30	7.7	10	12
40	2.7	3.5	4
Number of extracted bits	19	33	52

n is the number of antennas

TABLE V
LIST OF NIST TESTS ON PSEUDO RANDOMNESS.

Nr.	Title of test
1	The frequency test
2	Frequency test within a block
3	The runs test
4	Tests for the longest-run-of-ones in a block
5	The binary matrix rank test
6	The discrete Fourier transform (spectral) test
7	The non-overlapping template matching test
8	The overlapping template matching test
9	Maurer's "Universal Statistical" test
10	The linear complexity test
11	The serial test
12	The approximate entropy test
13	The cumulative sums (cusums) test
14	The random excursion test
15	The random excursions variant test

Unfortunately the considered system (as well as all PHY-based systems) is vulnerable against active adversary. It is a scenario where an adversary, say Mallet, is presented by Alice or Bob as legitimate users and performs with any of them the above mentioned protocol. It is obvious that then he is able to share reliable key after completing the protocol. Such problem has to be solved by some additional activity of legitimate users, in order to reject falsely shared key before its implementation for encryption of secure messages [27].

III. SIMULATION RESULTS FOR THE PROPOSED KEY SHARING PROTOCOL

In order to verify our theoretical discussion it was undertaken a simulation of the EVSkey protocol. The results of simulation for extraction of key bits from matrix eigenvalues are presented in Table IV, where is presented also the number of key bits for different number of antennas n calculated by (8).

We see from Table IV that the acceptable SNR is at least 30 dB even for the case when we mean to use later error correcting codes (see Section IV). As far as the lengths of share key string they are too small for implementation even for block ciphers like 3DES or AES. Thus one can be recommended to repeat key sharing session several times. (Such approach is also presented in Section IV.)

The generated key bits were investigated by NIST tests on pseudo randomness [28]. The list of NIST tests is presented in Table V, while the results of testing on pseudo randomness in Table VI with their numbering taken from Table V.

TABLE VI
RESULTS OF THE NIST-BASED TESTING FOR THE KEY BITS SEQUENCE EXTRACTED FROM THE MATRIX EIGENVALUES UNDER THE CONDITION SNR = 30 DB AND ALSO AFTER A SHIFTING AND SUMMATION PROCEDURE.

("1" – means that test is passed, "0" – that test is not passed).

Test number	Original one	After shift and addition mod 2
1	1	1
2	1	1
3	1	1
4	0	1
5	1	1
6	0	0
7	1	1
8	1	1
9	1	1
10	1	1
11	0	0
12	0	0
13	1	1
14	0	0
15	0	0

TABLE VII
SIMULATION RESULTS FOR PROBABILITY OF KEY (TRACE) COINCIDING AFTER A PERFORMANCE OF KEY SHARING PROTOCOL BASED ON QUANTIZATION BY (6) THE MATRIX TRACES ON AMPLITUDE (16 ANTENNAS).

The number of rings	Number of key bits	σ^2	P_{tr}
4	2	0.01	0.88
		0.001	0.90
		0.0001	0.98
8	3	0.01	0.82
		0.001	0.94
		0.0001	0.99
16	4	0.01	0.74
		0.001	0.90
		0.0001	0.98
32	5	0.01	0.68
		0.001	0.83
		0.0001	0.97
64	6	0.01	0.67
		0.001	0.78
		0.0001	0.92

In the same Table VI there are presented also the results of NIST-based testing after a shifting right on the 20 bits and addition $\mathrm{mod}\ 2$ with the original sequence.

We see that after the transformation procedure the key sequence occurs slightly better. The results of simulation for extraction of key bits from matrix traces are presented in Tables VII, VIII. Comparing the results in Table IV and Tables VII, VIII we see that extraction of the key bits from the matrix eigenvalues results in larger errors than for the trace-based extraction but the number of extracted bits is significantly less for the case of extraction from the traces than for the extraction from eigenvalues.

The key bits extracted from traces were investigated by the NIST tests given in Table V. The results of testing are shown in Table IX jointly with "shift and addition" transformation. We can see from this Table that now an additional transform is not

TABLE VIII
SIMULATION RESULTS OF THE BIT ERROR PROBABILITIES P' (IN PERCENTS) FOR EXTRACTION THEM FROM MATRIX TRACES WITH DIFFERENT SIZES OF QUANTIZATION LEVELS AND ANTENNA NUMBERS.

The number of antennas	The number of sectors	The number of rings	Number of key bits	σ^2	P'
4	8	8	6	0.01	14.7
				0.001	4.7
				0.0001	2.1
	16	4	6	0.01	14
				0.001	4
				0.0001	1.5
	32	4	7	0.01	21
				0.001	11
				0.0001	3
	16	8	7	0.01	18
				0.001	7
				0.0001	2
	8	16	7	0.01	19
				0.001	10
				0.0001	2
	32	8	8	0.01	19
				0.001	10
				0.0001	2
8	8	8	6	0.01	14.3
				0.001	4.4
				0.0001	1.1
16	8	8	6	0.01	12.3
				0.001	6.7
				0.0001	0.7

TABLE IX
RESULTS OF NIST-BASED TESTING FOR THE KEY BITS EXTRACTED FROM MATRIX TRACES UNDER CONDITION OF SNR = 30 DBAND ALSO AFTER A SHIFTING AND SUMMATION PROCEDURE.

Test number	Original one	After shift and addition mod 2
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11	1	1
12	1	1
13	1	1
14	0	0
15	0	0

necessary. This means that this case is superior to extraction from eigenvalues with point of key statistic view.

By comparing the results of Table III and Table VII we conclude that the quantization procedure based on (6) is acceptable. This is valid also for the case of 4 and 8 antennas.

IV. ERROR CORRECTION AND PRIVACY AMPLIFICATION

We assume that the length of the shared key should be at least 256 bits, taken into account for example that the length of key string for AES is 128 bits. This means that in order to provide the requested key length it is necessary to arrange

several sessions of key sharing protocol. Moreover in order to provide a good statistic of shared key bits it is necessary that states of channel matrices between sessions should be statistically independent. In order to short the number of such sessions the method of key bit extraction from matrix eigenvalues occurs preferential because it allows to extract more bits than matrix trace extraction during a single session (see Table IV and Tables VII, VIII). But anyway the values of bit error probabilities are too much for a good key agreement between legitimate users. This fact requires to correct errors by sending over public noiseless channel the check symbols of some good error correction code. But on the other hand a sending of check symbols over public (open) channel results in a leaking to Eve some information about key string. In order to guarantee that such leakage is limited by some value of Shannon information it is necessary to use so called *privacy amplification*. It can be provided by hashing of raw key string to more shorter final key string. It has been proved in [29] the enhanced *privacy amplification theorem.* This theorem says that using special two-stage hashing procedure the eavesdropper's expected Shannon information I_o about the final key shared by legitimate parties, satisfies the inequality:

$$I_o < \frac{1}{\gamma \ln 2} 2^{-(k-t_c-\ell-r)} \tag{10}$$

where k is the length of the raw key string shared by legitimate users after a completing of protocol, t_c is the Renyi (collision) information obtained by Eve, r is the number of check symbols sent from Alice to Bob in order to reconcile their key strings, ℓ is the length of the final key string after hashing, γ is a coefficient that approaches 0.42 for any fixed r, as k, ℓ and $k - \ell$ increase.

Since we assume that Eve is not nearby legal users, she has no eavesdropping at all, hence t_c can be removed ($t_c = 0$).

The probability P_d of error after decoding by some linear binary error correcting code with the number of information bits k, the number of check symbols r and for the probability of bit error after a completing at protocol *P'* has the following upper bound [29]

$$P_d \leq 2^{-k(1-R)} \left(1 + 2\sqrt{P'(1-P')}\right)^k \tag{11}$$

where $R = \frac{k}{k+r}$.

Using the formulas (10), (11) we can optimize the parameter r to provide the requested values I_o and P_d.

But of course for practical implementation it is necessary to use some constructive methods of encoding and decoding, say for the thing, the LDPC codes [30].

V. Conclusion

In the current paper we considered some extension of key sharing protocol proposed in [24]. It has been proved that key extraction can be performed not only from matrix eigenvalues but from matrix traces also. Moreover the extracted key bits occur for the last case even closer to pseudo random sequence in terms of NIST tests. But unfortunately the length of key strings is significantly less in the last case in comparison with extraction the key from matrix eigenvalues. Therefore this method is superior for practical implementation against matrix trace-based extraction.

We investigated how affect such parameters of key sharing protocol as the number of antennas, SNR in the legitimate channel and method of quantization. It was striked that key sharing protocol does not work if eavesdroppers has the same communication channels as legitimate users!

In fact it would be very strange to be the case because then legitimate users could share secret key without a presence of any fading channels and they could simply communicate through any channels with constant parameters.

We believe that key sharing between mobile unit is a promising approach because nothing restrictions on eavesdropping channels are suggested except of nearby locations of eavesdroppers against legal users.

The future work can be devoted to a modification of quantization procedures for the case of extraction from eigenvalues and investigation of constructive encoding and decoding for the most effective error correction in the shared key string.

References

[1] W. Diffie and M. E. Hellman, "New directions in cryptography," vol. 22, no. 6, pp. 644–654, 1976.

[2] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, ser. The CRC Press series on discrete mathematics and its applications. 2000 N.W. Corporate Blvd., Boca Raton, FL 33431-9868, USA: CRC Press, 1997. ISBN 0-8493-8523-7

[3] B. Alpern and F. B. Schneider, "Key exchange using 'keyless cryptography'." *Inf. Process. Lett.*, vol. 16, no. 2, pp. 79–81, 1983. [Online]. Available: http://dblp.uni-trier.de/db/journals/ipl/ipl16.html#AlpernS83

[4] M. M. Yung, "A secure and useful "keyless cryptosystem"," vol. 21, no. 1, pp. 35–38, Jul. 1985.

[5] A. Wyner, "Wire-tap channel concept," *Bell System Technical Journal*, vol. 54, pp. 1355–1387, 1975.

[6] A. Carleial and M. Hellman, "A note on wyner's wiretap channel (corresp.)," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 387–390, May 1977. doi: 10.1109/TIT.1977.1055721

[7] U. Maurer, "Secret key agreement by public discussion from common information." *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 733–742, 1993.

[8] I. Csiszár and J. Körner, "Broadcast channel with confidential messages." *IEEE Transactions on Information Theory*, vol. 24, no. 2, pp. 339–348, 1978.

[9] V. Korjik and V. Yakovlev, "Non-asymptotic estimates for efficiency of code jamming in a wire-tap channel," *Problems of Information Transmission*, vol. 17, pp. 223–22, 1981.

[10] L. H. Ozarow and A. D. Wyner, "Wire-tap channel II," in *Advances in Cryptology: Proceedings of EUROCRYPT 84, A Workshop on the Theory and Application of of Cryptographic Techniques, Paris, France, April 9-11, 1984, Proceedings*, 1984. doi: 10.1007/3-540-39757-4_5 pp. 33–50. [Online]. Available: https://doi.org/10.1007/3-540-39757-4_5

[11] V. Korjik and D. Kushnir, "Key sharing based on the wire-tap channel type ii concept with noisy main channel," in *Proc. Asiacrypt96*. Springer Lecture Notes in Computer Science 1163, 1996, pp. 210–217.

[12] V. Yakovlev, V. I. Korzhik, and G. Morales-Luna, "Key distribution protocols based on noisy channels in presence of an active adversary: Conventional and new versions with parameter optimization," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2535–2549, 2008.

[13] V. Korjik and M. Bakin, "Information-theoretically secure keyless authentication," in *Proc. IEEE Symp. on IT'2000*. IEEE, 2000, p. 20.

[14] C. H. Bennett, F. Bessette, G. Brassard, L. Salvail, and J. Smolin, "Experimental quantum cryptography," *J. Cryptol.*, vol. 5, no. 1, pp. 3–28, Jan. 1992. [Online]. Available: http://dl.acm.org/citation.cfm?id=146395.146396

[15] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," in *Proceedings of International Conference on Computers, Systems and Signal Processing*, December 1984.
[16] Z. Li, W. Trappe, and R. Yates, "Secret communication via multi-antenna transmission," in *Information Sciences and Systems, 2007. CISS '07. 41st Annual Conference on*, March 2007. doi: 10.1109/CISS.2007.4298439 pp. 905–910.
[17] J. W. Wallace and R. K. Sharma, "Automatic secret keys from reciprocal MIMO wireless channels: measurement and analysis." *IEEE Trans. Information Forensics and Security*, vol. 5, no. 3, pp. 381–392, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/tifs/tifs5.html#WallaceS10
[18] V. Yakovlev, V. Korzhik, P. Mylnikov, and G. Morales-Luna, "Outdoor secret key agreement scenarios using wireless MIMO fading channels," vol. 14, pp. 1–25, 01 2017.
[19] T. Aono, K. Higuchi, T. Ohira, B. Komiyama, and H. Sasaoka, "Wireless secret key generation exploiting reactance-domain scalar response of multipath fading channels," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 11, pp. 3776–3784, 2005.
[20] V. Yakovlev, V. I. Korzhik, Y. Kovajkin, and G. Morales-Luna, "Secret key agreement over multipath channels exploiting a variable-directional antenna," *Int. Jour. Adv. Computer Science & Applications*, vol. 3, no. 1, pp. 172–178, 2012.
[21] T. Dean and A. Goldsmith, "Physical-layer cryptography through massive MIMO," in *2013 IEEE Information Theory Workshop, ITW 2013, Sevilla, Spain, September 9-13, 2013*, 2013. doi: 10.1109/ITW.2013.6691222 pp. 1–5. [Online]. Available: http://dx.doi.org/10.1109/ITW.2013.6691222
[22] R. Steinfeld and A. Sakzad, "On massive mimo physical layer cryptosystem," in *2015 IEEE Information Theory Workshop - Fall (ITW)*, Oct 2015. doi: 10.1109/ITWF.2015.7360782 pp. 292–296.
[23] V. Korzhik, V. Starostin, and K. Akhrameeva, "Investigation of keyless cryptosystem proposed by Dean and Goldsmith," in *2017 21st Conference of Open Innovations Association (FRUCT)*, Nov 2017. doi: 10.23919/FRUCT.2017.8250182 pp. 194–201.
[24] D. Qin and Z. Ding, "Exploiting multi-antenna non-reciprocal channels for shared secret key generation," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2693–2705, Dec 2016. doi: 10.1109/TIFS.2016.2594143
[25] W. Feller, *An introduction to probability theory and its applications. Volume 1*, ser. Wiley series in probability and mathematical statistics. New York, Chichester, Brisbane: John Wiley & sons, 1968. ISBN 0-471-25711-7. [Online]. Available: http://opac.inria.fr/record=b1122219
[26] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001. ISBN 0130422320
[27] D. Dasgupta, A. Roy, and A. Nag, *Advances in User Authentication*, 1st ed. Springer Publishing Company, Incorporated, 2017. ISBN 3319588060, 9783319588063
[28] L. E. Bassham, III, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, E. B. Barker, S. D. Leigh, M. Levenson, M. Vangel, D. L. Banks, N. A. Heckert, J. F. Dray, and S. Vo, "Sp 800-22 rev. 1a. a statistical test suite for random and pseudorandom number generators for cryptographic applications," Gaithersburg, MD, USA, Tech. Rep., 2010.
[29] V. Korjik, G. Morales-Luna, and V. Balakirsky, "Privacy amplification theorem for noisy main channel," *Lecture Notes in Computer Science*, vol. 2200, pp. 18–26, 2001.
[30] K. Shalkoska, *Implementation of LDPC Algorithm: In C Programming Language*. LAP LAMBERT Academic Publishing, 2017. ISBN 9783330026049. [Online]. Available: https://books.google.com.mx/books?id=1yNcMQAACAAJ

APPENDIX

Proof of relation (7)

Let us consider an extraction of the key based on matrix traces. Assume that the elements of both channel matrices $P = [p_{ij}]_{1\le i,j\le n}$, $Q = [q_{ij}]_{1\le i,j\le n}$ are random, mutually independent and identically distributed: $p_{ij}, q_{ij} \sim \mathrm{CN}(0, \sigma_w^2)$. Similarly these conditions hold and for the noise matrices $N_1 = [n_{ij1}]_{1\le i,j\le n}$, $N_2 = [n_{ij2}]_{1\le i,j\le n}$: $n_{ij1}, n_{ij2} \sim \mathrm{CN}(0, \sigma_e^2)$. We admit also that channel matrices and noisy are mutual independent. The relation (3) entails

$$\begin{aligned} YX^{-1} &= PQ + PN_1X^{-1} + N_2X^{-1} \quad \text{and} \\ \mathrm{Tr}(YX^{-1}) &= \mathrm{Tr}(PQ) + \mathrm{Tr}(PN_1X^{-1}) + \mathrm{Tr}(N_2X^{-1}). \end{aligned}$$

It is easy to show that for large number of antennas ($n \gg 1$) due to Central Limit Theorem, the random variables

$$Z_A = \mathrm{Tr}(Y_{A2}X_A^{-1}) \quad , \quad Z_B = \mathrm{Tr}(Y_{B2}X_B^{-1})$$

have Gaussian distributions:

$$f_A(z) = f_B(z) = \frac{1}{\pi\sigma^2} e^{-\frac{|z|^2}{\sigma^2}} \tag{12}$$

where $\sigma^2 = DZ_A = DZ_B = n^2\sigma_w^2(\sigma_w^2 + \sigma_e^2)$.

Let us estimate the dependence of the random variables Z_A, Z_B using the notion of linear regression Z_A onto Z_B:

$$Z_B - E(Z_B) = \gamma \frac{\sigma_A}{\sigma_B} (Z_A - E(Z_A))$$

where

$$\gamma = \frac{1}{\sqrt{(DZ_A)(DZ_B)}} \mathrm{cov}\,(Z_A, Z_B)$$

is a correlation coefficient.

Since Z_A, Z_B are centered random variables with equal variances, the equation of linear regression Z_A onto Z_B has the form

$$Z_B = \gamma Z_A. \tag{13}$$

It is easy to show that $\mathrm{cov}\,(Z_A, Z_B) = n^2\sigma_w^2$. Thus we get

$$\begin{aligned} \gamma &= \frac{n^2\sigma_w^2}{n^2\sigma_w^2(\sigma_w^2 + \sigma_e^2) + n\sigma_e^2} \\ &= \left(1 + \frac{\sigma_e^2}{\sigma_w^2}\left(1 + \frac{1}{n\sigma_w^2}\right)\right)^{-1} \end{aligned} \tag{14}$$

Since the correlation coefficient γ is real-valued, it results that the random values Z_A, Z_B differ by modulus only.

If $n\sigma_w^2 \gg 1$ and the noise-to- signal ratio $\frac{\sigma_e^2}{\sigma_w^2}$ is small, then we get by (14)

$$\gamma = \frac{1}{1+\alpha} \approx 1 - \alpha \ , \ \alpha = \frac{\sigma_e^2}{\sigma_w^2}\left(1 + \frac{1}{n\sigma_w^2}\right) \approx \frac{\sigma_e^2}{\sigma_w^2} \ll 1.$$

Thus the dependence (13) between Z_A, Z_B is almost linear.

In order to get a uniformly distributed key, let us quantize the range of values Z_A (on the complex plane) in radial direction in such a way that the probability to hit Z_A into each of N rings $R_k = \{z \in \mathbb{C} |\ r_{k-1} \le |z| < r_k\}$, $r_0 = 0$, $r_N = +\infty$, occurs equally likely:

$$\Pr\,(r_{k-1} \le |z| < r_k) = \frac{1}{N} =: p \text{ for } k = 1, \ldots, N \tag{15}$$

Using (12) we are able to find the radial distribution function of Z_A:

$$F(r) = \Pr\,(|z| < r) = 1 - e^{-\frac{r^2}{\sigma^2}}$$

Thus (15) holds if and only if

$$\begin{aligned}\Pr\left(r_{k-1} \le |z| < r_k\right) &= F(r_k) - F(r_{k-1}) \\ &= e^{-\frac{r_{k-1}^2}{\sigma^2}} - e^{-\frac{r_k^2}{\sigma^2}} \\ &= p\end{aligned}$$

It results the relation

$$F(r_k) = kp = 1 - e^{-\frac{r_k^2}{\sigma^2}} \tag{16}$$

Eventually we get $r_k = \sigma\sqrt{-\ln(1-kp)}$.

Let us estimate now the probability of key coincidence for both legitimate users A and B. First we estimate the probability p_k to get Z_A and Z_B in the ring R_k. Taken into account that the dependence (13) is almost linear $Z_B \approx \gamma Z_A$, where $0 < \gamma \le 1$, we get $|Z_B| = \gamma|Z_A|$. Hence

$$\begin{aligned}p_k &= \Pr\left(Z_A \in R_k \ \&\ Z_B \in R_k\right) \\ &= \Pr\left(r_{k-1} \le |Z_A| < r_k \ \&\ r_{k-1} \le |Z_B| < r_k\right) \\ &= \Pr\left(r_{k-1} \le |Z_A| < r_k \ \&\ r_{k-1} \le \gamma|Z_A| < r_k\right) \\ &= \Pr\left(r_{k-1} \le |Z_A| < r_k \ \&\ \frac{r_{k-1}}{\gamma} \le |Z_A| < \frac{r_k}{\gamma}\right) \\ &= \Pr\left(\frac{r_{k-1}}{\gamma} \le |Z_A| < r_k\right).\end{aligned}$$

Using (16), we find that

$$\begin{aligned}p_k &= F(r_k) - F\left((\frac{r_{k-1}}{\gamma}\right) \\ &= e^{-\frac{r_{k-1}^2}{\gamma^2\sigma^2}} - e^{-\frac{r_k^2}{\sigma^2}} \\ &= (1-(k-1)p)^{\frac{1}{\gamma^2}} - (1-kp)\end{aligned}$$

Then the probability that even legal users get the same key (trace) under the condition $\gamma > \gamma_{cr} = \frac{r_{k-1}}{r_k}$ is equal to

$$\begin{aligned}p' &= \sum_{k=1}^{N} p_k \\ &= \sum_{k=1}^{N} \left((1-(k-1)p)^{\frac{1}{\gamma^2}} - (1-kp)\right).\end{aligned}$$

It is worth to note that a quantization problem of the matrix trace (in the case when legal users extract the key namely from it) can be solved trivially because the distribution (12) is independent of the "angle variable". This is valid also for all coefficients of characteristic polynomial including matrix eigenvalues. In fact, it is a consequence of circular symmetry of channel matrices and matrices of noises.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 577–583
DOI: 10.15439/2018F38
ISSN 2300-5963 ACSIS, Vol. 15

MATLAB Implementation of an Adaptive Neuro-Fuzzy Modeling Approach applied on Nonlinear Dynamic Systems – a Case Study

Roxana-Elena Tudoroiu
University of Petrosani
20 Universității Street,
332006, Petroșani,
Romania
tudelena@mail.com

Mohammed Zaheeruddin
Concordia University 1455
De Maisonneuve Blvd West,
Montreal, QC, H3G 1M8,
Canada
zaheer@encs.concordia.ca

Nicolae Tudoroiu
John Abbott College 21275
Lake Shore Road, Sainte-Anne-de-Bellevue, QC, H9X 3L9,
Canada
ntudoroiu@gmail.com

Dumitru Dan Burdescu
University of Craiova, 107
Decebal Bvd., 200440, Craiova,
Romania
dburdescu@yahoo.com

***Abstract*—In this paper one of the most accurate adaptive neuro-fuzzy modelling approach is investigated. It is suitable for modelling the nonlinear dynamics of any process or control systems. Basically, this new modelling approach is an improvement of a linear ARX polynomials models based on the least square errors estimation method that is preferred for its simplicity and faster implementation, since it uses typical functions from MATLAB system identification toolbox. For simulation purpose, to prove its effectiveness in terms of modeling accuracy, an appropriate case study of a centrifugal chiller is considered. The reason for this selection is given by the fact that the centrifugal chiller control system is one of the most seen in a large variety of applications in HVAC control systems. Since its dynamic model is of high complexity in terms of dimension and encountered nonlinearities, a tight control in closed-loop requires a suitable modelling approach**.

I. Introduction

THIS research paper investigates an alternative modelling design methodology for nonlinear systems dynamics, such as an adaptive neuro-fuzzy logic modeling approach, very useful for a large number of control systems applications from different fields, including also multimedia networks and communications. Basically, the new modeling approach is a combination of an artificial neural network (ANN) [1]-[3] and a fuzzy logic (FL) [4]-[6] modeling features. Nowadays considerable advances have been made in applying the ANN systems for problems found intractable or difficult for traditional computation [1]. Some representative preliminary results obtained and related to this field in our research work activity during the years can be found in [3]. The fuzzy logic modelling technique is a powerful tool for the formulation of expert knowledge and the combination of imprecise information from different sources [4]. The FL is in fact a control system modelling technique for a "*complicated system without knowledge of its mathematical description*", as is also stated in [4]. Fuzzy logic modelling technique was applied successfully in our research activity to design an improved hybrid fuzzy sliding mode observer estimator [5]. For simulation purpose and "proof-concept" considerations as a case study is chosen one relevant application from heating ventilation and air conditioning (HVAC) control systems, namely a centrifugal chiller system, that is one of the most widely used in this kind of applications. It is characterized by a great complexity and high nonlinear behavior, as is shown with many details in [7]. As is shown in [7], the centrifugal chillers have become the most widely used devices since they have high capacity, reliability, and require low maintenance. Furthermore, in this recent research paper work is completed a literature review in the field that reveals a significant amount of work done in a classic way on transient and steady state modelling for centrifugal chillers, such is revealed in [8]-[16]. Additionally, we try to complete in the following the literature review with the most recent sophisticated and intelligent approaches related directly to a combined neural networks and fuzzy logic or separately modelling methodologies applied to centrifugal chillers. An interesting integrated modelling methodology is used in [17] to improve the reconstruction of the performance map of axial compressor and fans, where the learning capability of ANN is integrated to the knowledge aspect of fuzzy inference system (FIS) to offer enhanced prediction capabilities rather than using a single methodology independently.

In [18] an application of combined neuro-fuzzy modelling techniques to develop a fault detection, diagnosis and isolation (FDDI) strategy for centrifugal chillers is presented.

A new modelling approach of steady state vapour-compression liquid chillers is presented in [19] that uses a generalized radial basis function (GRBF) ANN to predict chiller performance. As is mentioned also in [7] that centrifugal chillers are the most energy-consuming devices in HVAC applications, especially if they do not operate optimally, i.e. they cannot produce the required cooling load

This work was not supported by any organization

capacity, an improvement of their coefficient performance (COP) and to reduce the power consumption will be required. This objective is achieved in [20] by using for model prediction an ANFIS based Fuzzy Clustering Subtractive (FCS) and for classification and optimization an Accelerated Particle Swarm Optimization (APSO) algorithm.

The proposed technique "*reduces the total power consumption by 33.2% and meets the cooling demand requirements*", as is stated also in [20]. Also, "*it improves the cooling performance based on COP, thus resulting in a 15.95% increase in efficiency compared to the existing cooling system*". The studied ANFIS-based FCS outperforms the ANFIS-based fuzzy C-means clustering in terms of the regression. Then, the algorithm-based classifier APSO has better results compared to the conventional particle swarm optimization (PSO).

Thus, it is important to explore new modelling methodologies for HVAC centrifugal chillers dynamic systems. With this as motivation, the remainder of the paper is structured as follows. In Section 2 the ARX model of the centrifugal chiller based on the measured input-output measurements data set obtained by extensive simulations in open-loop is presented. In Section 3 the simulation results of both closed-loop control subsystems, evaporator and condenser based on ARX models built in Section 2 are shown. In Section 4 the neuro – fuzzy model of centrifugal chiller based on the same open-loop input-output measurements data set used in Section 2 is presented. In Section 5 performance analysis of both models are compared in terms of modeling accuracy. Finally, the Section 6 concludes the relevant contributions of this research paper.

II. THE ARX MODELS OF CENTRIFUGAL CHILLER

A. The Simulink Model of the Open Loop Centrifugal Chiller

Basically, a centralized centrifugal chiller control system can be considered as an interconnection of two main closed-loops control subsystems, the first one is a chilled water temperature control loop inside an evaporator, and the second one is a refrigerant liquid level control loop inside a condenser [7]. Since its dynamic model is of high complexity in terms of dimension and encountered nonlinearities, a tight control of the both closed-loops requires a suitable modelling approach. The centrifugal chiller efficiency can be improved by implementing advanced model-based controller design strategies. Consequently, the development of high-fidelity centrifugal chiller dynamic model has become a priority task of our research. For interested readers a complete dynamic model of centrifugal chiller under our consideration is given in Annex 1 of our research work [7], pp. 299-305.

In first step of the control design, based on the mentioned modelling development, a MATLAB SIMULINK model for a centrifugal chiller centralized system in open-loop is built, as is shown in Fig. 1 and Fig. 2, similar to those introduced in [7], p. 286.

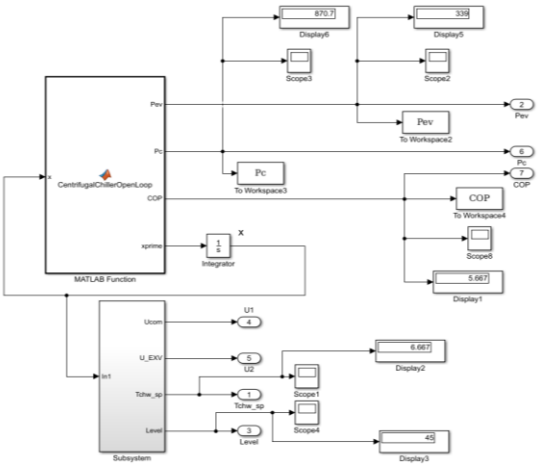


Fig.1 SIMULINK model of centralized centrifugal chiller in open-loop (see also (7), p. 286)

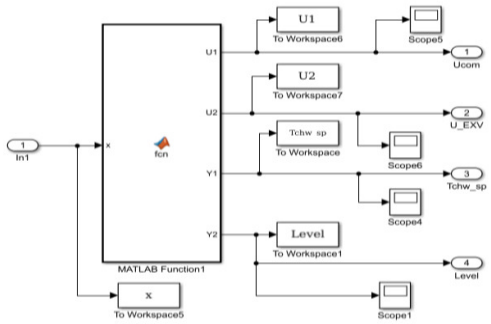


Fig.2 SIMULINK model of the bottom Subsystem block *(*see also (7), p. 286*)*

The SIMULINK centrifugal chiller open-loop model is an useful architecture structure support for ARX models, and also to build several closed-loop control strategies as the one based on two single input - single output (SISO) ARX models described in the next section for the overall system, in fact a distributed system, since in "real-life" there exist some interferences between the both loops [7]. Through extensive open-loop simulations is generated the most appropriate input-output data set required to build the linear SISO polynomial ARX models for both open-loops of the centrifugal chiller, as is shown in Fig. 3 for chilled water temperature and Fig.4 for refrigerant liquid level respectively. Consequently, a good open-loop model capable to capture entire dynamics of the overall chiller system under various operating conditions is essential to give more flexibility for closed-loop control design strategies.

B. The SISO ARX Model of Temperature and Liquid Refrigerant Level Open-Loops

Basically, the *arx* MATLAB *function* [7]–[8] estimates the parameters of two polynomials discrete-time models known as autoregressive with an exogenous input (ARX) and a more simple polynomial autoregressive without exogenous input (AR) that estimates the parameters of the scalar time

series, by means of the well-known least squares method the most used in control systems identification and parameters estimation, such as those well documented in [7]-[8]. Theoretically, the *arx* MATLAB function uses a prediction-error method and specified polynomial orders [7]-[8]. Also, a pure transport delay of the signal flow in the feedback path from the measurement sensors to the controllers is specified in each ARX structure. The ARX model is "*inherently linear and the most significant advantage is that we can perform model structure and parameter identification rapidly*", as is stated in [9]. To get the ARX model the data set is divided in two segments, first segment consisting of 1000 samples required for the prediction phase, and the second one consisting of the remaining 2600 samples required for validation phase respectively. The information obtained on the discrete-time chilled water temperature ARX model in MATLAB command window is:

$$A(z)y(t) = B(z)u(t) + e(t) \tag{1}$$

where the coefficients of the polynomials $A(z)$, $B(z)$ are estimated by the well-known least square errors (LSE) procedure, given in the next two equations:

$$A(z) = 1 - 1.096z^{-1} - 0.7723z^{-2} + 0.8689z^{-3} \tag{2}$$

$$\begin{aligned} B(z) &= 7.98z^{-2} - 8.074z^{-3} - 8.775z^{-4} + 8.31z^{-5} + \ldots \\ &2.009z^{-6} - 1.44z^{-7} + 0.002313z^{-8} - 0.009972z^{-9} + \ldots \\ &-0.002079z^{-10} + 0.001031z^{-11} \end{aligned} \tag{3}$$

Sample time: 1 second

Parameterization:

Polynomial orders:

$n_a = 3$ ($A(z)$ degree), $n_b = 10$ ($B(z)$ degree),

$n_k = 2$ (delay)

Number of free coefficients: 13

Status:

Estimated using ARX on time domain data.

Fit to estimation data: 100% (prediction focus)

FPE: 2.821e-11, MSE: 2.694e-11

The argument variable z^{-1} is the equivalent of the time-discrete delay operator q^{-1}, as is shown in [7]-[8], i.e. $z^{-1}y(t) = q^{-1}y(t) = y(t-1)$, and $e(t)$ is an additive white noise.

The simulation results obtained in both prediction phase (in the top graph side) and validation phase respectively (in the bottom graph side) are shown in Fig.5. These results reveal a good performance in the both phases for chilled water temperature. Similarly, for refrigerant liquid level SISO ARX open-loop the following information provided in MATLAB command window is very useful for controller design:

$$A(z)y(t) = B(z)u(t) + e(t) \tag{4}$$

$$\begin{aligned} A(z) &= 1 - 2.02z^{-1} + 0.5349z^{-2} + 0.8555z^{-3} + \ldots \\ &-0.2345z^{-4} + 0.08802z^{-5} - 0.4596z^{-6} + \ldots \\ &0.1761z^{-7} + 0.131z^{-8} - 0.07358z^{-9} + \ldots. \\ &0.001818z^{-10} \end{aligned} \tag{5}$$

$$\begin{aligned} B(z) &= -15.85z^{-1} + 19.63z^{-2} + 4.138z^{-3} - 3.039z^{-4} + \ldots \\ &-8.471z^{-5} - 0.677z^{-6} + 5.626z^{-7} - 0.2662z^{-8} + \ldots \\ &-1.1z^{-9} \end{aligned} \tag{6}$$

Sample time: 1 seconds

Parameterization:

Polynomial orders:

$n_a = 10$ ($A(z)$ degree), $n_b = 9$ ($B(z)$ degree),

$n_k = 1$(delay)

Number of free coefficients: 19

Status:

Estimated using ARX on time domain data.

Fit to estimation data: 99.98% (prediction focus)

FPE: 6.911e-08, MSE: 6.502e-08

The simulation results obtained in both prediction phase (in the top graph side) and validation phase respectively (in the bottom graph side) are shown in Fig.6. These results reveal a poor performance in the both phases for refrigerant liquid level, caused by the presence of the oscillations around the level set point of 45% of the liquid in Condenser.

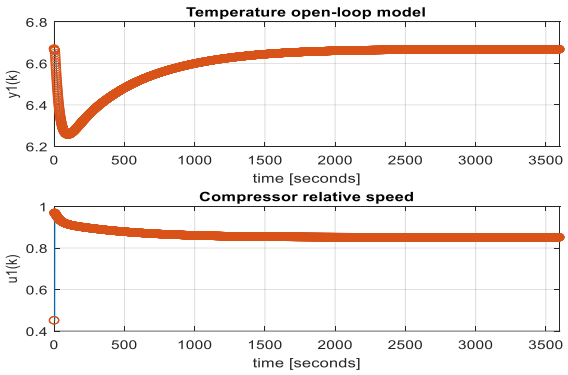


Fig.3 Temperature open-loop model

Concluding, the performance in Fig. 7 and Fig.8 appears to be very good for chilled water temperature, and worst for refrigerant liquid level. However, if a better performance level is desired, we might want to switch to a nonlinear model. In particular, we are going to use a neuro-fuzzy modeling approach ANFIS, to see if we can push the performance level of an off-line trained ANN with a fuzzy inference system (FIS), as can be seen in Section 4.

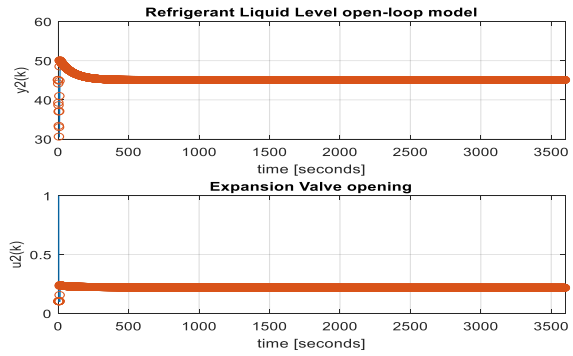


Fig.4 Refrigerant liquid level open-loop model

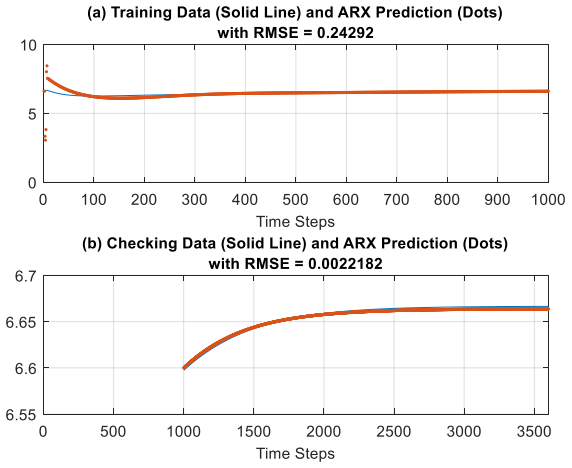


Fig.5 Temperature SISO ARX open-loop model

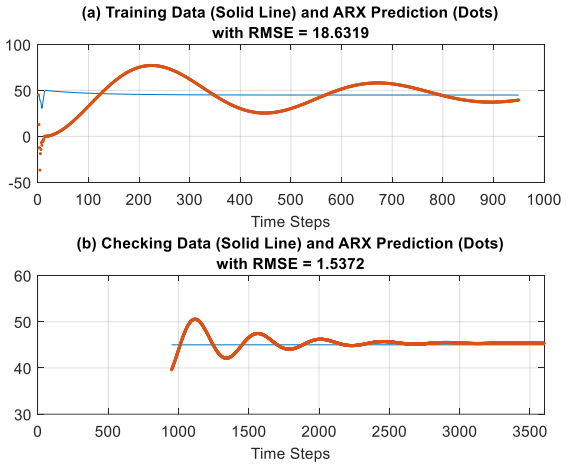


Fig.6 Refrigerant liquid level SISO ARX open-loop model

III. CENTRIFUGAL CHILLER SISO ARX MODELS CLOSED-LOOP SIMULATION RESULTS

In this section we show the great potential of both ARX models for control design in feedback closed-loop. Precisely, we show the results of the standard proportional integral derivative (PID) control used in each closed-loop to control the chilled water temperature and the refrigerant liquid level. The main idea is that even if the ARX models performance is not the best one, but it is satisfactory, two well-tuned PID controllers, i.e. their parameters $k_P, k_I,$ and k_D have optimal values, as is shown also in [7], compensate the modeling mismatches, by performing very well in closed-loop. The Simulink model of the both closed-loops is shown in Fig. 7, and the simulation results that reveal a very good performance in terms of robustness to the changes in set points and convergence with a very fast transient are shown in Fig. 8, for chilled water temperature, and Fig.9 for refrigerant liquid level.

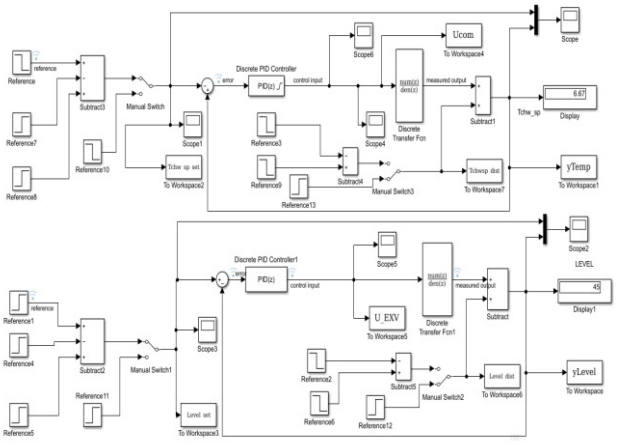


Fig.7 SIMULINK SISO ARX model of centrifugal chiller in closed-loop

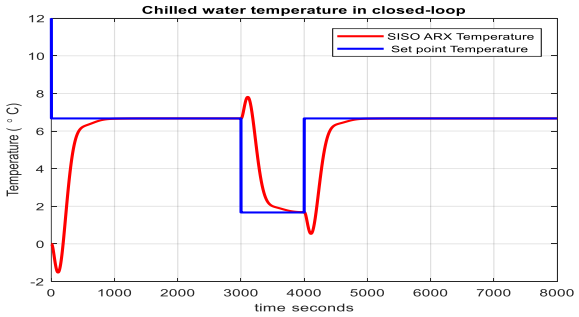


Fig.8 Chilled water Temperature SISO ARX closed-loop model

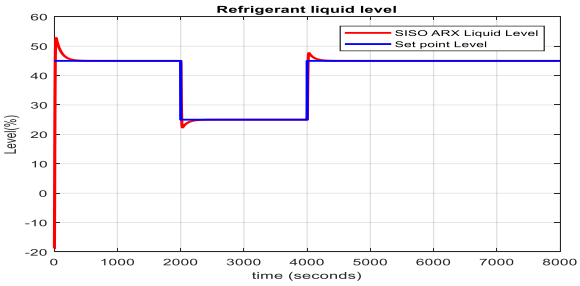


Fig.9 Refrigerant liquid level SISO ARX closed-loop model

IV. THE ANFIS MODEL OF CENTRIFUGAL CHILLER

Basically, the architecture of an adaptive neuro-fuzzy inference system (ANFIS) attached to the centrifugal chiller nonlinear model is a combination of an ANN that drives a fuzzy inference system (FIS). The ANN is trained online or offline based on an input-output measurements data set collected as a result of extensive open-loop simulations to find the most appropriate nonlinear relationship between the inputs and the outputs of the both open-loops of centrifugal chiller control system.

Essentially, the proposed ANN structure has the following three main features:

- Ability to find the solution of such problems for which algorithmic method is expensive or does not exist
- Ability to learn by experience, thus is not need to program them at much extent
- High accuracy and a considerable fast processing speed than traditional systems.

The proposed ANN computation is performed by "*of dense computing mesh nodes and connections*" [1], as is shown in Fig. 10 [2]. It is a driven data based, operating "*collectively and simultaneously*" [1], [3] to learn (e.g. in supervised or un-supervised learning mode) how to match the input and output training (offline or online) measured data set (prediction phase) of the centrifugal chiller control system under consideration, thus to establish an accurate relationship (i.e. model), validated on the supplemental segment data set (validation phase). The basic processing elements of the proposed ANN structure are called artificial neurons, or simply nodes, that typically operate in parallel and are configured in regular architectures, similar to those shown in Fig.10.

Within the ANN architecture the neurons perform similar to summing and nonlinear mapping junctions.

Basically, the ANN architecture structure contain input, hidden and output layers, as well as several feedback connections within the layer and toward adjacent layers.

Each connection strength is expressed by a numerical value called weight, which can be modified during the online or offline training in a prediction phase of computational process.

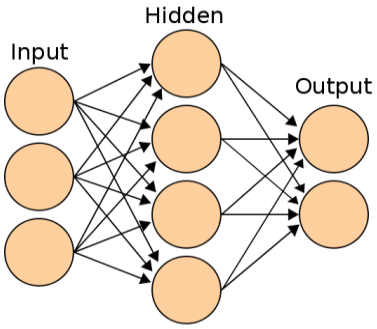


Fig.10 A simple ANN architecture (reproduced from [2])

Fuzzy logic modelling technique is a "*computational paradigm that is based on how humans think*", as is stated in [6]. According also to [6] in fuzzy logic, any statement assumes a probabilistic value between 0 and 1, representing the membership degree of an element that belongs to a given set.

The first step in ANFIS modeling system identification is the input selection to determine which variables should be the input arguments to the ANFIS model. For simplicity, as is suggested in [9] we assume that there are 10 input candidates for the ANN [3]:

$(y(k-1), y(k-2), y(k-3), y(k-4))$,

$(u(k-1), u(k-2), u(k-3), u(k-4), u(k-5), u(k-6))$

and the output to be predicted is $y(k)$. A heuristic approach to input selection is called sequential forward search, in which each input is selected sequentially to optimize the total squared error (RMSE). This can be done by using the MATLAB function *seqsrch*; the results are shown in Fig. 11 for chilled water temperature, and Fig.12 for refrigerant liquid level respectively.

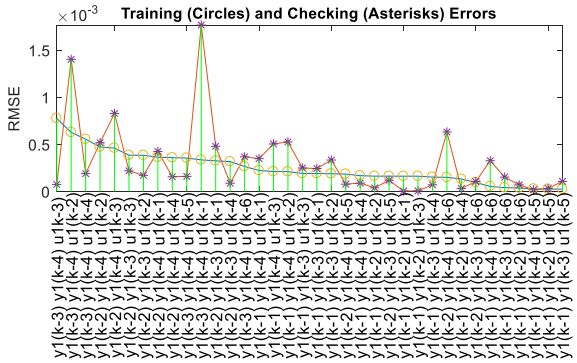


Fig.11 Temperature SISO ANFIS open-loop model training phase

Legend: ANFIS info
Number of nodes: 34
Number of linear parameters: 32
Number of nonlinear parameters: 18
Total number of parameters: 50
Number of training data pairs: 1000
Number of checking data pairs: 2601
Number of fuzzy rules: 8
Minimal training RMSE = 0.000025
Minimal checking RMSE = 0.000120906

$[\, n_a \;\; n_b \;\; n_k \,] = 10 \;\; 9 \;\; 1$

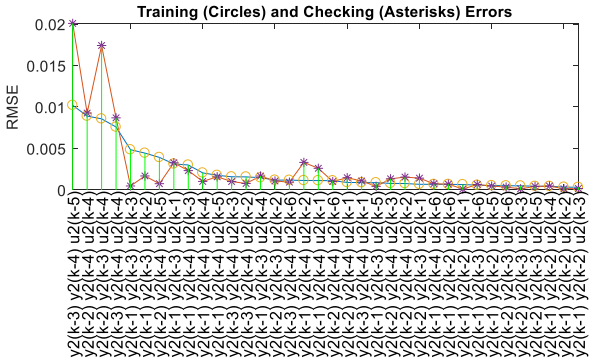


Fig.12 Refrigerant liquid level SISO ANFIS open-loop model training phase

Legend: ANFIS info
Number of nodes: 34
Number of linear parameters: 32
Number of nonlinear parameters: 18
Total number of parameters: 50
Number of training data pairs: 950
Number of checking data pairs: 2651
Number of fuzzy rules: 8
Minimal training RMSE = 0.000320
Minimal checking RMSE = 0.000274694

$$[\,n_a \;\; n_b \;\; n_k\,] = 10 \;\; 9 \;\; 1$$

The inputs are selected with a training RMSE of 0.000320, and checking RMSE of 0.000120906, for chilled water temperature, and a training RMSE of 0.000274694, for refrigerant liquid level respectively.

The simulation results for the estimated nonlinear centrifugal chiller ANFIS models in open-loop are shown in Fig.13, for chilled water temperature, and Fig.14 for refrigerant liquid level.

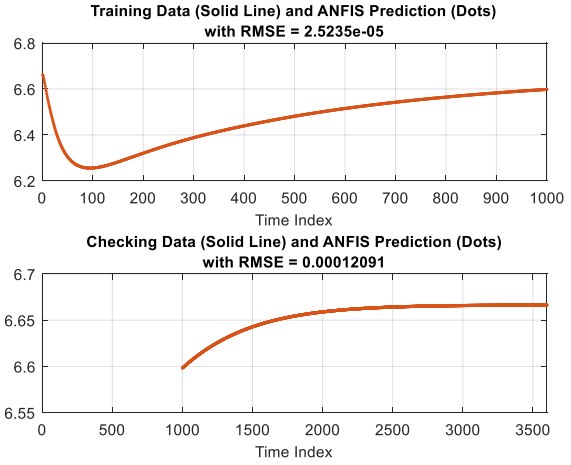


Fig.13 Temperature SISO ANFIS open-loop model

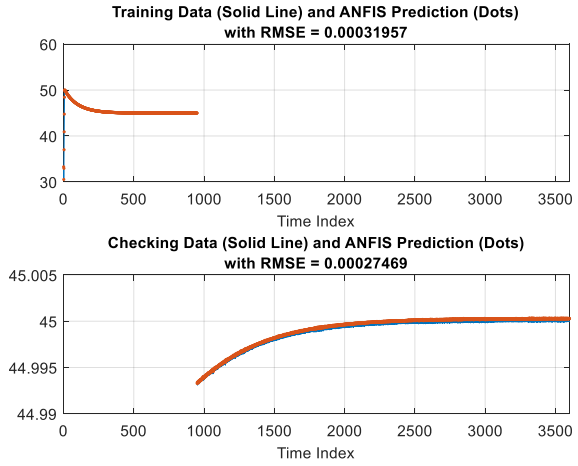


Fig.14 Refrigerant liquid level SISO ANFIS open-loop model

As mentioned in Section 2 for ARX models we follow the same procedure to split the original data set obtained by extensive open-loop simulations in two data subsets, the first one for training phase (the top side of the graphs), and the second one for prediction (estimation) phase (the bottom side of the graphs) respectively. These modeling aspects can be seen also in Fig.13 and Fig.14. The simulation results reveal a great accuracy modeling performance for the nonlinear ANFIS models compared to the performance of linear ARX models, and thus they have also a great potential for control design.

To simplify the ANFIS modeling design we suggest the use of the MATLAB function *anfis* provided by MATLAB Fuzzy Logic Toolbox, that is a hybrid learning algorithm to identify the membership function parameters of single output, Sugeno type FIS. A combination of LSE and backpropagation gradient descent methods [3] are used for training FIS membership function parameters [5] to model a given set of input-output data set, as is mentioned in MATLAB Fuzzy Logic Toolbox [8].

V. Conclusion

This research paper is an interesting MATLAB application of a modeling design approach that uses the most accurate nonlinear adaptive neuro-fuzzy ANFIS models, as an alternative to linear polynomial ARX models. If we want to choose between ARX and ANFIS models we need to think in terms of fast implementation or precision.

For fast implementation we can choose the ARX models since these are inherently linear and the most significant advantage is that we can perform model structure and parameter identification rapidly. If a high modeling accuracy performance is desired, we are going to use a neuro-fuzzy modeling approach ANFIS, to push the performance level with a fuzzy inference system. In the future work we will extend the investigations area to apply the nonlinear ANFIS models for many other similar applications in different fields, and to explore their great potential in closed-loop control systems, and also in sliding mode control [3], [10],[23].

References

[1] J.M. Zurada, *Introduction to artificial neural networks*, 1st ed. Ed. St. Paul, USA, MN: West Publishing company, 1992, ISBN 0-3 14-93391 -3.

[2] [Internet]. Wikipedia. Available online on the website: https://en.wikipedia.org/wiki/Artificial_neural_network, accessed at May 2nd 2018.

[3] M. Zaheeruddin, N. Tudoroiu, "Neuro - PID tracking control of a discharge air temperature system", Elsevier, Energy Conversion and Management, vol. 45, pp.2405–2415, 2004, DOI:10.1016/j.enconman.2003.11.016.

[4] B. Bouchon-Meunier, M. Detyniecki, M-J. Lesot, C. Marsala, M. Rifqi, "Real-World Fuzzy Logic Applications in Data Mining and Information Retrieval". Available on website at: https://pdfs.semanticscholar.org/b86c/2c39d3457bd439b2e280eff8134768fcbe90.pdf, accessed at May 2nd 2018.

[5] S.M. Radu, E-R Tudoroiu, W. Kecs, N. Ilias, N. Tudoroiu, "Real Time Implementation of an Improved Hybrid Fuzzy Sliding Mode Observer Estimator", *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 1, January 2017, pp.214-226, ISSN: 2415-6698, DOI: 10.25046/aj020126, www.astesj.com.

[6] Fuzzy Logic Tutorial, available online on the web site: http://www.massey.ac.nz/~nhreyes/MASSEY/159741/Lectures/Lec2012-3159741-FuzzyLogic-v.2.pdf.

[7] N. Tudoroiu, M. Zaheeruddin, S. Li, E-R. Tudoroiu, "Design and Implementation of Closed-loop PI Control Strategies in Real-time MATLAB Simulation Environment for Nonlinear and Linear ARMAX Models of HVAC Centrifugal Chiller Control Systems", *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 2, April 2018, pp.283-308, ISSN: 2415-6698, DOI: 10.25046/aj030233,www.astesj.com.

[8] S. Bendapudi, J. E. Braun, et al., "Dynamic Model of a Centrifugal Chiller System - Model Development, Numerical Study, and Validation", *ASHRAE Trans.*, vol. 111, pp.132-148, 2005.

[9] A. Beyene, H. Guven, et al., (1994)," Conventional Chiller Performances Simulation and Field Data", *International Journal of Energy Research*, vol.18, pp.391-399, 1994.

[10] J. E. Braun, J. W. Mitchell, et al., "Models for Variable-Speed Centrifugal Chillers", *ASHRAE Trans.* , New York, NY, USA, 1987.

[11] M. W. Browne, P. K. Bansal, "Steady-State Model of Centrifugal Liquid Chillers", *International Journal of Refrigeration*, vol. 21, no.5, pp. 343-358, 1998.

[12] J.M. Gordon, K. C. Ng, H.T. Chua, "Centrifugal chillers: thermodynamic modeling and a diagnostic case study", *International Journal of Refrigeration*, vol.18, no. 4, pp.253-257, 1995.

[13] Li Pengfei, Li Yaoyu, J. E. Seem, "Modelica Based Dynamic Modeling of Water-Cooled Centrifugal Chillers", *International Refrigeration and Air Conditioning Conference*, Purdue University, pp.1-8, 2010. *http://docs.lib.purdue.edu/iracc/1091*.

[14] P. Popovic, H. N. Shapiro, "Modeling Study of a Centrifugal Compressor", *ASHRAE Trans.,* Toronto, 1998.

[15] M. C. Svensson, "Non-Steady-State Modeling of a Water-to-Water Heat Pump Unit", in *Proceedings of 20th International Congress of Refrigeration*, Sydney, 1999.

[16] H. Wang, S. Wang, "A Mechanistic Model of a Centrifugal Chiller to Study HVAC Dynamics", in *Building Services Engineering Research and Technology*, vol. 21, no.2, pp. 73-83, 2000.

[17] M. Gholamrezaei and K. Ghorbanian, "Application of integrated fuzzy logic and neural networks to the performance prediction of axial compressors", J Power and Energy, vol. 229(8), pp. 928–947, 2015, DOI: 10.1177/0957650915596877.

[18] Q. Zhou, S. Wang, F. Xiao, "A Novel Strategy for the Fault Detection and Diagnosis of Centrifugal Chiller Systems", HVAC&R RESEARCH Journal, vol. 15(1), pp.57-75, 2009.

[19] D. J. Swider, M.W. Browne, P. Bansal, V. Kecman, "Modelling of vapour-compression liquid chillers with neural networks", Applied Thermal Engineering Journal, vol. 21(3), pp. 311-329, 2001, DOI: 10.1016/S1359-4311(00)00036-3.

[20] Elnazeer Ali Hamid Abdalla, Perumal Nallagownden, Nursyarizal Bin Mohd Nor, Mohd Fakhizan Romlie, Sabo Miya Hassan, "An Application of a Novel Technique for Assessing the Operating Performance of Existing Cooling Systems on a University Campus", Energies Journal, vol.11(4), 719, 2018; DOI:10.3390/en11040719.

[21] [Internet]. Available online on website: https://www.mathworks.com/help/ident/ref/armax.html, MATLAB R2017b Documentation. Accessed at February 3rd, 2018.

[22] [Internet]."ANFIS Model Identification", available online on website: https://www.mathworks.com/help/fuzzy/examples/nonlinear-system-identification.html#d119e1365. Accessed at March 25th, 2018.

[23] R-E Tudoroiu, W. Kecs, M. Dobritoiu, N. Ilias, S-V Casavela, N.Tudoroiu," Real-Time Implementation of DC Servomotor Actuator with Unknown Uncertainty using a Sliding Mode Observer", ACSIS, vol.8, pp.841-848,DOI: 10.15439/2016F95, Poland, 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 585–588
DOI: 10.15439/2018F182
ISSN 2300-5963 ACSIS, Vol. 15

Feature Extraction of Binaural Recordings for Acoustic Scene Classification

Sławomir K. Zieliński
Faculty of Computer Science, Białystok University of Technology, Białystok, Poland
Email: s.zielinski@pb.edu.pl

Hyunkook Lee
Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom
Email: h.lee@hud.ac.uk

***Abstract*—Binaural technology becomes increasingly popular in the multimedia systems. This paper identifies a set of features of binaural recordings suitable for the automatic classification of the four basic spatial audio scenes representing the most typical patterns of audio content distribution around a listener. Moreover, it compares the five artificial-intelligence-based methods applied to the classification of binaural recordings. The results show that both the spatial and the spectro-temporal features are essential to accurate classification of binaurally rendered acoustic scenes. The spectro-temporal features appear to have a stronger influence on the classification results than the spatial metrics. According to the obtained results, the method based on the support vector machine, exploiting the features identified in the study, yields the classification accuracy approaching 84%.**

I. Introduction

DUE to a growing popularity of binaural technology [1], large repositories of audio material with binaural sound will soon be created. This will inevitably give rise to challenges concerning the management of spatial audio content. The method proposed in this paper could potentially be used for automatic indexing, search and retrieval of binaural recordings according to their spatial properties, helping to manage future audio repositories.

Most of the studies in the area of acoustic scene classification (ASC) aim to identify an environment where a given scene was recorded [2]-[4]. Little work has been done towards the classification of the recordings according to their spatial characteristics. The key idea underlying this work is, therefore, to extract the features from binaural recordings and to develop a prototype classifier allowing for classification of the spatial properties of acoustic scenes.

Taking advantage from feeding binaural signals to the input of ASC algorithms does not constitute a new approach. Chu et al. developed an environment-aware robotic system equipped with binaural microphones [5]. Trowitzsch et al. demonstrated benefits from using a binaural signal processor for detection of environmental sounds [6]. More recently, such researchers as Han and Park, as well as Weiping at al., exploited binaural signals in their ASC algorithms submitted to the DCASE2017 Challenge [7], [8]. However, to the best of the authors' knowledge, no-one has yet attempted to classify spatial properties of auditory scenes evoked by binaural recordings.

This study extends and builds on the recent work by Zieliński [9]. In contrast to the aforementioned study, which was focused on the classification of five-channel surround sound recordings, the experiment described in this paper was devoted to the classification of binaural audio content.

II. Taxonomy of Basic Spatial Audio Scenes

Information provided at the output of the proposed classifier identifies one of the four basic spatial scenes, labeled as *FB*, *FF*, *BF*, and *BB*. These scenes constitute the typical distribution patterns of foreground and background audio content around the listener in the horizontal plane (see Table I). Foreground sound objects represent easily identifiable, important and clearly perceived audio sources, whereas background objects normally represent reverberant, unimportant, unclear, ambient, "foggy" and distant sound sources. A taxonomy of the acoustic scenes adopted in this study was inspired by Rumsey's simplified spatial audio scene-based paradigm [10].

Table I.
The Basic Spatial Audio Scenes

Acoustic Scene	Description
Foreground-Background (*FB*)	A listener perceives foreground audio content in the front and background content behind the head.
Foreground-Foreground (*FF*)	A listener is surrounded by foreground audio content.
Background-Foreground (*BF*)	A listener perceives background audio content in the front and foreground content behind the head.
Background-Background (*BB*)	A listener is surrounded by background audio content.

This work was supported by a grant S/WI/3/2018 from Bialystok University of Technology and funded from the resources for research by Ministry of Science and Higher Education.

III. Corpus of Binaural Recordings

In total 600 binaural recordings were gathered for the purpose of this experiment. Most of the selected excerpts were extracted from the recordings available in the Internet, while 28 recordings, which constitutes 4.7% of all the items, were obtained through a binaural processing of the commercially available 5.0 surround sound recordings. The gathered sound clips represented such recording genres as classical music, pop music, jazz, electronic music, nature, documentary, drama, ambient recordings, and film soundtracks. During the selection procedure, care was taken that each excerpt exemplified a single spatial scene (*FB*, *FF*, *BF* or *BB*). The recordings were annotated manually by the first author. The average duration of the acquired audio samples was equal to 20 seconds. The recordings were stored in uncompressed two-channel audio files with a sampling rate of 44.1kHz and a 16-bit resolution. The available recordings in the audio corpus were split into the two subsets intended for the training (75% of items) and validation purposes (25% of excerpts), respectively.

IV. Feature Extraction

In total 1012 features were extracted for the purpose of this study. They could be divided into two broad categories: spatial and spectro-temporal. An overview of the extracted features was given in Table II. The rms-based metrics and binaural cues were classified in this study as spatial features, whereas the spectral features, the Mel-frequency cepstral coefficients (MFCCs) and the discrete cosine transformed amplitude modulation spectrogram coefficients (DCT AMS) were categorized as the spectro-temporal metrics. The procedure used to extract the features was outlined below.

Let x and y denote the left and right ear signals of the binaural recordings, respectively. Some of the metrics were extracted directly from the above signals whereas the other features were calculated based on m and s signals, where $m = x + y$ and $s = x - y$. Prior to calculating the metrics, the signals were split into 20 ms time frames with a 10 ms overlap. In order to save the computation time the duration of the analyzed time-blocks of the recordings was reduced to 7 seconds.

For each time frame, a ratio between the rms values of the x and y signals was computed. This way the obtained descriptors constituted a crude approximation of the interaural level differences (*ILD*). Similarly, for every time frame, a ratio between m and s signals was also calculated. It was assumed by the authors that this ratio could also be considered to be a simple descriptor of spatial characteristics.

All the metrics, including those described in the remainder of the paper, were calculated for every time frame of the signals. Then, they were summarized using the absolute mean values and standard deviations. In order to account for temporal fluctuations of the rms ratio across the time frames, the standard *delta* metrics [11] were also computed in a similar way as explained above.

There are three fundamental cues responsible for the spatial perception of sound: interaural level difference (*ILD*), interaural time difference (*ITD*), and interaural coherence (*IC*) [1], [12]. These cues were computed separately for each output of a 40-channel gammatone filter bank using their corresponding rate-maps. The rate-maps constitute a representation of auditory nerve firing rates [13] and are used in ASC algorithms [6]. The standard *delta* metrics [11] were also computed based on the *ILD*, *ITD*, and *IC* cues. The binaural cues were estimated using the publically available software package developed as an auditory front-end of the TWO!EARS system [14].

The following spectral features were included in the study: *centroid*, *spread*, *brightness*, *high-frequency content*, *crest*, *decrease*, *entropy*, *flatness*, *irregularity*, *kurtosis*, *skewness*, *roll-off*, *flux*, and *variation*. They all constitute the standard metrics commonly used in music information retrieval algorithms [15]. The above spectral features were extracted separately from the x and y signals. Then, the differences between the obtained spectral descriptors (difference features) were computed for each time frame. In addition, the same procedure was also applied to the m and s signals.

Mel-frequency cepstral coefficients (MFCCs) are commonly used in the ASC algorithms as spectral descriptors [4]. In our study, the first 20 coefficients were extracted for the m and s signals, respectively, and summarized using means and standard deviations. The similar calculations were also performed for the *delta*-MFCC coefficients. Moreover, the same procedure was also applied to the difference values between the MFCC coefficients obtained for the m and s signals, respectively.

The last group of features included in this study was derived from the amplitude modulation spectrograms (AMSs) [16]. First, the AMSs were calculated for the m and s signals, respectively. Then, the modulation spectrograms were transformed using the discrete cosine transform (DCT). As a result, for each time frame 600 DCT coefficients were produced. In order to compress the data, only the first 40 coefficients were preserved (the value adjusted during the pilot experiments). Finally, the DCT coefficients were summarized across time frames using the mean values and standard deviations.

V. Experiments and Results

The following five algorithms were selected and compared in terms of their ability to classify the spatial scenes: (1) *k*-nearest neighbors algorithm (*k-nn*), (2) multinomial re-

TABLE II.
OVERVIEW OF THE EXTRACTED FEATURES (1012 METRICS IN TOTAL)

	Spatial Features		**Spectro-Temporal Features**		
Feature Acronym	RMS	Binaural Cues	Spectral Features	MFCC	DCT AMS
No. of Features	8	492	112	240	160

gression with a least absolute shrinkage and selection operator (*lasso*) [17], (3) random forest, (4) neural network, and (5) support vector machine (*svm*).

The training data consisted of 451 observations and 1012 variables (features). A standard 10-fold cross-validation was performed during the supervised training procedure.

Fig. 1. shows the average classification accuracy results obtained using a single classification algorithm, namely the random forest. The classifier employing a subset of only 8 features based on the rms estimators produced the worst results, with the mean accuracy below 60%. This outcome shows that such simplistic metrics are inadequate, on their own (that is used in isolation from the other features), to reliably discriminate between the audio scenes. Far better results could be obtained by using a set of 492 features based on the binaural cues, with an accuracy reaching approximately 70%. Spectral features (112 metrics), when used on their own, yielded a similar level of accuracy. Slightly better accuracy could be obtained employing solely the MFCC features (240 metrics). DCT-AMS features (160 metrics) used in isolation from the other descriptors produced slightly disappointing results with the accuracy level of approximately 65%. The best classification outcome was obtained by incorporating all the features simultaneously (1012 metrics), yielding a mean classification accuracy of approximately 78%.

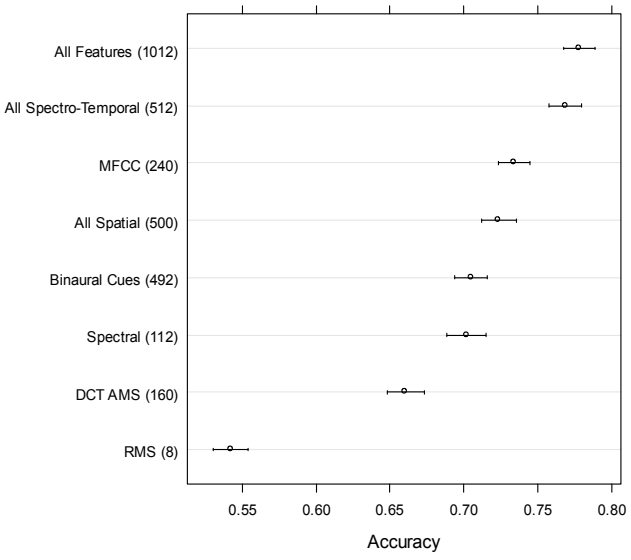


Fig. 1 Classification accuracy obtained using lasso regression for selected groups of features. The results show means and associated 95% confidence intervals. Numbers in brackets denote a quantity of features in each group.

Note that a conglomerate of all the 500 spatial features produced markedly worse results compared to those obtained using the combined group of all the 512 spectro-temporal features. This surprising outcome showed that the spectro-temporal features might be better at discriminating between the spatial scenes than the spatial metrics. This observation was confirmed during the validation test described below.

In order to reduce the risk of overfitting, a backward stepwise selection technique [17] was applied to the test data. An overview of the obtained results, including the accuracy levels, the number of retained features and the values of the model parameters, were presented in Table III. The obtained results show that the best models obtained for the lasso regression method, random forest, and support vector machines produced very similar results, with the accuracy level being equal to approximately 79.8%. The main difference between these models was the number of the selected features. For the lasso regression method, 116 features were selected, whereas for the random forest only 33 metrics were retained. The best model obtained for the support vector machine was based on 490 selected features. The worst outcomes were produced by the neural network and *k*-nn algorithms. The best models selected for each classifier during the feature selection procedure were subsequently used in a validation test.

During the validation test, based on the test dataset, the best classification accuracy results were obtained using the support vector machine (83.89%), followed by the random forest (77.18%), and the lasso regression method (76.51%). The neural network and the method based on the *k*-nearest neighbors produced the worse accuracy results, at the level of 75.17%. The confusion matrix obtained for the support vector machine (the winning method) was presented in Fig. 2. It can be seen that the algorithm could make a particularly good distinction between the *BB* scene and the remaining three scenes (sensitivity of 90.7%).

TABLE III.
OVERVIEW OF THE BEST MODELS OBTAINED THROUGH THE PROCEDURE OF FEATURE SELECTION

Classifier	**Accuracy (%)**	**No. of Features**	**Parameters**
k-nn	73.17	445	$k = 7$
lasso regression	79.84	116	*Alpha* = 0.55 *Lambda* = 6.460145×10^{-3}
random forest	79.81	33	*No. of trees* = 500 *mtry* = 17
neural network	77.64	394	*No. of hidden layers* = 1 *No. of hidden units* = 3 *Weight decay* = 0.1
svm	79.83	490	*Kernel – radial basis function (RBF)* *Sigma* = 1.822721×10^{-3} *Cost* = 1

VI. DISCUSSION AND CONCLUSIONS

The aim of this study was to identify the features useful for discrimination of the four basic spatial audio scenes of binaural recordings, labeled as *FB*, *BF*, *FF*, and *BB* (see Table I). The obtained results showed that spatial audio scenes could be classified using a mixture of spatial and spectro-

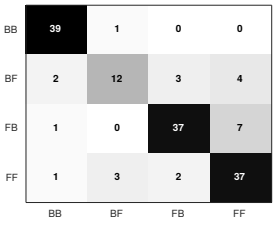


Fig. 2 Confusion matrix for the best classification algorithm (SVM, accuracy 83.89%, 490 features)

temporal metrics with an accuracy exceeding 80%. This outcome indicates that the standard spectro-temporal descriptors combined with the fundamental binaural cues (*ITD*, *ILD*, and *IC*) are adequate for the aforementioned task. Moreover, it provides evidence that the task of spatial audio scene classification may be successfully undertaken without employing a blind source separation algorithm or any other sophisticated techniques aiming to isolate and/or localize audio sources in complex binaural audio scenes. Such an approach could simplify the design of spatial audio scene classifiers.

It was surprising to observe that the spectro-temporal features appeared to have a stronger influence on the classification results than the spatial metrics. This effect, which requires further investigation, could have been caused by an unintended correlation between the spectral and spatial characteristics of the audio recordings used in this study.

Out of the five machine-learning algorithms compared in this study, the support vector machine exhibited the best classification performance, reaching an accuracy of 83.89% upon the validation test. While this result can be considered as satisfactory at this stage of research, there is still scope for improvements. In order to enhance the proposed method, a model accounting for a well-known binaural precedence effect [18] could be incorporated in future studies.

REFERENCES

[1] J. Blauert, *The Technology of Binaural Listening*. Springer, New York, 2013, ch. 1. https://doi.org/10.1007/978-3-642-37762-4

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M.D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag*., vol. 32, pp. 16-34, 2015. https://doi.org/10.1109/msp.2014.2326181

[3] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379-393, 2018. https://doi.org/10.1109/taslp.2017.2778423

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia,* vol. 17, no. 10, pp. 1733–1746, Oct. 2015. https://doi.org/10.1109/tmm.2015.2428998

[5] S. Chu, S. Narayanan, C.C.J. Kuo, and M. J. Matarić, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. of IEEE International Conference on Multimedia and Expo, IEEE, Toronto*, Canada, July, 2006. https://doi.org/10.1109/icme.2006.262661

[6] I. Trowitzsch, J. Mohr, Y. Kashef, and K. Obermayer, "Robust Detection of Environmental Sounds in Binaural Auditory Scenes," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1344-1356, 2017. https://doi.org/10.1109/taslp.2017.2690573

[7] Y. Han and J. Park, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," *Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, November, 2017.

[8] Z. Weiping, Y. Jiantao, X. Xiaotao, L.Xiangtao and P. Shaohu, "Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion," *Workshop on Detection and Classification of Acoustic Scenes and Events*, Munich, Germany, November, 2017.

[9] S.K. Zieliński, "Feature extraction of surround sound recordings for acoustic scene classification," In: Rutkowski L., Scherer R., Korytkowski M., Pedrycz W., Tadeusiewicz R., Zurada J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2018. Lecture Notes in Computer Science, vol. 10842. Springer. https://doi.org/10.1007/978-3-319-91262-2_43

[10] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651-666, 2002.

[11] L. Rabiner, B.-H. Juang, B. Yegnanarayana, *Fundamentals of Speech Recognition*, Pearson Education, 2008.

[12] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*. The MIT Press, London, 1996, ch. 3.

[13] G.J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[14] A. Raake *et al.*, "Two!ears—Integral interactive model of auditory perception and experience," *Proc. DAGA*, 2014.

[15] G. Peeters, B.Giordano, P. Susini, N. Misdariis, and S. McAdams, The Timbre Toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am*., vol. 130, no. 5, pp. 2902–2916, 2011. https://doi.org/10.1121/1.3642604

[16] T. May, and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *J. Acoust. Soc. Am*., vol. 136, no. 6, pp. 3350–3359, 2014. https://doi.org/10.1121/1.4901711

[17] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, London, 2017, ch. 6.

[18] A.D. Brown, G.C. Stecker, and D.J. Tollin, "The Precedence Effect in Sound Localization," *J. Assoc. Res. Otolaryngol*., vol. 16, no. 1, pp. 1–28, 2015. https://doi.org/10.1007/s10162-014-0496-2

International Conference on Innovative Network Systems and Applications

MODERN network systems encompass a wide range of solutions and technologies, including wireless and wired networks, network systems, services and applications. This results in numerous active research areas oriented towards various technical, scientific and social aspects of network systems and applications. The primary objective of Innovative Network Systems and Applications (iNetSApp) conference is to group network-related events and promote synergy between different fields of network-related research. To stimulate the cooperation between commercial research community and academia, the conference is co-organised by Research and Development Centre Orange Labs Poland and leading universities from Poland, Slovak Republic and United Arab Emirates.

The conference continues the experience of Frontiers in Network Applications and Network Systems (FINANS), International Conference on Wireless Sensor Networks (WSN), and International Symposium on Web Services (WSS). As in the previous years, not only research papers, but also papers summarising the development of innovative network systems and applications are welcome.

- CAP-NGNCS'18—1st International Workshop on Communications Architectures and Protocols for the New Generation of Networks and Computing Systems
- INSERT'18 - 2nd International Conference on Security, Privacy, and Trust
- IoT-ECAW'18—2nd Workshop on Internet of Things—Enablers, Challenges and Applications
- WSN'18 - 7th International Conference on Wireless Sensor Networks

AREA SUPERVISORY COMMITTEE

- Awad, Ali Ismail, INSERT'18
- Furtak, Janusz, IoT-ECAW'18
- Hamrioui, Sofiane, CAP-NGNCS'18
- Ševčík, Peter, WSN'18

2nd International Conference on Security, Privacy, and Trust

ADMITTEDLY, information security works as a backbone for protecting both user data and electronic transactions. Protecting communications and data infrastructures of an increasingly inter-connected world have become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The 2nd International Conference on Security, Privacy, and Trust (INSERT'18) focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The conference is an umbrella for all information security technical aspects, user privacy techniques, and trust. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. INSERT'18 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the INSERT'18 can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing elaborating integrated security solutions, e.g. integrate different risk-based management system.
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas

TOPICS

Topics of interest include but are not limited to:

- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Social engineering and human aspects in security
- Individuals identification and privacy protection methods
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management
- Tools supporting security management and development
- Trust in emerging technologies and applications
- Ethical trends in user privacy and trust
- Digital forensics and crime science
- Security knowledge management
- Privacy Enhancing Technologies
- Misuse and intrusion detection
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

EVENT CHAIRS

- **Awad, Ali Ismail,** Luleå University of Technology, Sweden
- **Bialas, Andrzej,** Institute of Innovative Technologies EMAG, Poland

PROGRAM COMMITTEE

- **Banach, Richard,** University of Manchester, United Kingdom
- **Bun, Rostyslav,** Lviv Polytechnic National University, Ukraine
- **Clarke, Nathan,** Plymouth University, United Kingdom
- **Cyra, Lukasz,** DM/OICT/RMS (UN)
- **Daszczuk, Wiktor Bohdan,** Warsaw University of Technology, Poland
- **Felkner, Anna,** Research and Academic Computer Network NASK
- **Furnell, Steven,** Plymouth University, United Kingdom
- **Furtak, Janusz,** Military University of Technology, Poland

- **Gawkowski, Piotr,** Institute of Computer Science, Warsaw University of Technology, Poland
- **Geiger, Gebhard,** Technical University of Munich, Faculty of Economics
- **Grzenda, Maciej,** Orange Labs Poland and Warsaw University of Technology, Poland
- **Hämmerli, Bernhard M.,** Hochschule für Technik+Architektur (HTA), Switzerland
- **Hasssaballah, M.,** South Valley University, Egypt
- **Kapczynski, Adrian,** Silesian University of Technology, Poland
- **Krendelev, Sergey,** Novosibirsk State University,JetBrains research, Russia
- **MD Faisal, Mohammad,** Integral University, India
- **MD Rafiqul, Islam,** School of Computing and Mathematics, Charles Sturt University
- **Misztal, Michal,** Military University of Technology, Poland
- **Pańkowska, Małgorzata,** University of Economics in Katowice, Poland
- **Rot, Artur,** Wroclaw University of Economics, Poland
- **Stokłosa, Janusz,** WSB University in Poznan, Poland
- **Suski, Zbigniew,** Military University of Technology, Poland
- **Szmit, Maciej,** University of Lodz; IBM GSDC, Poland
- **Wahid, Khan Ferdous,** Airbus, Germany
- **Yahya, Eslam,** Ohio State University, Columbus
- **Zamojski, Wojciech,** Wrocław University of Technology
- **Zieliński, Zbigniew,** Military University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 593–602
DOI: 10.15439/2018F226
ISSN 2300-5963 ACSIS, Vol. 15

Secure Cloud Computing: Risk Analysis for Secure Cloud Reference Architecture in Legal Metrology

Alexander Oppermann, Marko Esche, Florian Thiel
Physikalisch-Technische Bundesanstalt (PTB)
Department 8.5 Metrological IT
Abbestr. 2-12, 10587 Berlin, Germany
Email: {alexander.oppermann, marko.esche, florian.thiel}@ptb.de

Jean-Pierre Seifert
Technische Universität Berlin,
Security in Telecommunications,
Ernst-Reuter-Platz 7, 10587 Berlin
Email:jpseifert@sec.t-labs.tu-berlin.de

***Abstract*—In the field of Legal Metrology, a risk assessment is demanded by European directives for certain measuring instruments. In this paper, a previously published reference cloud architecture will be subjected to such an assessment to demonstrate its suitability for providing adequate software protection. A specially tailored and standardized method is used to identify essential threats and common attack vectors for the reference architecture. With the help of calculated probability score and risk factors, the fulfillment of the essential requirements of the applicable European directives are shown. Furthermore, Attack Probability Trees are applied to more complex scenarios to identify suitable countermeasures to increase the resilience level where necessary.**

I. Introduction

LEGAL METROLOGY'S raison d'être is to establish trust between all stakeholders such as customers, manufacturers and users of measuring instrument. Since none of the involved parties alone can guarantee the validity and integrity of measurements, a Notified Body, e.g. the Physikalisch-Technische Bundesanstalt (PTB) in Germany, is obligated to inspect measuring instruments. The essential requirements of the Measuring Instruments Directive (MID) [1], such as reproducibility, repeatability, durability and protection against corruption of measuring instruments and measurements, have to be fulfilled before entering the market. Enhancing public trust in measuring instruments is vital for Legal Metrology, especially in a world with new and increasingly complex technologies in use.

New technologies, like Cloud Computing enable manufacturers and users of measuring instruments to provide improved services to customers that are more flexible and comfortable to, for example, access meters via mobile devices or enable improved service via intelligent data services. However, Legal Metrology faces a radical change through the transformation of well-contained measuring instruments nowadays to future distributed measuring systems. In 2016, the stated transition and security implications for Legal Metrology were described, concluding with a proposition for a Secure Cloud Reference Architecture focusing on these challenges [2]. By fulfilling the essential requirements of the MID and the applicable WELMEC (Western European Legal Metrology Cooperation) guide 7.2 [3] a level of legally adequate security is met. The introduced architecture further tackles threats, such as a malicious insider and data manipulation in the cloud, via fully homomorphic encryption (FHE) [4]. Moreover, exposing FHE to real-world requirements, four application scenarios were developed and applied to Smart Meter Gateway (SMGW) tariffs. These tariff applications were derived from the SMGW's technical guide of the Federal Office for Information Security (BSI) in Germany.

In this paper, a risk analysis is applied to the Secure Cloud Reference Architecture to fulfill the legal requirements (see Section II). This risk analysis is based on software risk assessment for measuring instruments under legal control proposed by WELMEC Working Group 7 [5]. By objectifying the derived probability score for identified threats while following at the same time the guidelines of ISO/IEC 27005, ISO/IEC 15408 and ISO/IEC 18045, this risk assessment method enables comparability and standardizes the otherwise highly subjective assessment process. Furthermore, potential countermeasures are identified and quantified using Attack Probability Trees (AtPT) [6] for derived assets to be suitable protected.

The remainder of this paper is structured as follows. Section II sketches the Secure Cloud Reference Architectures and describes the considered parts for this risk assessment. Section II-A explains the derived assets and applies the risk assessment method and its shortcomings. In order to introduce the AtPT to tackle the further assets in Section III, Section IV gives an overview of the results, conclusions and further work.

II. Secure Cloud Reference Architecture

The distributed measuring instrument and its reference architecture are described in [2] and a summarized overview of its modules can be seen in Figure 1 and 2. The architecture uses virtualization techniques, in order to separate software modules subject to legal control from legally non-relevant ones. The purpose of a reference architecture is to provide a generic software framework which manufacturers can adopt in their products to provide adequate software protection in line with MID requirements.

This approach benefits not only from decreased idle times and an improved cost-efficiency ratio for employed servers, but also facilitates software update processes for manufacturers in the legally non-relevant software part. This improved update

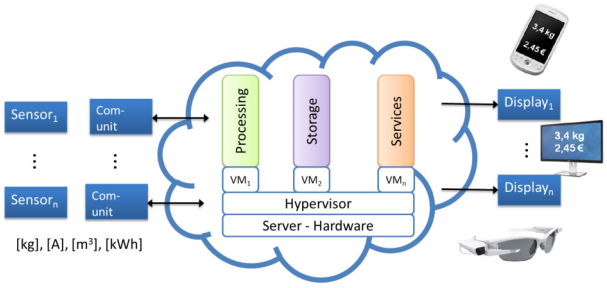


Fig. 1. Overview of the distributed measuring instrument. The measurement device is reduced to only a sensor and communication unit, while processing and storage will be moved to the cloud environment.

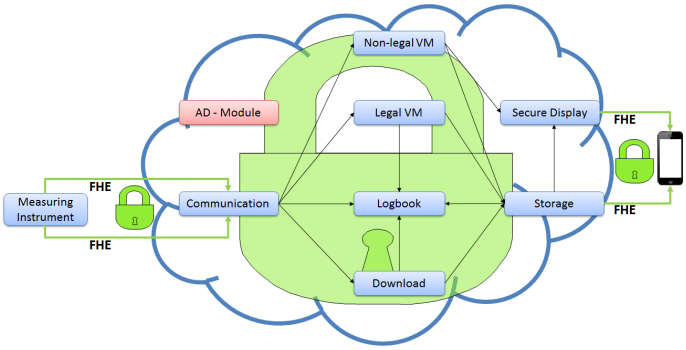


Fig. 2. Overview of the cloud reference architecture. The measurement data is completely secured throughout the whole process via FHE. The legally relevant processes are separated into different virtual machines with well defined communication paths.

process avoids a re-verification of the instrument by the market and user surveillance body and thus decreases downtimes and costs for the manufacturer.

The basis of the Infrastructure as a Service layer (IaaS) is built with the help of the Openstack framework. Core functionalities are mapped to physical devices, such as server, storage and network. Through separation into logical smaller entities via subnetworks, the network and thus the IaaS layer constitute the first low level separation between legally relevant and non-relevant processes.

The Platform as a Service layer (PaaS) consists of a microservice pattern build with Spring Boot and Spring Cloud as well as the Netflix software stack. By reducing services to their core functionality and at the same time minimizing the software lines of code (SLOC), the microservice pattern enables to maintain a clean code base. Furthermore, it offers flexible scaling and efficient resource pooling by cutting idle times of the underlying hardware. Deploying and developing services independently of each other fosters productivity within the software development team and encourages creativity. Nevertheless, stability and downtime will not be a threat to the architecture because of a rigorous separation. A stepwise transition of software versions is encouraged by running different releases side-by-side. The high level separation allows each microservice to be written in the best problem-fitting programming language.

The communication of messages is realized via RESTful API. An active message queue (ActiveMQ) guarantees reliability and pseudo resilience for messages. Messages can be stored in a queue and will be delivered later in time, in case of unavailability of services.

Fully homomorphic encryption (FHE) enables computation of encrypted messages without decrypting them first [7]. The smart meter gateway tariffs application are protected by FHE [4] and hosted at the Software as a Service layer (SaaS). Measurements are encrypted directly in the sensor unit to be processed securely in a centralized cloud structure.

In the next paragraphs, a brief description is given of the most significant legally relevant processes and virtual machines (VM). A summarized overview of the topology is illustrated in Figure 2. Increasing the portability, distributivity and scalability by separating the services via VM another security layer is introduced. The described services represent the common ground of all fields in Legal Metrology to fit a generic reference architecture.

a) Logbook: All relevant activities around the measuring system, i.e. arrival, processing, saving of measurement data, user activities, software updates etc are logged via the logbook service hosted in the Logbook VM.

b) Legal Processing: The Legal VM uses the most of all CPU cores available, because it is responsible for processing encrypted measurement data.

c) Download Manager: After an integrity and authenticity verification of the signed software update, the Download Manager will forward the software update, as intended from the manufacturer, to the dedicated machine.

d) Storage Manager: A database stores measurement data for a long period of time. The Storage Manager will make measurement data available via an REST-interface to other authorized services.

e) Monitoring Service: Detecting anomalies within the system, via continuously monitoring the behaviour of all VMs, is an important part of the security mechanisms of the cloud reference architecture. The Monitoring Service provides APIs for real time monitoring.

A. Derivation of Assets to be protected

Esche et al. [5] developed a risk assessment method based on ISO/IEC 27005 [8] and WELEMEC Risk Assessment Guide [9]. The approach consists of three stages and is shortly summarized in the following paragraphs. This algorithm is here applied to the secure cloud reference architecture (see Section III).

Every measuring instrument that undergoes conformity assessment has to fulfill the essential software requirements listed in Annex I of the MID before being put on the market. From these requirements three relevant assets are selected here that are noteworthy to be protected for all kind of measuring instruments, i.e. *measurement data*, *software that is critical for measurement characteristics* , and *metrologically relevant parameters* stored or transmitted. For each, the MID requires integrity and authenticity protection. Consequently, these assets

TABLE I
FORMAL DEFINITION OF THREATS

ID	Threat Intention	Description
B1	Integrity of transmitted measurement data	An attacker alters measurement data during transmission.
B2	Authenticity of transmitted measurement data	An attacker creates tampered measurement data, that will be assigned wrongly to a verified measuring instrument.
B3	Evidence of an intervention	An attacker prevents legally relevant events from being registered in the logbook.
B4	Integrity of Parameters	An attacker alters persistence saved parameters, e.g. connection parameters.
B5	Availability of the Logbook Service	An attacker prevents a legally relevant service from answering requests.

must be secured against intentional or unintentional changes. By fulfilling this demand, integrity and authenticity of these assets are guaranteed. In addition, the MID requires evidence of an intervention, i.e. events registered in a logbook, to be available during verification.

a) Threat definition: A threat is any invalidation of a security property of a given asset. To define a threat, aside from the asset definition, several attacker models should be taken into account, for example, inside attacker and external attacker. Usually the market participant with the highest skill level can be used as a reference model. Additionally, different access levels and their associated roles within a measuring instrument take an important part in the risk assessment. In Table I these five assets are linked to a threat intention and short description of what an attacker wants to achieve. The assets itself will be further described in separate tables, where the attack vectors (technical steps needed to implement a threat) are broken down into atomic attacks with a time, expertise, knowledge, window of opportunity and equipment column that are individually scored (see Section III), according to [10]. This procedure has the advantage of objectifying the risk assessment procedure based on scores for well-defined features of any attack. This enables manufacturers and Notified Bodies alike, to be able to compare the same threats for different measuring instruments.

b) Identification of Attack Vectors: The second risk assessment phase is the least formalized stage. It starts with the examination of the manufacturer's documentation of the measuring instrument. Followed by creating a collection of possible attack vectors, needed to realize the prior identified threats from stage one. The collection comprises attack vectors reaching from simple to very complex structured attacks.

c) Calculating Probability Score and Risk Score: In phase three, the interim results from stage one and two are combined, i.e. an adverse action with at least one associated attack vector. Thereafter, the likelihood of implementing such an attack has to be calculated. The evaluation is based on the following five features [11] that lay the foundation to score and identify the resources that all attacks have in common:

- Elapsed Time (0-19 points)
- Expertise (0-8 points)
- Knowledge of the TOE (0-11 points)
- Window of Opportunity (0-10 points)
- Equipment (0-9 points)

The amount of *elapsed time* represents the time needed to implement a specific attack by any chosen attacker. The score ranges from 0 (equals 1 day) to 19 (more than half a year). *Expertise* represents the skill set of an attacker, where 0 is a layman and 8 is given when an attacker has to have competence in more than one field. *Knowledge of the Target of Evaluation (TOE)* scores the needed information on an attacked measuring instrument. It starts with publicly available knowledge (0) and ends with critical insider knowledge (11), that usually resides with the manufacturer. The *window of opportunity* evaluates the possibility available to an attacker, where 0 represents unlimited access, which would be common for measuring instruments connected to the Internet. If the access is difficult, a value of 10 should be given. In case it is impossible to obtain access, no rating is done and the attack vector would be removed from the list. The last category scores the *equipment* needed to carry out the attack. Standard available hardware or software is described by 0, where 9 represents multiple bespoke devices or software.

After successfully calculating the sum yielded by the five categories for the chosen attack, a probability score is matched to the different ranges of the total sum. In Table II the Common Criteria evaluation is also included in the final probability score calculation, so that a basic resistance results in a total sum of 10-13 points while 24 or more points represent a high resilience against the rated attack. Finally, the resistance evaluation is associated with the probability score, where 1 represents an unlikely occurrence while 5 stands for high probability to occur.

The final risk will be calculated by multiplying the impact score for the threat with the probability score, that is issued in Table II, of the most likely realized attack vector:

$$\text{risk score} = \frac{\text{impact score}}{5} \cdot \text{probability score} \qquad (1)$$

TABLE II
CALCULATION OF A TOE AND ASSOCIATION OF A PROBABILITY SCORE ACCORDING TO [5]

Sum of Points	TOE Resistance	Probability Score
0-9	No rating	5
10-13	Basic	4
14-19	Enhanced Basic	3
20-24	Moderate	2
>24	High	1

TABLE III
ATTACK VECTORS FOR THREAT B1

Attack-ID	Attack Vector	Time	Expertise	Knowledge	Window of Opportunity	Equipment	Sum	Damage
A3	Manipulate data in transit	19	8	11	10	0	48	1
A4	Exchange processing unit	7	6	11	4	0	29	1

TABLE IV
PREREQUISITES FOR ATTACK VECTOR A3

Attack-ID	Attack Vector	Time	Expertise	Knowledge	Window of Opportunity	Equipment	Sum	Damage
A3.1	MITM-attack	1	6	11	10*	0	28	1
A3.2	decrypt-encrypt data	19	8	11	0	0	38	1

III. METHODOLOGY OF ASSETS

In this section, the risk assessment algorithm will be applied to the secure cloud reference architecture, that were both briefly introduced in the previous section. The threats listed in Table I will be treated sequentially and will pass the three stages of risk assessment. Afterwards, in Subsection III-C the Attack Probability Tree (AtPT) is introduced to describe more complex attack scenarios, by introducing a prescribed way to construct attack vectors in a standardized and compact way. At the end, suitable countermeasures for attack vectors will be discussed briefly.

A. Integrity of transmitted measurement data

The threat intention of the attacker is to undermine the integrity of transmitted measurement data by manipulating measurement data during transmission. The sensor unit will be considered, that collects the data and encrypts them with a protected public key via FHE before sending them to the cloud reference architecture. The transmission is secured by Transport Layer Security (TLS) and additionally by a x.509 certificate at the cloud service endpoint, so that the sensor unit usually knows the receiver. An insider attack is assumed with the attacker having the access rights of an administrator. For this threat, two attack vectors are taken into consideration, namely A3 and A4 (see Table III). A3 needs two prerequisites A3.1 and A3.2 (see Table IV), in order to be feasible.

To manipulate the data in transit, the attacker has to carry out an active Man-In-The-Middle attack (MITM) (see Table IV A3.1), that means the connection has to be rerouted via the attacker's interception device and the TLS-connection has to be captured during key exchange. Furthermore, the certificate has to be forged by, for example, getting the private key of the server and the client to establish active sessions at both ends with the impersonated certificates needed for authentication. The client's improper validation of the certificate would be a big advantage for the attacker.

The time needed to execute such an attack would be less than a day (1), if the attacker is an expert (6) and has critical knowledge of the system (11). While the window of opportunity is difficult (10), since the manipulation has to be carried out during transmission within the boundaries of transmission delay. There is no special equipment needed (1), that exceeds standard hardware. So the total sum of points for this attack (48) leads to high TOE resistance (see Table II).

Even if A3.1 (MITM) is successfully established, the data itself is still encrypted by FHE. Lattice based cryptography is provable secure and provides worst-case security that is still not broken by quantum algorithms. Therefore, the maximum time of more than half a year (19) assumed for A3.2. The attacker has to have expertise on several fields (8) to decrypt and/or break cryptography as well as having critical system knowledge (11) at disposal. Once, the cryptography is broken, the window of opportunity is unnecessary (0). From the authors' point of view standard hardware (0) is sufficient. This yields a total sum of 38 points and again implies high resilience against the attack vector.

The two attack vectors A3.1 and A3.2 both need to be executed to form A3. The result is shown in Table III and implies a high resilience (48) for this attack vector. According to Table II, the sum score translates to a probability score of 1. Since this threat has potential influence on all future measurement values, the impact score is 5 and the subsequent risk ($\frac{\text{impact score}}{5} \cdot \text{probability}$) also takes on a value of 1. PTB does not accept technical solutions with a risk greater than 3. This solution qualifies for PTB certification.

Another attack vector is to exchange the FHE-processing unit (A4) in the cloud, in order to manipulate the data during processing. First, the attacker needs to have access to the software repository, to manipulate the FHE-processing unit and then deploy the manipulated software into the cloud service. Furthermore, the hash of the manipulated software has to match the comparative hash, that the market surveillance monitor evaluates. Given the bonus of an insider attacker with the access level of an administrator, it should be feasible, yet the time frame for execution is less than two months (7). The attacker needs to be at least an expert (6) in IT and the window of opportunity is moderate (4), since a lot of security mechanisms have to be worked around. No special hardware (0) is needed. This yields a total sum of 29 and means a high TOE resistance and a probability score of 1. The threat influences all future measurements, the impact score is 5 and the resulting risk has a value of 1.

TABLE V
ATTACK VECTORS FOR THREAT B2

Attack-ID	Attack Vector	Time	Expertise	Knowledge	Window of Opportunity	Equipment	Sum	Damage
A1	Manipulate sensor unit	4	8	11	0	7	30	1
A2	Replace sensor unit	4	8	11	0	7	30	1
A3	Spoof identity	19	6	11	0	0	36	1

TABLE VI
PREREQUISITES FOR ATTACK VECTOR A3

Attack-ID	Attack Vector	Time	Expertise	Knowledge	Window of Opportunity	Equipment	Sum	Damage
A3.1	Steal key from vault	1	6	11	0	0	18	1
A3.2	Obtain certificate	19	6	0	0	0	25	1
A3.3	Generate false data	19	6	11	0	0	36	1

B. Authenticity of transmitted measurement data

The threat intention of B2 is to attack the authenticity of transmitted measurement data. In Table V three attack vectors A1-A3 are summed up, while the third is composed of three sub attack vectors displayed in Table VI.

The easiest way of attacking the authenticity is to manipulate the origin of the measurement data: the sensor unit itself (A1). The idea behind this attack vector is just to compromise the authenticity, thus it is enough to break the seal and replace the physical sensor with a tampered one, that calculates, for example, a smaller measurement value. Breaking the seal implicates forging a new seal, so that the instrument does not seem to be manipulated to market surveillance.

The time needed for this invalidation of authenticity (A1) is less than a month (4) and the attacker needs to be expert on several fields (8), since forging an official calibration seal needs knowledge and special equipment (7). Furthermore, replacing the physical sensor requires critical knowledge (8). The window of opportunity is unlimited (0) for this attack vector, because the instrument in the field is not subject to constant surveillance. In total, the attack vector reaches 30 points and represents a TOE with high resistance with an associated probability score of 1, which translate to a risk level of 1 because of its influence of all future measurement values (impact score of 5). However, it is noteworthy that in Legal Metrology there is no higher protection level achievable than a sealed hardware solution.

The second attack vector A2 deals with obtaining security features from the original sensor unit (physical sensor + communication unit) and replacing this unit with a tampered one that is identically constructed. Hereby, the attacker extracts, for example, the protected key (public key) needed for encryption from the original sealed instrument and then stores this security feature in an identical, but tampered unit. A2 differs from A1 since it does not involve tampering original hardware, but buying malfunctioning hardware on purpose and putting it into use. The scores are the same as for the previous attack vector. It is again considered very hard to forge an official verification seal, which is reflected in the total sum of 30 points and offers high resilience.

With the last attack vector A3 the identity of the sensor unit will be spoofed by masquerading the IP address of the attacker's sensor unit, for example, by faking the source address field in the TCP header. In order to be successful at the cloud service endpoint, the attacker has to first obtain the protected key from the software vault in the cloud service, in order to be able to encrypt its fake measurement data (A3.1). Given the fact that an insider attacker with the privileges of an administrator is considered, the access to the cloud architecture is self-evident. The attacker will retrieve the information in less than a week (1). The postulated skill set of an expert (6) is needed in an IT related area and critical knowledge (11) of the system is demanded. A3.1 yields in total 18 points, which is considered as an enhanced basic resistance level.

As a next step (A3.2), the attacker has to get his hands on the private key of the x.509 certificate. It is assumed that this is very time consuming (>6 month) (19) but feasible for an expert (6), in order to forge the x.509 certificate and overcome the authentication barrier. The attack vector A3.2 has a total sum of 25 points and achieves high resilience against this threat.

As a last action, the attacker has to generate false measurement data with the stolen key from A3.1 and authenticates himself against the cloud service endpoint with a forged certificate, in order to achieve the objective to compromise the authenticity of the measurement data. Because of the logical AND operation of A3.1 and A3.2 the highest value will run into A3. That leads to the time frame of more than 6 months (19), an expert level (6) and the requirement of critical system knowledge (11), which totals into 36 points and reaches a high resistance level. The probability score evaluates to 1 with an associated risk level of 1 because of the influence of all future measurements (impact score 5).

For threat intention B3 the same risk assessment procedure is carried out and noted in tables. Yet, this methodology is limited and it quickly becomes extremely difficult to map all requirements and dependencies for all possible attack vectors. As a solution, Esche et al. introduced the attack probability tree that visualizes in a very compact manner the attack vectors and make it easy to deduce a probable attack path. Furthermore, it enables to derive the attacker motivation. In the next section a short theoretical introduction of the AtPT will be given and

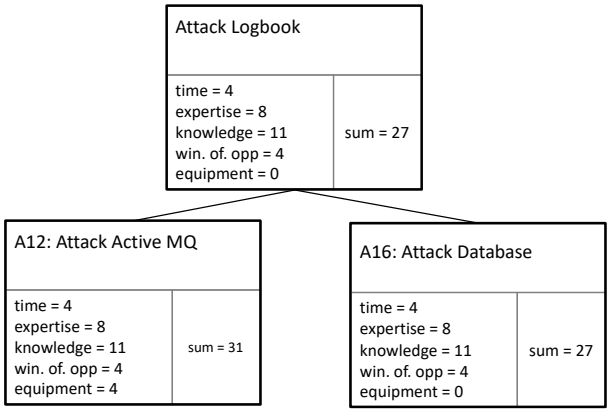


Fig. 3. AtPT for threat intention B3. View: root node and two attack vectors.

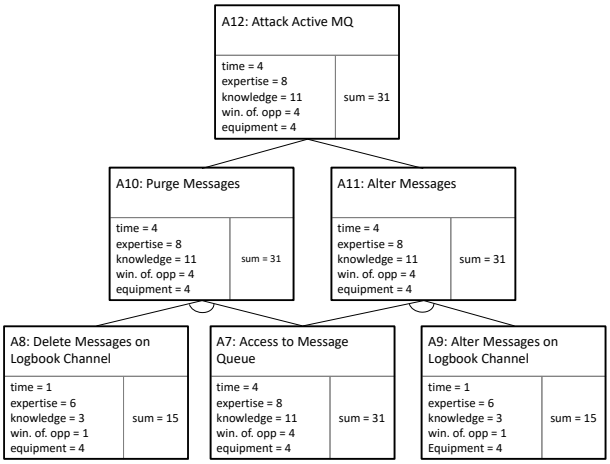


Fig. 4. AtPT for threat intention B3. View: Subtree of attack vector Active Message Queue.

subsequently applied to B3 until B5.

C. Attack Probability Tree

Esche et al. introduced attack probability trees (AtPT) as an extension of attack trees by Mauw and Oostdijk [12] to tackle two main objectives: developing a method to standardize the deduction of attack vectors and to efficiently visualize the interdependencies of attack vectors in order to easily derive attacker motivation and as a result the most likely attacker path. [6]. Additionally, each node embodies features with its own score, such as time, expertise, knowledge, window of opportunity and equipment, that have been previously collected in tables. Furthermore, the logical relationship between parent and child attacks are visualized and attack nodes are linked either by an AND- or OR-statement.

Information enter the tree via the leaves, so that parent nodes' and finally the root's attributes can be calculated from the bottom to the top. The rules for both statements and each attribute/point score are extensively described in [6]. Briefly summarized: for AND-statements, the *maximum* for each attribute chosen; for OR-statements, the *smaller* sum score indicates the threat to select. A great side-effect of AtPTs is the reduction of required time for revaluation of individual attacks, because of the possibility of reusing attack nodes, that are common among different attacks without recalculating attributes.

The following subsections use the AtPT approach for risk assessment of the cloud reference architecture. Nevertheless, the corresponding tables were generated, as introduced in the previous sections. However, due to space constrains, they are not published here.

D. Evidence of an Intervention

In this scenario an attacker prevents legally relevant events from being registered in the logbook. The threat intention is to attack the availability of the evidence of an intervention. In case of a successful manipulation, the user cannot present all relevant logbook entries that market surveillance demands.

In this paper, only the AtPT for a logbook attack is presented. Another attack scenario with the same attack attributes is evaluated for the storage service of the instrument, with a similar-looking AtPT. Because of the complexity of the attack, the AtPT is divided into four subtrees (see Figures 3-6), that will be described in the next paragraphs.

An AtPT is read from the root to the leaves. For attacking the logbook, two possibilities are available. Either the attacker aims for the active message queue (Active MQ) or for the database of the logbook service (see Figure 3). Since these two attack vectors are alternatives, they are linked by an OR-connection. If the two vectors would be needed to be executed together, they would be linked by an AND-connection graphically expressed by an arc.

When attacking the Active MQ (A12), an attacker could either purge messages (A10) or alter message (A11) on the logbook channel. For both actions, access to the message queue is required (A7) with the combination of deleting a message (A8) or changing a message (A9) on the logbook channel represented by an arc below the linked nodes (see Figure 4).

The actual scores in Figure 4 are calculated from the bottom to the top, for example, attack vector A10 consists of nodes A8 and A7. Since the latter two nodes are linked by an AND-statement the greater value is put across to A10. The time to purge a message takes less than a month (4) and stems from A7 accessing the message queue. Furthermore, it is required to be an expert in several areas (8), to have critical knowledge of the system (8) and the window opportunity is moderate (4). These attributes stem also from A7. However, the equipment to purge messages on the active MQ is specialized (4), since the software is an expert tool written in python without a graphical user interface. Yet it is indeed publicly available.

Now one could argue, that using a specialized software and obtaining access to the message queue needs less time than proposed here. However, the whole AtPT does not end with obtaining access to the message queue (A7), but rather continuous and becomes more detailed in how the access could be obtained in a malicious way.

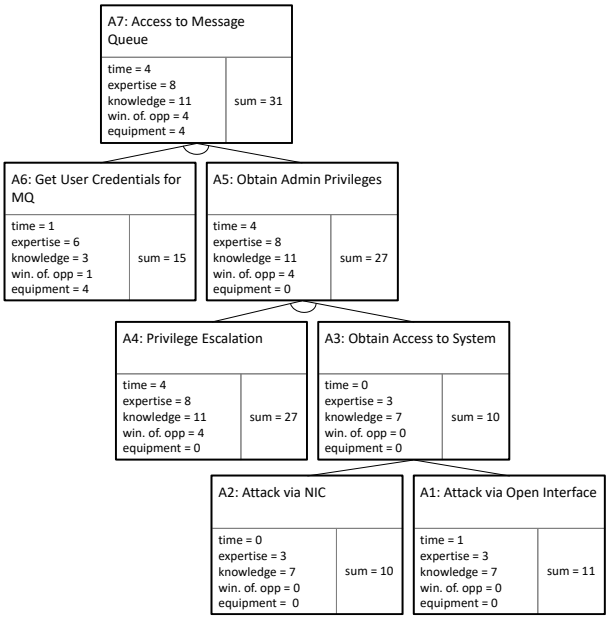


Fig. 5. AtPT for threat intention B3. View: Subtree of attack vector Access to Message Queue.

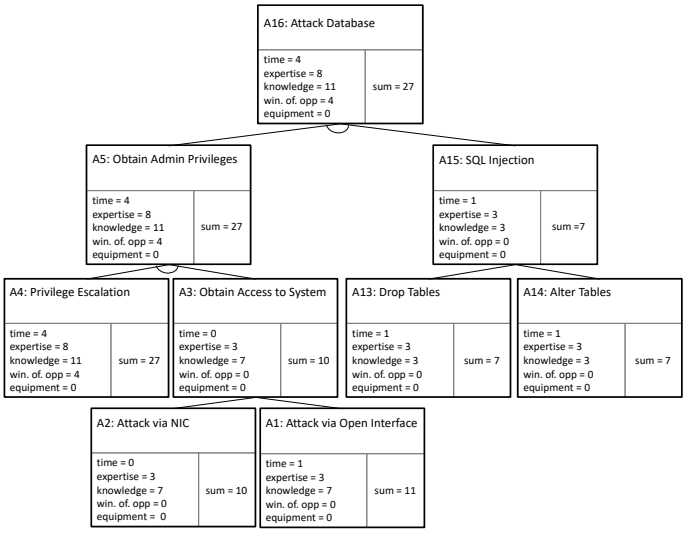


Fig. 6. AtPT for threat intention B3. View: Subtree of attack vector Attack Database.

In Figure 5 an exemplary attacking path is detailed. Node A7 consists of obtaining administrator privileges in the virtual machine (A5), that runs the active MQ or is at least in the same subnet. With these new privileges the specialized software can be executed, which triggers node A6 to get the credentials for the message queue.

To get hold of the user credentials, less than a week (1) is estimated. An expert level (6) and restricted knowledge of the measuring system is required. The window of opportunity for an inside attacker is easy (1) even so specialized software (4) is needed. Node A6 holds a total sum of 15 points which would be considered as an enhanced basic resistance level. However, A6 is to be evaluated in conjunction with A5 through the AND-connection.

The attack vector A5 depends again on a privilege escalation through exploiting Common Vulnerabilities and Exposures (CVE) of the underlying system (A4) and obtaining access to the virtual machine (A3). To accomplish a privilege escalation, the attack is assessed with less than a month (4), expertise on more than one field (8), critical system knowledge and moderate window of opportunity (4). Further, no special equipment (0) is expected. A privilege escalation is considered as a difficult endeavor with 27 points in total that translate to a high resilience. This corresponds again to a probability score of 1 with an impact score of 5 and results into a risk of 1.

Obtaining access to a virtual machine and therewith to the distributed measuring system (A3) is possible in two ways that are alternatives (OR-connection). Either the system is penetrated through a network interface card (NIC) (A2) or via an open physical interface (A1), such as a USB port. Considering the fact that an inside attacker with administrator privileges is assumed, that logs remotely into the measuring system for maintenance reasons, this attack is achievable in less than a day (0). To clarify, it is assumed that the inside attacker does not automatically have administrator privileges on the remote machine, but as an employee of the manufacturer. Furthermore, to login remotely requires only a proficient expertise and sensitive system knowledge (7). The window of opportunity is negligible (0), since this can belong to the attacker's daily routine. No special equipment is needed (0). The TOE resistance is basic (10 points in total).

The attack via an open interface (A1) differs from A2 only in the time attribute. It is assumed that the attacker has to physically approach the hardware to carry out the attack. That takes additional time (less than a week (1)) and is more inconvenient than opening a SSH-shell from the desktop pc in the office.

To sum up, the attack path just described consists of A2, A3, A4, A5, A6, A7 then a decision has to be made if the messages should be altered or deleted. However, in terms of likelihood the nodes do not differ, but practically spoken deletion is often easier. The path would continue via A8, A10.

To completely describe the AtPT for compromising the evidence of an intervention via a logbook attack, the alternative path via the database attack vector (A16) has to be described, as shown in Figure 6. For attacking the database, administrator privileges (A5) are needed combined with an attack against the database such as SQL injection (A15) or via command line interface (CLI). The path down to the leaves for A5 is already described in the previous paragraphs. Its TOE resistance depends on leaf A4, that describes the privilege escalation via a CVE. Attack Vector A15 is divided into dropping tables (A13) or modifying tables (A14).

The scores for A13, A14 are equal and subsequently A15 is identical as well. For both attacks, less than a day is assumed, only a proficient expertise level (3) is needed, no special equipment (0) is required and the window of opportunity

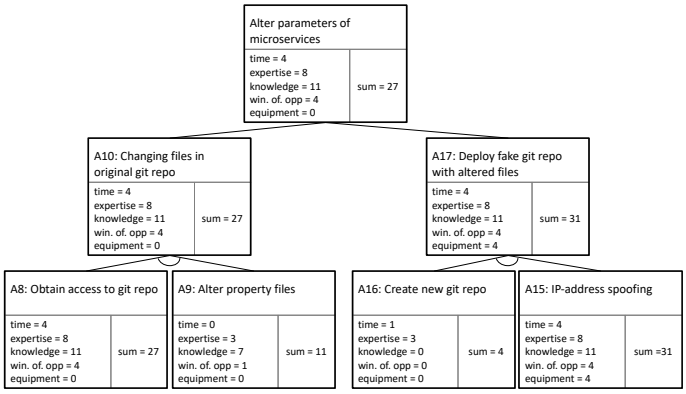


Fig. 7. AtPT for threat intention B4. View: Root, Alter parameters of microservices.

is unlimited (0). In total, the database attacks combine to 7 points, which translate to no resistance at all (no rating). However, since A5 and A15 are connected via an AND-statement the parent node A16 receives the TOE resistance high, since the attacks depend on the privilege escalation to be carried out.

The most likely attack path would be via the database, since no special software is needed, thus less time is required for learning and incorporating the software. To compromise a database, no new software has to be deployed so that the effort on the attacker side is less than attacking the message queue, especially if the intention is to just compromise the integrity of the measuring instrument.

E. Integrity of Parameters

Threat intention B4 aims for harming legally relevant software parameters to violate the security properties integrity and authenticity. In the following paragraphs the presented scenario offers an attacker to alter persistent saved parameters of the logbook service by attacking the configuration service. Two possible attack scenarios are presented via an AtPT. The tree is compartmentalized into several subtrees, because of its size (see Figures 7-9). As already pointed out, the subtree consisting of the node A1-A5 could be reused for several attack scenarios without revaluation. Due to space constraints it was renounced to map the whole subtree of A5 downwards in Figure 8. A complete subtree can be seen in Figure 5.

It is proposed that the attacker changes microservice property files in the original git repository to attack the microservice architecture (A10) and provides, for example, false message queue groups. That could lead to loss of messages in the legal relevant logbook. Aiming for the configuration basis can cause fundamental harm and chaos to the whole system.

To be able to carry out attack vector A10, it is assumed that the attacker has to obtain access to the original git repository (A8) and is able to alter the property files (A9). Nodes A8 and A9 are linked via an AND-connection to A10 (see Figure 7).

In order to obtain access to the git repository (A8), a SSH-key has to be created (A7) and placed into the specific folder for the git repository to be evaluated (A6). A7 and A6 are

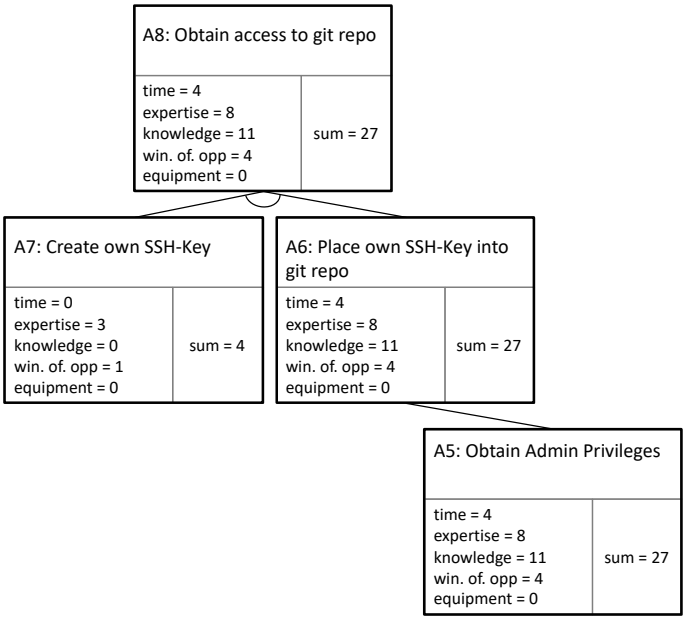


Fig. 8. AtPT for threat intention B4. View: Subtree of attack vector A8.

linked via an AND-statement to A8 (see Figure 8). Attack vector A6 is linked to subtree A5, that describes accomplishment of obtaining administrator privileges (a complete subtree is shown in Figure 5).

The score for subtree A5 has been described in the previous section, so that the evaluation starts with A6. All of the attributes stem from A5 and the difficulties to obtain administrator privileges, which are prerequisites for A6. To create an SSH-key (A7) takes less than a day (0) with proficient expertise (3). How to do this is public knowledge (0) and tutorials are easily to find on the Internet. Furthermore, assuming that an inside attacker is already in the system, the window of opportunity is easy to accomplish (1) and also no special equipment (0) is necessary. This yields a total sum of 4 points and a TOE resistance with no rating.

Attack vector A8 receives the point score from A6, since because of the AND-connection only the maximum of both attributes will be passed upwards. That leads to a total sum of 27 points for obtaining access to the original git repository and implies high resilience.

Altering property files can be done in less than a day (0), with only proficient expertise (3), an easy window of opportunity (1) and with any text editor (equipment = 0). That totals in 11 points and matches basic resilience for this attack.

A10 receives the attributes in total from A8 and thus defines the total score of 27 points and a high resilience for this attack path (see Figure 7).

The alternative attack vector for B4 is to deploy a fake git repository with already altered property files (A17). This attack vector splits into first creating and deploying a new git repository (A7) and then tricking the system into trusting and pulling the files from the fake git repository via IP spoofing (A15). The spoofing attack itself is subdivided into carrying

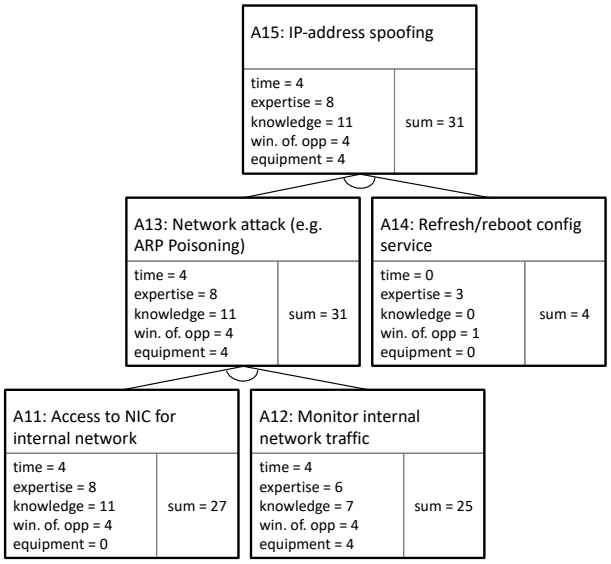


Fig. 9. AtPT for threat intention B4. View: Subtree of attack vector IP-address spoofing.

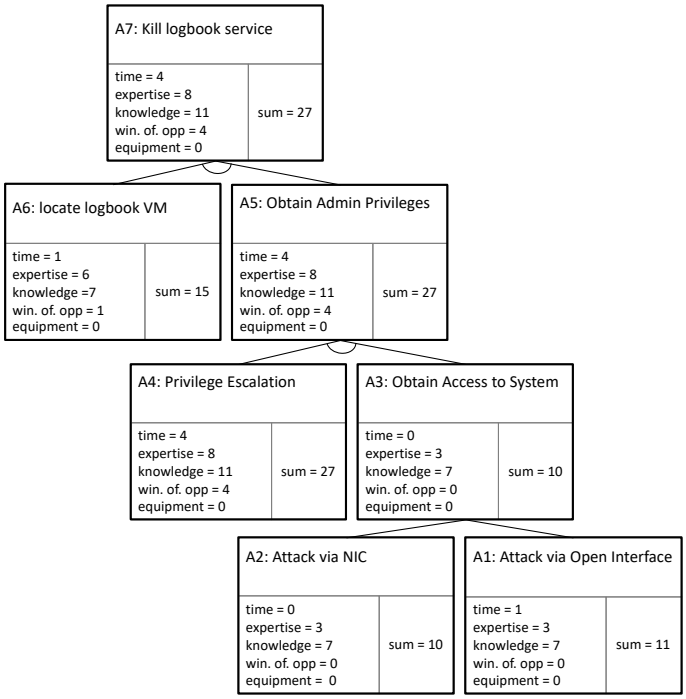


Fig. 10. AtPT for threat intention B5 to violate the availability security property.

out a network attack such as Address-Resolution-Protocol-Poisoning (ARP-poisoning) (A13), in order to replace the IP address and then rebooting the configuration service (A14). To be able to carry out a network attack, the attacker is assumed to have full access to the internal network of the distributed measuring system (A11). To monitor the internal network traffic (A12), the attack vector A11 is necessary (see Figure 9). Once again subtree A5 is required to successfully implement A11.

The score of gaining full access to the internal network interface card (NIC) and thus to the internal network (A11) is inherited from A5 and the struggle of obtaining administrator privileges. To gain a full picture of the structure of the internal network with its services (A12) takes less than a month (4) with an assumed expertise in networking (6) and a sensitive knowledge of the measuring system (7). A moderate window of opportunity (4) is predicted, because it is difficult to explore a supervised internal network undetected. Furthermore, specialized software is needed to monitor network traffic (4). This yields in total 25 points and maps to a high resistance to attacks with probability score of 1 and a risk of 1.

Carrying out network attacks, such as ARP-poisoning (A13), requires less than a month (4) for experts on several fields (8) with sensitive knowledge of the system (11), a moderate window of opportunity (4) and specialized software (4). For most network attacks, it is not necessary any more to write specialized software. There exists publicly available grey software, that can be used to detect vulnerabilities or can be misused to attack computer systems. This attack vector combines to 31 points and a high resistance factor. From here on, no significant changes to the resilience are contributed until the final attack vector A17. Minor actions are required to finally deploy a fake git repository but both acquire only 4 points in total with a negligible threat resistance (see Figure 7).

The most probable attack path will be via A10, changing the properties files in the original repository, since it is the least complex one and without the hassle of deploying software and monitoring traffic etc. This is also reflected in the total score of 27 against 31 points.

F. Availability of Service

Threat intention B5 targets the availability of a legally relevant logbook service. In Figure 10 the complete AtPT is illustrated with the already introduced subtree A5, that enables access to the measuring system and comes along with an escalation of privileges. This subtree in conjunction with the localization of the logbook service's virtual machine (A6) enables the final attack vector that kills the logbook service (A7).

The final score stems from the difficulty to gain access to the system and to elevate the privilege level (subtree A5), which totals 27 points. With an associated risk level of 1 and a probability score of 1. Locating the logbook's virtual machine is with 15 points in total an enhanced basic TOE resistance level, but has no significant influence on the final score of killing the logbook service (A7).

G. Effect of Attacker Motivation

Esche et al. described in [10] possibilities to represent attacker motivation during risk assessment. The presented AtPTs are created for a highly motivated attacker. In order to reconsider these trees with a low or medium motivated attacker, the expertise and equipment score have to be replaced

TABLE VII
MAPPING OF EXPERTISE AND MOTIVATION LEVEL ACCORDING TO [10]

Expertise	Score	Motivation	Score
Layman	0	no motivation	9
Proficient	3	low	6
Expert	6	moderate	3
Multiple Expert	8	high	0

with a higher motivation score according to Table VII if they are originally smaller. This will result in a decreased probability score for a lower motivation and vice versa for a highly motivated attacker. It is noteworthy, that the likeliest attacker path can shift, when the motivation is adjusted.

H. Suitable Countermeasures

To find the best suitable place for countermeasures in an AtPT, it is recommended to locate an inverted subtree for mitigating attack vectors and increasing the impact of applied countermeasures. An inverted tree is usually any leaf that is connected to more than one node of the previous level. Subsequently, the size of an inverted tree matters, since the greater it is, more parent nodes are impacted. In the trees for B3 and B4, A7 and A16 depend on A5 as well as A6 and A11 depend on A5. Subtree A5 is of general importance, because it describes the unauthorized access to the measuring system and privilege escalation. A countermeasure specifically tailored for A5 will exacerbate to obtain administrator rights. This node will have the biggest impact on all three threat scenarios from B3-B5.

A suitable countermeasure is to strengthen the access rights and to enforce a least privilege policy. For example, one could implement Security Enhance Linux (SELinux) for virtual machines (VM), that provides a mandatory access control system and security policies. Instead of using a standard Linux, the kernel extension SELinux provides by default a least privilege policy that denies everything except if it is specifically allowed by access policies (enforcing mode). All violations against these rules are logged and an alarm can be triggered. To obtain administrator privileges by an escalation of access rights would need significantly more time (less than 2 months (7)) with SELinux in place. Furthermore, if the attacker is able to bypass SELinux via switching form enforcing to permissive mode it needs to be done on every VM with a bespoke software (7). However, rolling out SELinux to the measuring system would mean a lot of configuration overhead, but it would elevate the security score by 10 points to 37. This security enhancement would propagate via the inverted tree to the top of each AtPT.

IV. SUMMARY

In this paper, a secure cloud reference architecture for distributed measuring instruments under legal control was presented and subjected to a especially tailored risk assessment method for software in Legal Metrology. After formally introducing the risk analysis, five threats for the reference architecture were described and evaluated extensively. The first two threats were assessed using the traditional method via tables. However, this approach seemed infeasible for more complex threats. Therefore, the Attack Probability Tree (AtPT), that eases the handling of more complex attacks, was introduced and applied. It was shown that adequate protection of the essential requirements formulated by the MID is provided by the secure cloud reference architecture. Therefore, the architecture is qualified to be implemented in measuring systems under legal control.

The detailed analysis of the threat intentions using AtPTs revealed for all formulated threats and attacked security properties a high resilience factor. Nevertheless, through the inverted subtree method for AtPTs the optimal entry point for countermeasures was identified. The implementation of countermeasures reduced the risk to the level provided by physical sealing and increases the resilience to attacks.

Future work will focus on different attacker motivation and therewith diverse attack paths. Furthermore, the formalization of creating AtPTs has to be optimized and standardized.

ACKNOWLEDGMENT

This Paper would not exist without the help and fruitful discussions with our colleague Federico Grasso Toro.

REFERENCES

[1] European Parliament and Council, "Directive 2014/32/EU of the European Parliament and of the Council," *Official Journal of the European Union*, 2014.
[2] A. Oppermann, J.-P. Seifert, and F. Thiel, "Secure cloud reference architectures for measuring instruments under legal control." in *CLOSER (1)*, 2016, pp. 289–294.
[3] "WELMEC 7.2 Software Guide," *WELMEC European cooperation in legal metrology, Welmec Secretariat, Delft, Standard*, 2015.
[4] A. Oppermann, A. Yurchenko, M. Esche, and J.-P. Seifert, "Secure cloud computing: Multithreaded fully homomorphic encryption for legal metrology," in *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Springer, 2017, pp. 35–54.
[5] M. Esche and F. Thiel, "Software risk assessment for measuring instruments in legal metrology," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015, pp. 1113–1123.
[6] M. Esche, F. G. Toro, and F. Thiel, "Representation of attacker motivation in software risk assessment using attack probability trees," *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on (pp. 763-771). IEEE.*, 2017.
[7] C. Gentry *et al.*, "Fully homomorphic encryption using ideal lattices." in *STOC*, vol. 9, 2009, pp. 169–178.
[8] ISO27005:2011(e), "Information technology - security techniques - information security risk management." *International Organisation for Standardisation, Geneva, CH*, vol. Standard, Jun. 2011.
[9] "Welmec 5.3 Risk Assessment Guide for Market Surveillance: Weigh and Measuring Instrument," *WELMEC European cooperation in legal metrology, WELMEC Secretariat, Ljubljana*, May 2011.
[10] M. Esche and F. Thiel, "Incorporating a measure for attacker motivation into software risk assessment for measuring instruments in legal metrology," *18. GMA/ITG-Fachtagung Sensoren und Messsysteme 2016,Nürnberg, Germany*, vol. 1, no. 1, pp. 735–742, Mai 2016.
[11] ISO/IEC18045:2012, "Common Methodology for Information Technology Security Evaluation," *International Organisation for Standardisation, Geneva, CH*, Sep. 2012.
[12] S. Mauw and M. Oostdijk, "Foundations of attack trees," in *International Conference on Information Security and Cryptology*. Springer, 2005, pp. 186–198.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 603–606
DOI: 10.15439/2018F97
ISSN 2300-5963 ACSIS, Vol. 15

Probabilistic Block Cipher

Nikita Zbitnev, Dmitry Shishlyannikov, Dmitry Gridin
JetBrains Research, Novosibirsk, Russia
Novosibirsk State University, Novosibirsk, Russia
Email: nikita.zbitnev.a@gmail.com, dmitry.shishlyannikov.a@gmail.com, dmitry.gridin.v@gmail.com

***Abstract*—This paper is devoted to the description of a new block cipher that will be applicable in the post-quantum era and will not need a lot of resources. The main advantages: probabilistic encryption, the cipher block chaining mode, the ability to transfer to distributed systems. All this combined with the use of PRNG, working on the Cremona transformations, has significantly increased the cryptographic strength and increased the scope of this encryption.**

I. Introduction

NIST recommendations for block encryption consist in increasing the size of the key. We believe that instead of increasing the key, it is necessary to build completely probabilistic encryption in such a way that the output file size would be unpredictable, that is, to make the output file size unpredictable, even if the key is the same. In addition, it is necessary to build a pseudo-random number generator that receives parameters depending on the time, or other parameters of the computer. To achieve this, the finite rings - adapted Cremona transformation is used. The case is that this kind of transformation in real numbers is in use for fractals constructing.

II. Algorithm description

A. Formulation of the problem

It is necessary to encrypt a block of text B, which has n symbols of b-bit size each with secret key.

B. Key

The key is used to generate matrices and to fill in the tables of frequencies in Huffman coding. The assumed key size is:

$$l = 2 * b * n \tag{1}$$

To extract all the necessary data from the key we would use a PRNG (hereinafter the Generator), which is described in more detail below.

C. Preparation phase

1) Module Choice

Let us choose a set of prime integers $M_1 < M_2 < \ldots < M_m$ such as:

$$M = M_1 \cdot M_2 \cdot \ldots \cdot M_m \in [2^b, 2^{b+1}] \tag{2}$$

It is necessary for having unique decomposition any symbol of b-bit size in remainders by these numbers [2].

2) Generation of Permutations

To provide nonlinear encryption it is necessary to obtain permutations for arithmetic operations for each module from the secret key.

To the module M_i a permutation is a set of prime integers of the type: $\langle a_0, a_1, a_2, \ldots, a_{M_i-1} \rangle$, where $a_i \in [1, M_i - 1]$ and $a_i \neq a_j$, $i \neq j$. This set is obtained by Generator by the induction algorithm [1].

The permutation changes the addition and multiplication tables for each module [1].

3) Generation of Matrices

For each module invertible matrices $A_1, A_2, \ldots, A_m$ of $n \times n$ size are generated [1].

4) Generation of Huffman Table

In order to increase the cryptographic strength of the algorithm at one of the stages of encryption, we use the Huffman algorithm [1], [3], in which the occurrence frequency for each symbol are obtained by Generator. This choice may be justified by the fact that the combination of Huffman coding and insertion of fake symbols which are described below leads to the changes in the size and content of the encrypted message at every other ciphering though the key is not changed. This helps to resist entropic methods of cracking and known-plaintext attacks [4].

D. Move-To-Front

In order to add the block chaining mode - the phenomenon in which changing the block of source text leads changing not only the corresponding encrypted block, but all following blocks, the technology Move-To-Front [3] is used.

Let us present the Huffman table in the form of two arrays: one contains symbols, which are need to be encrypted, a dictionary, and the other is bit representations. Initially, the dictionary will be represented as a sequence from 0 to M_m. The encoding consists of sequentially traversing all the characters of the encrypted vectors. The symbol is searched for in the dictionary, and its bit representation is written to the output stream. After that, the symbol is removed from its position and inserted into the beginning of the dictionary. The second array remains unchanged.

This work was supported by JetBrains Research

For decoding it is necessary to do similar actions with encoded text, in this case the initial state of the dictionary must be identical to the initial state of the encoding dictionary.

E. Fake Symbols

As mentioned above, the algorithm has a stronger cryptography if encrypted data look different at every other ciphering with the same key. We use fake symbols, which change the look of the encrypted message for this purpose.

Obviously, with a symbol size of b, the encrypted characters can not have a value greater than 2^b. Also, according to the Chinese remainder theory, from the remainders of the division of an integer we can collect a number which is smaller than the multiplication of all modules M [2]. Then, if a number λ is added into a block would satisfy following condition $2^b < \lambda < M$ it will be clear that the number is not a symbol of the source file when the block is deciphered and collected.

Each time a new block is read, a decision, whether to insert a fake symbol, made by Generator. The symbol λ itself is also obtained by means of the Generator [1].

F. Ciphering

We will choose modules: $M_1, \dots, M_m$. Secret key is used to generate permutations, matrices $A_1, A_2, \dots, A_m$ for each module and a Huffman table. Read the block B of size $n * b$ and ciphered as follows:

1. The block $B = q_1, q_2, \dots, q_n$ is represented in the form of a vector $\vec{B} = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{pmatrix}$. The coordinates of the vector is remainders of $M_1, \dots, M_m$. We would have vectors $v_1 = \begin{pmatrix} v_{1_1} \\ v_{1_2} \\ \vdots \\ v_{1_n} \end{pmatrix}, v_2 = \begin{pmatrix} v_{2_1} \\ v_{2_2} \\ \vdots \\ v_{2_n} \end{pmatrix}, \dots, v_m = \begin{pmatrix} v_{m_1} \\ v_{m_2} \\ \vdots \\ v_{m_n} \end{pmatrix}$, where $v_{i_j} = q_j \ mod \ M_i$.
2. Matrices are multiplied by the relevant vectors (in every module), all operation we do by tables of multiplication and addition generated previously, vectors of such form are obtained
$$v_i' = A_i \cdot v_i, \quad \text{where } i = 1..m. \tag{3}$$
3. The resulting vectors are recorded a special order in accordance with the Huffman table. The order is defined by the sequence of numbers $P = \{p_1, p_2, \dots, p_m\}$, where $p_i \in [1, m]$ – position of this module, $p_i \neq p_j$ if $i \neq j$. So, the vectors are recorded as a sequence: $v_{p_1}, v_{p_2}, \dots, v_{p_m}$.

When moving from block to block, matrices, Huffman tables and the order of record of the encrypted vectors are changed. Matrices are changed through string/column exchange operations or transpositions to save their invertibility.

The order of record of the encrypted vectors is changed through the right circular shift of P to the right to number obtained from the Generator.

The Huffman table is modified according to the Move-To-Front algorithm described above, based on the result of vectors multiplication.

G. Deciphering

We will choose modules: $M_1, \dots, M_m$. Secret key is used to generate permutations, matrices $A_1, A_2, \dots, A_m$ for each module and a Huffman table. Inverse matrices $A_1^{-1}, \dots, A_m^{-1}$, all operation we do by tables of multiplication and addition generated previously.

1. We read m encrypted vectors $v_1', \dots, v_m'$, decoding each symbol in accordance with the Huffman table and change the Huffman table according to the MTF dictionary in the opposite direction.
2. An invertible matrix is multiplied by the relevant vector, all operation we do by tables of multiplication and addition generated previously: $A_i^{-1} \times v_i', \ i = 1, \dots, m$. Result of this operation is deciphered vectors $v_1, \dots, v_m$.
3. According to the Chinese remainder theory the obtained vectors are collected into the block B. This is a deciphered block. If it has a symbol the numerical value of which is more than 2^b, the symbol is assumed as fake and is thrown out of the block.

When moving from block to block, the changes in the reading order and in Huffman tables are done similarly to the ciphering procedure. In case of invertible matrices, string permutations are replaced by column permutations. Transposition remains unchanged.

III. PRNG AND CREMONA TRANSFORMATION

A fairly large part of the encryption job is tied to a pseudo-random number generator. In order to exclude the possibility of such an important detail to become a weak point, we decided to take the Cremona transforms [5], [6] as the basis of the principle of the generator's operation. The Cremona transformation is an invertible polynomial mapping from the vector space R^n onto itself, which is given by polynomial functions:

$$\begin{gathered} h_1(x_1, x_2, \dots, x_n), \\ h_2(x_1, x_2, \dots, x_n), \\ \dots \\ h_{n-1}(x_1, x_2, \dots, x_n), \\ h_n(x_1, x_2, \dots, x_n); \end{gathered}$$

invertibility means that equations system:

$$\begin{gathered} h_1(x_1, x_2, \dots, x_n) = a_1, \\ h_2(x_1, x_2, \dots, x_n) = a_2, \\ \dots \\ h_{n-1}(x_1, x_2, \dots, x_n) = a_{n-1}, \\ h_n(x_1, x_2, \dots, x_n) = a_n; \end{gathered} \tag{4}$$

is solvable for any right-hand side.

We fix some module m such that raising to the power $k_1, \dots, k_n$ is a one-to-one mapping modulo m. These degrees can be the same.

We will introduce the mapping:

$$\begin{gathered} p_1 = (u_1)^{k_1} + f_1(u_2, u_3, \dots, u_n), \\ p_2 = (u_2)^{k_2} + f_2(u_3, \dots, u_n), \\ \dots \\ p_{n-1} = (u_{n-1})^{k_{n-1}} + f_{n-1}(u_n), \\ p_n = (u_n)^{k_n}; \end{gathered} \tag{5}$$

Here $f_1(u_2, u_3, \dots, u_n)$, $f_2(u_3, \dots, u_n)$, …, $f_{n-1}(u_n)$ are arbitrary polynomials in the indicated variables. We call this mapping an upper-triangular Cremona mapping and denote this mapping **F**.

The lower-triangular Cremona map is defined similarly. Let the raising to the power $s_1, s_2, \dots, s_n$ to be a one-to-one mapping modulo m. Lower-triangular mapping Cremona:

$$\begin{gathered} q_1 = (v_1)^{s_1}, \\ q_2 = (v_2)^{s_2} + g_1(v_1), \\ \dots \\ q_{n-1} = (v_{n-1})^{s_{n-1}} + g_{n-2}(v_1, v_2, \dots, v_{n-2}), \\ q_n = (v_n)^{s_n} + g_{n-1}(v_1, v_2, \dots, v_{n-2}, v_{n-1}); \end{gathered} \tag{6}$$

we will denote this map **G**.

It is obvious, that both maps are invertible.

We can take a superposition with linear invertible mappings that are represented by $n \times n$ matrices, and these matrices are invertible modulo m. In addition, we can use invertible affine mappings that are determined by an invertible matrix and the vector **w**. Let us give a simple example. Consider the case:

$$\begin{gathered} n = 2, \\ m = 11; \end{gathered}$$

Upper-triangular transformation:

$$\begin{gathered} p_1 = (u_1)^3 + 2 * (u_2)^2, \\ p_2 = (u_2)^5; \end{gathered}$$

Lower-triangular transformation:

$$\begin{gathered} q_1 = p_1, \\ q_2 = p_2 + (p_1)^2; \end{gathered}$$

Superposition:

$$\begin{gathered} q_1 = (u_1)^3 + 2 * (u_2)^2, \\ q_2 = (u_2)^5 + [(u_1)^3 + 2 * (u_2)^2]^2 = \\ = (u_2)^5 + (u_1)^6 + 4 * (u_1)^3 * (u_2)^2 + 4 * (u_2)^4; \end{gathered}$$

Suppose given a matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 5 \\ 10 & 2 \end{pmatrix}$$

Consider the superposition of **GAF**.

AF:

$$\begin{gathered} \begin{pmatrix} 3 & 5 \\ 10 & 2 \end{pmatrix} \begin{pmatrix} (u_1)^3 + 2(u_2)^2 \\ (u_2)^5 \end{pmatrix} = \\ = \begin{pmatrix} 3(u_1)^3 + 6(u_2)^2 + 5(u_2)^5 \\ 10(u_1)^3 + 9(u_2)^2 + 2(u_2)^5 \end{pmatrix} \end{gathered}$$

GAF:

$$\begin{gathered} 3(u_1)^3 + 6(u_2)^2 + 5(u_2)^5 \\ 10(u_1)^3 + 9(u_2)^2 + 2(u_2)^5 + \\ + [3(u_1)^3 + 6(u_2)^2 + 5(u_2)^5]^2 \end{gathered}$$

This composition can be used any number of times, thus increasing the degree of polynomials. In this case, if the degrees of the transformations p_i and q_i are less than or equal to 2, then the possibility to restore the initial data will be preserved. That as a result will make it possible to infinitely complicate the relationship between the initial data. The best application of this fact can be found in the generation of public keys and the exchange of keys.

By increasing the dimensionality of the matrix, we can also control the length of the resulting vector, which makes the algorithm easily scalable.

Cremona transformations can be used not only to generate a sequence of polynomials of any degree, but also to generate pseudo-random vectors. Consider the following example:

$$\begin{gathered} m = 11, \\ x = u^7 + v^4 + 5, \\ y = v^5, \\ p = x, \\ q = x^2 + 2y, \\ x_1 = 10, \\ y_1 = 8, \\ x_{i+1} = {x_i}^7 + y_i^4 + 5, \\ y_{i+1} = ({x_i}^7 + y_i^4 + 5)^2 + 2 * y_i^5; \end{gathered}$$

$\binom{10}{8}; \binom{4}{9}; \binom{0}{5}; \binom{10}{5}; \binom{9}{5}; \binom{3}{6}; \binom{8}{9}; \binom{8}{6}; \binom{1}{10}; \binom{3}{1}; \binom{0}{6}; \binom{10}{1}; \binom{1}{0}; \binom{2}{5}; \binom{6}{0}; \binom{9}{1}; \binom{6}{7}; \binom{1}{3}; \binom{6}{6}; \binom{7}{2}; \binom{1}{2}; \binom{7}{7}; \binom{10}{10}; \binom{1}{7}; \binom{5}{2}; \binom{9}{9}; \binom{10}{3}; \binom{4}{2}; \binom{0}{1}; \binom{2}{7}; \dots$

This sequence has a period of about 30 operations, but if you change the input data and the transformation degree, we can achieve a lager period. This fact has yet to be explored in more detail. Since for the generator we do not need the invertibility of operations, the degrees of the transformations p and q can be arbitrary. However, the largest period is observed in powers that can be uniquely inverted for most field elements.

Generators of this kind will be useful if there will be necessity for generating a set of numbers for further work, for example, determining the insertion of fake symbols into a sequence of blocks at once, and not on each block separately.

IV. Cryptanalysis

A. Probabilistic encryption

The basis of probabilistic encryption in our algorithm is fake symbols. These symbols can easily be added to the source text, using any PRNG to determine the character and its position, and it is also easy to detect them while decrypting. The algorithm also does not impose any restrictions on the number of fake characters in one block, which allows you to change their number depending on our needs. Therefore, on small texts you can use not only fake symbols, but also the fake blocks, distributing useful information among random places of ciphertext.

The number of multipliers, the source text block divided by, also affects the size of the ciphertext. In combination with the Huffman table, this fact allows you to fully control the size of the encrypted text, that means you can both increase the size of the output text, and reduce.

Thus, the size of the resulting text becomes unpredictable, which forces the attacker to pick up the initial parameters before the attack begins, that is completely restore the algorithm. So, the problem of hacking is a full search of options. Even on the same source text and the same key, the

ciphertext will be different. This is ensured by the use of a PRNG, based on some non-persistent parameter, for example, the time, the processor clock rate. As a source of entropy, any time-dependent variable will do.

B. Avalanche effect and block chaining mode

If any bit in any block is changed, the entire block changes on average by 45% after encryption [4]. This property increases encryption strength, due to the fact that attacks, based on small changes in the source text, stop working.

Block chaining is provided by changing the Huffman table for all following blocks, depending on the results of the encrypted block. So, changing any bit in the block, changes the result of encryption not only of this block, but of all the following. Depending on the Huffman table, changes in subsequent blocks are in range from 40% to 60% and on average provide good block chaining.

C. Cremona transformations

In the existing version, the Cremona transformation is used to increase the entropy of the Generator seeds, which allows use the Generator based on the current time not so often. The key is sufficient to create a whole set of generators, based on these transformations, which can be switched during operation, ensuring the Generator operation unpredictability.

Another option is to generate fake characters at once for a set of blocks. This way will allow to access to the generator less often, which will significantly increase the speed of encryption. Moreover, the fact of working in the fields allows not to worry about a great increase in the transformations degrees, which will provide an acceptable generation rate.

V. PERFORMANCE

The sizes of the read character and block are 16 bits and 8 characters respectively. The key is 256 bits. Modules are: 5, 7, 11, 17, 19. Processor: Intel Core i7-4720HQ 2.6 GHz. The test results for files of different sizes are presented below:

TABLE I.
TEST RESULTS

Algorithm	**File Size (Bytes)**	**Encrypt time (s)**	**Decrypt time (s)**	**Encrypt file size (Bytes)**
Our algorithm	1 048 576 (1 MB)	0.24 – 0.27	0.23 – 0.25	1 726 399 – 1 727 072
AES-256	1 048 576 (1 MB)	0.035	0.068	1 048 576
Our algorithm	104 857 600 (100 MB)	25.12 – 25.96	23.87 – 24.42	172 611 865 – 172 622 351
AES-256	104 857 600 (100 MB)	2.63	5.96	104 857 600

It can be seen that our algorithm loses in speed. This is due to the fact, that the algorithm is probabilistic and, in fact, a greater amount of data is encrypted, than is fed to the input. Also, the current version of the algorithm is just an early prototype, and we are still working on optimization. Nevertheless, even now, because of the great flexibility of the settings, you can achieve a significant speed increase.

VI. CONCLUSION

In this article, a modification of a fully probabilistic cipher based on the theory of information compression and Cremona transformation was presented. This modification can be useful in various areas, since it has a modular structure. Probabilistic encryption guarantees high cryptographic strength for any application. In future research, we hope to optimize this algorithm to make it lightweight, and explore the application of Cremona transformations more. We hope that in the near future there will be an increased interest in probabilistic encryption, so that there is an opportunity to actively develop this direction and compare this algorithm with analogues.

ACKNOWLEDGMENT

We greatly thank our scientific director Sergey Krendelev for support and comments that improved the paper. We would also like to show our gratitude to the team of key exchange, under the direction of Sergey Krendelev, for sharing their ideas with us during the course of this research.

REFERENCES

[1] S Krendelev, N Zbitnev, D Shishlyannikov and D Gridin, “Block cipher based on modular arithmetic and methods of information compression” IOP Conf. Series: Journal of Physics: Conf. Series 913 (2017) 012009 https://doi.org/10.1088/1742-6596/913/1/012009

[2] Vinogradov I M “Elements of Number Theory”, 5th ed Kravetz S, Dover, 1954

[3] Nelson M 1995 “The Data Compression Book”, 2nd Edition IDG Books Worldwide Inc

[4] Schneier B “Applied Cryptography Second Edition”, John Wiley & Sons Inc, 1996

[5] S. Cantat “The Cremona group in two variables”, Proceedings of the sixth European Congress of Math., pp. 211–225, Europ. Math. Soc., 2013

[6] S. Cantat “The Cremona Groups”, to appear in Proceedings of 2015 Summer Institute on Algebraic Geometry, AMS Proceedings of Symposia in Pure Mathematics

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 607–616
DOI: 10.15439/2018F11
ISSN 2300-5963 ACSIS, Vol. 15

Volatile memory-centric investigation of SMS-hijacked phones: a Pushbullet case study

Mark Vella
Department of Computer Science
University of Malta
Msida, Malta
Email: mark.vella@um.edu.mt

Vishwas Rudramurthy
Department Of Computer Science and Engineering
Channabasaveshwara Institute Of Technology
Gubbi, Tumkur, India
Email: vishwas.rudramurthy@cittumkur.org

***Abstract*—Cloak-and-Dagger attacks targeting Android devices can completely hijack the UI feedback loop, with one possible consequence being that of hijacking SMS functionality for cybercrime purposes. What is of particular concern is that attackers can decouple stealth activities from SMS hijacking. Consequently the latter could be pulled off using completely legitimate apps that normally would allow users to manage text messages from their personal computers (SMSonPC), but this time all hidden away under attacker control. This work proposes a digital investigation process aiming to uncover SMS-hijacked devices. It uses bytecode instrumentation in order to force the dumping of volatile memory areas where evidence for the hijack can be located. Eventually both the malware that conceals the SMS-hijacking and the compromised or smuggled SMSonPC app can be identified. Preliminary results are presented using a case study based on the popular SMSonPC app: Pushbullet.**

I. Introduction

THE Cloak-and-Dagger set of attacks demonstrates how through the abuse of two permissions, Android malware can take control of the entire User Interface (UI) feedback loop [1]. Essentially what this means is that through malicious crafting an attacker can snoop on or even take full control over a user's intentions when interacting with a smart-phone using touch screen taps and swipes, and conversely of all the device's reactions to them. The consequence is a Man-in-the-Middle (MiTM) posture for an attacker sitting in between users and their devices. This is a critical game changer in the sense that up till this point it was generally thought that UI attacks were more about forcing users to send clicks to marketing referral web-sites, rather than completely hijacking a device. This role up till this point was reserved to rooting/jail-braking malware that takes advantage of memory corruption errors inside firmware. The two abused permissions relate to accessibility (a11y) and overlay drawing (draw-on-top) functionality. The former permits an app to access the UI widgets of a second app, whilst the latter permits an app to draw overlays on top of on another app's UI. Their combined abuse can be disastrous due to the long-term stealth an attacker can attain.

The threat that we are concerned with in this work leverages these two permissions to silently install or compromise one of those apps that let users send/read SMS text messages from their personal computers (PCs). These apps are gaining popularity since in the larger context they let users manage all of their smart devices (phones, tablets, wearables and what not) from a single machine[1]. In the specific case of text messages, typing them on a PC keyboard is of particular convenience whenever possible, and for the rest of this paper we will refer to apps that offer this functionality as *SMSonPC*.

Once Cloak-and-Dagger malware tricks victims into giving up or stealing their login SMSonPC credentials, it moves on to activate a11y and conceal SMS-related activity by abusing draw-on-top permissions. At this point the SMSonPC app provides an attack vector to hijack the device's SMS functionality in a highly stealthy manner. What is of major concern is that the SMSonPC app in question is totally legitimate, possibly installed by the user in the first place. Furthermore, the use of draw-on-top and a11y features have been picked up by popular apps and at this point it can be very difficult to make amends from Android's end.

While Google Play's screening has been tightened accordingly, it is a well known fact that persistent attackers tend to succeed in eventually having their malware included in this trusted app store. Android Oreo also includes tightened security, yet its fragmented adoption is still expected to stand in the way[2][3]. Moreover, mitigations only address overlay drawing and which could potentially be replaced by social engineering tricks nonetheless. Further details with respect to the hijacking procedure and existing digital investigation options are provided in section II.

The idea behind the proposed digital investigation process (section III) is that in the event of a suspected SMS-hijacking, or else on a routine basis, users will be able to investigate their devices for possible infection. This approach aims directly at the core of the issue: long-term stealth. During a first stage those apps that look suspicious, either because of the aforementioned requested permissions or else due to SMS functionality, are extracted from the device in order to have their bytecode instrumented. The injected bytecode forces the dumping of those volatile memory areas where evidence uncovering the hijack could be located, without necessarily requiring device rooting. During a second stage

[1]https://www.androidauthority.com/apps-send-text-sms-pc-ways-740669/
[2]https://www.wired.com/story/cloak-and-Dagger-android-malware/
[3]https://developer.android.com/about/versions/oreo/android-8.0-changes.html#all-aw

of the investigation forensic analysis is conducted upon the collected memory dumps. They are combined with the context provided by text messages from flash/SIM memory along with any suspicious destination numbers as obtained from operator billing logs. In the interim, the device is used normally except for the additional recording of potential artifacts that can uncover both the malware that sets up and conceals the SMS-hijack, as well as the compromised/smuggled SMSonPC app. Basically any text message flows inferred to originate/end from/at SMSonPC apps without the device's owner consent, and in the presence of a draw-on-top/a11y app, indicate an ongoing SMS-hijack.

A case study using Pushbullet, a popular SMSonPC app, is used for initial exploration of this technique in terms of its effectiveness and practicality (section IV). The proposed SMS-hijack investigation process along with the preliminary results from this case study are the primary contributions of this work.

II. SMSonPC hijacking

The essential ingredients for the stealthy SMS-hijack being considered in our threat model consist of an SMSonPC app combined with a number of Cloak-and-Dagger attack techniques. In this work we focus on the abuse of Pushbullet to serve this purpose.

A. The Pushbullet SMS-hijack scenario

In order to make use of Pushbullet users need to install a controlling application on their PC in the form of a stand-alone native application or a browser extension. Otherwise they may simply log into a web interface[4]. Whichever client option, a device is instructed to send a text message by means of what is called an ephemeral message, which is possibly encrypted, and an example of which is shown in Listing 1. This is a JSON-formatted object which is sent to the Pushbullet server by the controlling application. In this case the instruction is to send a `Hello!` text message to `+1 303 555 1212` on behalf of user-id `ujpah72o0` through her Pushbullet-registered device with identification `ujpah72o0sjAoRtnM0jc`. This is an example of a Pushbullet `push` event, intended for dispatch to the identified Android device, and which accesses its SMS services as specified by the `messaging_extension_reply` type using the `com.pushbullet.android` package.

Listing 1
A Pushbullet ephemeral message instructing a phone to send an SMS text message.

```
{
  ``push": {
    ``conversation_iden": ``+1 303 555 1212",
    ``message": ``Hello!",
    ``package_name": ``com.pushbullet.android",
    ``source_user_iden": ``ujpah72o0",
    ``target_device_iden": ``ujpah72o0sjAoRtnM0jc",
    ``type": ``messaging_extension_reply"
  },
  ``type": ``push"
}
```

[4]https://www.pushbullet.com

Cloak-and-Dagger is really a collection of attacks [1] that abuse Android draw(ing)-on-top of opaque or transparent overlays and a11y system services. They require the `SYSTEM_ALERT_WINDOW` and `BIND_ACCESSIBILITY_SERVICE` permissions respectively. Since attacks #3 and #4 (as identified in [1]) are able to hijack the device's virtual keyboard, they could be used to steal Pushbullet credentials during installation/configuration. The prior attack succeeds by placing multiple transparent "pass-through-clicks" overlays per keyboard button and then snoops on keystrokes by having all the overlays capture clicks outside their region. Subsequently it identifies the tapped button using a clever Z-order trick. The latter attack abuses accessibility services by listening to keyboard button click notifications. Attack #5 provides an alternate hijacking strategy and combines the two permissions. It exploits accessibility services to detect that the user has navigated to the Pushbullet app, and then proceeds to exploit draw-on-top by displaying a fake but authentic-looking Pushbullet log-in screen. At that instance it lures users to send their credentials directly to the attacker.

Through Cloak-and-Dagger an attacker can even move on from compromising a user-installed Pushbullet installation to the silent installation of a covert one. Specifically through attack #8, using the standard Android API an attacker can initiate an installation of Pushbullet as well as programmatically confirming the same action when prompted, all the while covering this activity through a draw-on-top overlay. Subsequently, through a11y services, the malware can proceed to cover its tracks by accessing the "recent windows" view and dismissing all of its content. The final step is to launch attack #9, i.e. navigating to app settings and outright enabling all the permissions required by Pushbullet. Consequently the user won't get prompted to grant permissions when subsequently Pushbullet is launched remotely by an attacker to disclose or send SMSes on the device's owner behalf, and thereby maintaining stealth. It is noteworthy that both draw-on-top and a11y features come along with mechanisms to protect from abuse, yet the Cloak-and-Dagger attacks don't simply bypass these protections but also go as far as abusing them. For example the aforementioned Z-order trick exploits the same security flag that informs a clicked widget about whether the click passed through an overlay drawn on top of it.

Having obtained access to draw-on-top and a11y permissions through deceit, along with a compromised or smuggled SMSonPC app through Cloak-and-Dagger, an attacker can now proceed with mischief. For example, the device can be turned into a crime text messaging proxy or even into a spying device by leaking message content. Maintaining stealth in the former case can be achieved by deleting all sent messages, once again possibly through Cloak-and-Dagger means. In the latter case it is a question of whether the SMSonPC has been smuggled or compromised. In the first case, it is simply a question of keeping the SMSonPC app installation concealed from the device owner, while the second case also requires that attackers hide their tracks within the SMSonPC controlling

app. Whatever the scenario the end result is that of a stealthy SMS-hijack.

B. Android SMSonPC apps

Any SMSonPC app requires the `SEND_SMS` and `READ_SMS` permissions in order to be able to interact with the device's SMS features. The `INTERNET` permission is also required to provide a communication link with the SMSonPC server. This also applies to Pushbullet. In particular, the `SEND_SMS` permission provides access to the SMS manager service and through which app components can send text messages by calling `SmsManger.getDefault().sendTextMessage()`. An alternate method forgoes permissions by instead delegating message sending to a privileged app by means of a `startActivity(intent)` call, where the `intent` argument would have been associated with an SMS-related action. Reading of inbox/draft/outbox/sent messages on the other hand requires access to the SMS provider (`android.provider.Telephony.Sms`), which is populated from an SQLite database file that persists text messages, and which can be accessed through `getContentResolver.query()` calls.

As of Android Kitkat[5] only a designated default messaging app is actually permitted to write to this provider. This app also has exclusive privileges to handle incoming text messages. However it is then obliged to inform all interested apps of a newly delivered message, as well as to be delegated with message sending duties by unprivileged apps. It is perfectly possible that an SMSonPC app is also the designated default messaging app. That would facilitate even further the deletion of sent messages as part of the crime-proxy's functionality.

C. Limitations with existing digital investigation options

In the event of an SMS-hijack incident, existing options for digitally investigating it encompass examining the phone's SIM and flash memory for all stored text messages. This process comprises forensic imaging followed by the decoding steps concerning the manner with which text messages are encoded. SIM memory uses GSM-7 or the now obsolete GSM-8 or UCS2 encodings [2]. Android phones store text messages inside SQLite database files where UTF-8 or UTF-16 string encoding can be employed [3]. In this case there is the added difficulty that Android does not allow flash memory imaging without prior device rooting. In many cases this could be problematic due to warranty voiding, as well as it leaves the device's protection again future re-infection weakened. A more practical solution would be to simply install an SMS backup/recovery app that extracts all tables/columns individually from the SMS provider's SQLite database file. Such apps only require the `READ_SMS` permission to function, in addition to permission to copy messages to the some target destination.

[5]https://android-developers.googleblog.com/2013/10/getting-your-sms-apps-ready-for-kitkat.html

In any case the SMS crime-proxy text messages would have been cleared up using Cloak-and-Dagger steps that interact with the default messaging app. In the spying device's case the context associated with text messages inside the SMS provider only identifies the creator rather than the reading apps and is therefore useless. In fact both scenarios could only be fully reconstructed by tracing and preserving the entire sequence of events that lead to sending/deletion/reading of specific messages. Artifacts found inside volatile memory could potentially serve this purpose, however the ones concerning text messages are expected to be short-lived and all existing volatile memory dumping techniques require device rooting [4], further complicating matters.

III. Volatile memory-centric investigation

The proposed SMS-hijack investigation process is based directly on those components involved in the sequence of events when sending and reading SMS text messages, as controlled by a Cloak-and-Dagger malware. The volatile memory of these components is a candidate source for investigation-relevant artifacts, specifically the text messages themselves. The interfacing between these components is also of interest since the relevant code execution presents candidate triggers, indicating the presence of text messages within the memory areas of interest at that point in time.

A. Abused SMS components

Event sequence mapping for message sending/reading flows as abused during an SMS-hijack incident was carried out directly upon Android's source code[6], with guidance from literature sources that describe its core inter-process communication [5] and telephony stacks [6]. Figure 1 depicts the components and interfacing involved when covertly sending text messages. Firstly, the Cloak-and-Dagger malware itself needs to conceal from the user any activity related to SMS being conducted by the SMSonPC app. Draw-on-top overlays require a `TYPE_SYSTEM_OVERLAY` layout and which has now been deprecated by the more restrictive `TYPE_APPLICATION_OVERLAY`. However the new permission is only relevant for user-installed apps that do not need to be compatible with Android versions older than Oreo. Attackers interact remotely with the SMSonPC app to send instructions for sending/retrieving text messages, e.g. Listing 1, typically through HTTP(S).

The direct route for sending text messages is through the `SMSManager` service and which is hosted by the phone process `com.android.phone`. Most inter-process communication in Android happens through `Binder`, a Remote Procedure Call (RPC) mechanism, in order to trigger `SmsManager.sendTextMessage()`'s code execution. Message dispatching includes two important steps. First, the outgoing message is written to the `mmssms.db` SQLite database file by calling into the Linux Virtual File System (VFS) and eventually writing to the phone's

[6]https://source.android.com/

flash memory. This step is mandatory unless the originator of the message is the default messaging app. Secondly, the message is dispatched to the baseband processor. The `RIL.sendMessage()` triggers a chain of events that cause the outgoing message to formatted in a communication protocol-independent manner (PDU format). It is sent through a UNIX domain socket to the `rild` native daemon that in turn interfaces with a vendor-specific library. When its `ProcessCommandBuffer` event loop receives the RIL request number 25 (`RIL_REQUEST_SEND_SMS`), this library initiates message sending by calling into the baseband driver code via an `ioctl`. Eventually the baseband processor physically sends the text message onwards to the operator's core network via the closest base transceiver station. The operator logs the event for billing purposes even though this excludes message content [2].

An alternate indirect path is possible whenever an unprivileged SMSonPC app interacts with the default messaging app using intents. Intents are resolved by the `ActivityManagerService`, which is hosted by the `System Server` daemon and reachable through Binder RPC. The continuation path is similar to that of the direct path, except that at this point the persistence of outgoing messages is at the discretion of the default messaging app.

The primary components involved with the reading of text messages are shown in Figure 2. The SMS provider inside the phone process, as populated from `mmssms.db` through a call to `SQLiteQueryBuilder.query()`, is central to this operation. This provider can also get populated from the SIM memory through a `SmsManager.getDefault().getAllMessagesfromICC()` call, and which in turn sends a RIL request number 28 (`RIL_REQUEST_SIM_IO`). The task of obtaining a reference to a `ContentResolver` instance for calling `query()` is mediated by `ActivityManagerService`. Ultimately, the retrieved message is covertly leaked to the attacker via HTTP(S).

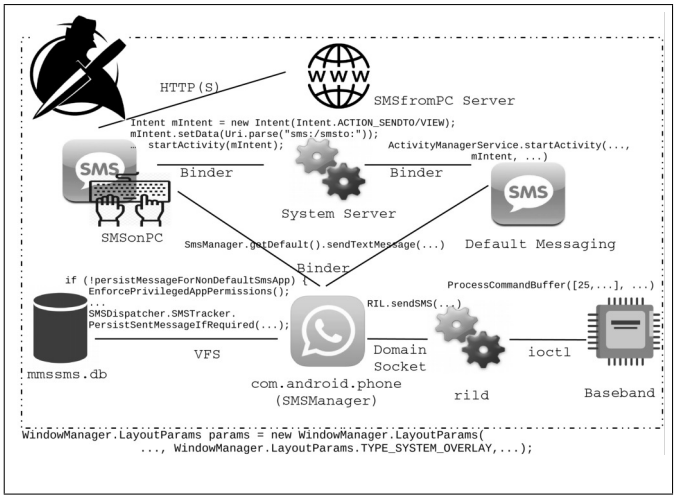


Fig. 1: Android components abused to send SMS text messages and concealed by Cloak-and-Dagger.

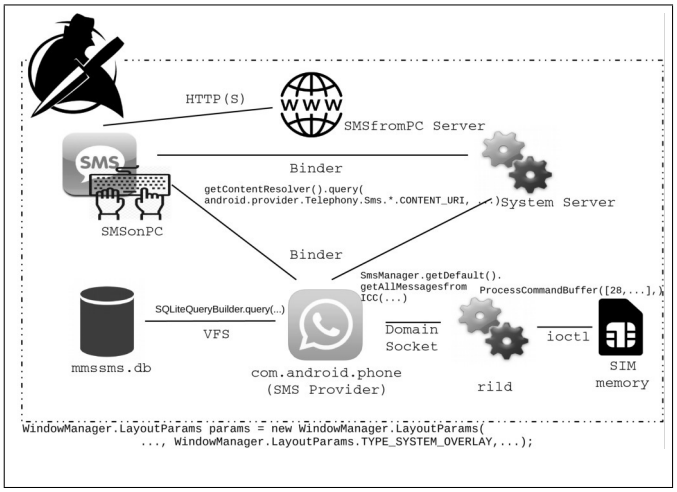


Fig. 2: Android components abused to read SMS text messages and concealed by Cloak-and-Dagger.

B. Observations

Potentially, any text message flow originating from an SMSonPC app and which after passing through system components terminates in flash/SIM memory, coupled with suspicious draw-on-top and a11y activities, should raise an alert of a possible SMS-hijack. The same argument applies for the inverse route. The device owner can confirm whether the observed flows had their consent or otherwise, at which point the suspicious app is identified as the Cloak-and-Dagger malware while the SMSonPC app is confirmed to have been compromised. Given that fully tracing these flows for prompt SMS-hijack detection is expected to be particularly expensive in terms of runtime overheads, an alternate practical approach is to defer detection during memory forensics analysis [7]. The idea is to keep track just of the key in-memory artifacts related to these flows, as occurring inside the memory space of apps and intermediate system components, and from which to infer their occurrence.

The main challenge however is presented by the brief permanence in memory of the said artifacts, calling for an event-driven collection approach. Bytecode instrumentation is a key enabler, whereby injected bytecode is responsible for initiating memory dumps at the appropriate SMS-hijack triggers. The most obvious solution is to focus on SMSonPC apps, since they can be statically instrumented through app repackaging and without the need of device rooting. Furthermore, instrumenting system components is still deemed a less desirable option due to the instability that it might incur. Yet, on a non-rooted device the injected bytecode would only have the faculty to dump the Dalvik (Java) heap, and would therefore miss those artifacts that would rather reside on the native heap whenever native components are employed. In cases where device rooting is viable, rather than having to analyze all native heaps of a myriad of SMSonPC apps it could suffice to inspect just that of the Android phone process. This heap is expected to be relatively constant across devices.

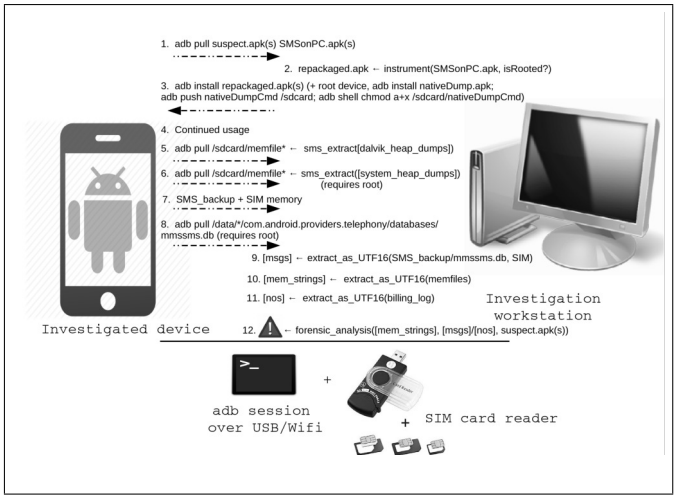


Fig. 3: The volatile memory-centric process for digitally investigating SMS-hijacking.

As can be observed from Figures 1 and 2 the phone process is central to both the SMS sending and reading flows. Also, due to the native Binder RPC mechanism it must process text message flows on its native heap.

C. The SMS-hijack investigation process

The proposed digital investigation steps carried upon devices with a suspected SMS-hijack, as shown in Figure 3, are based on the observations just made. They require the inspected device to be connected to an investigation workstation over an Android Debug Bridge (adb) session. The first 3 steps are concerned with identifying suspicious apps and instrumenting the trigger points for dumping volatile memory. The focus is on those packages (in apk format) that request relevant permissions. They must either request access to draw-on-top/a11y services that render them Cloak-and-Dagger suspects, or else they request SMS permissions and therefore are potentially abused SMSonPC apps. These apps are discernible from the device's `/data/system/packages.xml`.

Once pulled from the device (step 1), the potential SMSonPC apps are repackaged (step 2) with memory-dumping instrumentation and re-installed on the device in place of the original ones (step 3). Algorithm 1 describes the `instrument()` function applied on each `SMSonPC.apk`, and which takes into account whether the device will be rooted in step 3 as indicated by the `isRoot` flag. Lines 1-4 populate the filters for SMS or native method call-related trigger points as per component interfacing described in section III-A. These are defined using smali syntax[7] which is an assembly language for Dalvik bytecode[8]. The additional wild cards $\star$ are meant to match any smali statements in a non-greedy manner.

Trigger points are instrumented with Dalvik heap-dumping bytecode (`dalvik_dump_instr` on lines 5-12) and possibly also that of the native heap of the phone process (`systemmem_dump_instr` on lines 13-20). The

Algorithm 1: Step 2: `instrument`

Input: Potentially abused app: SMSonPC.apk, Root mode: RootFlag
Output: Repackaged and cracked app: repackaged.apk

```
1  [sms_trigger_filters] ← "invoke-direct{*},
     Landroid/telephony/SMSManager;
     ->send(Multipart)TextMessage(*)V";
2  [sms_trigger_filters] ← "const-string "sms*
     invoke-static{*},
     Landroid/net/URI;->parse(*)Landroid/net/URI;*
     invoke-direct{*},
     Landroid/content/Context;->startActivity(*)V";
3  [sms_trigger_filters] ← "(sget-object v*,
     Landroid/provider/Telephony/Sms/*CONTENT_URI:|
     "content://mms-sms")* invoke-direct{*},
     Landroid/content/ContentResolver; ->query(*)V";
4  [native_trigger_filters] ← "invoke*{*},
     *->nativeMethod(*)";
5  dalvik_dump_instr ← "invoke-static{},
     Ljava/lang/System;->currentTimeMillis()J
6  move-result-wide vA
7  invoke-direct{vA}, Ljava/lang/long;
     ->toString()Ljava/lang/String;
8  move-result vC
9  const-string vD, "/sdcard/hdump_hprof_"
10 invoke-direct{vD,vC}, Ljava/lang/String;
     ->concat(Ljava/lang/String;)Ljava/lang/String;
11 move-result vE
12 invoke-static{vE}, Landroid/os/Debug;
     ->dumpHprofData(Ljava/lang/String;)V";
13 systemmem_dump_instr ← "const-string vA,
     "com.inspect.nativeDump"
14 const-string vB ,"com.inspect.nativeDumpSrvc"
15 new-instance vC,
     Landroid/content/ComponentName;
16 invoke-direct{vC, vA, vB},
     Landroid/content/ComponentName;
     -><init>(Ljava/lang/String;
     Ljava/lang/String;)V;
17 new-instance vD, Landroid/content/Intent;
18 invoke-direct{vD}, Landroid/content/Intent;
     -><init>()V;
19 invoke-direct{vD, vC}, Landroid/content/Intent;
     ->setComponent(Landroid/content/ComponentName;
     )Landroid/content/Intent;
20 invoke-direct{p0, vD},
     Landroid/content/Context;
     ->startService(Landroid/content/Intent;)
     Landroid/content/ComponentName;";
21 [smali_class_files] ← unpack_apk(SMSonPC.apk);
22 foreach smali_class_file in [smali_class_files] do
23     smali_class_file ←
         crack_anti_tamper(smali_class_file);
24     smali_class_file ← unpack(smali_class_file);
25     while trigger_point ←
         getNextTriggerPoint(smali_class_file,
         [sms_trigger_filters] ∪ [native_trigger_filters]) do
26         InstrMethodStart(smali_class_file, trigger_point,
             dalvik_dump_instr);
27         if RootFlag then
28             InstrMethodStart(smali_class_file,
                 trigger_point, systemmem_dump_instr);
29     end
30     repackaged.apk ← smali_class_file;
31 end
32 return apkSign(repackaged.apk);
```

[7]https://github.com/JesusFreke/smali
[8]https://source.android.com/devices/tech/dalvik/dalvik-bytecode

latter relies on device rooting as well as the the installation of `nativeDump.apk` and `DumpCmd`, as per step 3 of Figure 3. `nativeDump.apk` exposes the `nativeDumpSrvc` service component reachable through `startService()` calls from the instrumented apps. In turn, `nativeDumpSrvc`'s implementation calls the `su; /sdcard/DumpCmd` shell command sequence. `DumpCmd` is a native process responsible for dumping the phone process's native heap via `/proc/<pid_suspect/phone>/maps` and `/proc/<pid_suspect/phone>/mem`. All memory dumps are placed on external (common app) storage area on the file-system (`/sdcard`). Within the same location the `sms_extract()` component is responsible to extract just the SMS-related memory areas, saving on space requirements. This file-system location facilitates later retrieval from the investigation workstation without requiring device rooting.

For both instrumentation code, the Dalvik VM register numbers `vA-vD` have to be adjusted so that no clashes occur, possibly also requiring an adjustment to the `.locals` smali directive. This directive declares the number of Dalvik VM registers needed to store the local variables of a class method (excluding method parameters). Furthermore, not shown in the instrumentation bytecode of Algorithm 1 is exception handling code, as well as an additional snippet that combined with `AndroidManifest.xml` permission entries for `READ_EXTERNAL_STORAGE` and `WRITE_EXTERNAL_STORAGE` requests access to the external storage. This operation would be required only by those apps not already including this functionality, with the instrumentation bytecode placed inside the `onCreate()` method of the app's main activity.

The trigger filters are applied for each smali representation of the compiled app classes (`smali_class_file`), as obtained through apk unpacking (`unpack_apk()`) (lines 21-31) by calling `getNextTriggerPoint()`. The identified trigger points are then instrumented by calling `InstrMethodStart()`. This is a routine that attempts to inject the instrumentation bytecode at the very start of the method containing the trigger point. This approach avoids having to renumber Dalvik registers to address clobbering, and therefore not running the risk of exceeding the highest register usable by most Dalvik opcodes (`v15`). Instrumentation for native heap dumping is only carried out in case of device rooting. Each, possibly instrumented, smali class file is eventually added to the repackaged app `repackaged.apk`. Finally, the app is signed (line 32) and is ready to be deployed back to the investigated device.

The pending explanation concerns lines 23-24. These are two pre-processing operations that would have to be applied in case the SMSonPC apk is hardened with anti-tampering (`crack_anti_tamper()`) and packed (`unpack()`) code. Their implementation is orthogonal to our work, rather the investigator must seek the assistance of third-party tools in order to successfully pre-process the said class files, with good disassembly skills coming in very handy.

Once step 3 (Figure 3) is complete, the device is returned to its owner for continued usage during step 4. Its duration is bounded by the space available for memory dumps (the `memfiles`). On investigation resumption, steps 5 and 6 take care of retrieving them from the device. Step 7 retrieves the available SMS text messages from flash memory using any SMS backup app, as well as those in SIM memory using an appropriate card reader. Additionally, on rooted devices text messages can rather be extracted directly from the `mmssms.db` SQLite database files in step 8. Steps 9-10 proceed with extracting and normalizing to UTF-16 the text message details as well all strings from the memory dumps. Step 11 on the other hand performs the same operation for those suspicious destination numbers obtained from the billing log.

The normalized content is now ready to be used for forensic analysis. In the case of a text message leakage investigation, the text messages from step 9 are central to the investigation starting point. In the case of a crime-proxy attack, where sent messages are deleted for stealth, the suspicious destination numbers from step 11 become essential. At this point, the aim of the investigator is to trace the messages/numbers inside the dumped strings, and from which to attempt to maximize the identification of SMS-hijack related artifacts. The non-comprehensive list includes: sent/leaked message times, crime-proxy message content, SMSonPC account details in case it has been smuggled, identification of the implicated SMSonPC app in case of multiple candidates, and ultimately the Cloak-and-Dagger malware itself.

IV. Case study: Pushbullet

In order to assess the potential of the proposed SMS-hijack investigation process we present a case study involving the widely used Pushbullet SMSonPC app. The chosen scenario is a simulated crime-proxy attack. Its objectives are to: *i)* Report on the instrumentation step (Algorithm 1) as applied to Pushbullet; *ii)* Measure the storage requirements needed for memory dumps, and *iii)* the overheads imposed by bytecode instrumentation; and finally *iv)* Report on the artifacts identified during the forensic analysis step.

The case study assumes a Cloak-and-Dagger malware to have stealthily installed Pushbullet and set up the device to act as an SMS crime-proxy. Eventually a sequence of suspicious outgoing text message destination numbers show up on a detailed break-down of the device owner's phone bill. Step 1 of the investigation identifies a suspicious app that requests draw-on-top and a11y permissions, as well as the Pushbullet app as the possibly abused SMSonPC app. At this stage the investigator is required to conduct the follow-up investigation steps. The full setup consists of an Android Virtual Device (Goldfish), Android Nougat (for Intel Atom), Pushbullet version 17.7.19-288 and Android Debug Bridge 1.0.39. Apktool 2.3.1 was used to assist bytecode instrumentation. Bash scripting was used for prototyping the instrumentation tool as well as native heap dumping.

A. Pushbullet instrumentation

The first two trigger filters from Algorithm 1, i.e. those relevant to SMS crime-proxy, identified two trigger points both inside `com.pushbullet.android.sms.h`'s `void a(String, String, String)` static method. The result of `InstrMethodStart()`'s execution is shown in Listing 2. Line 1 identifies the instrumented method and line 2 shows that the requested number of Dalvik VM registers has been increased from 8 to 12. The registers utilized by the instrumentation bytecode are in the `v2-v9` range, since attempts to make use of `v0` and `v1` resulted in compiler (`dex2oat`) errors. As compared to the abstracted version presented earlier in Algorithm 1, the injected instrumentation does not hard-code the location of the external common storage (line 10), makes use of the convenient `StringBuilder` class (line 15), had to resort to using `const-string/jumbo` (line 21) due to the large number of strings used by Pushbullet, and makes use of a try/catch block (line 31). On successful execution, control flow skips the exception handler and goes straight into the original entry point of the non-instrumented method (line 43), as indicated by the original `.prologue` directive (line 42).

In terms of obscured trigger points Pushbullet shows no signs of packed code or SMS-related native code. This situation simplifies matters with respect to trigger point coverage and avoids the need to dump native heaps. The use of ProGuard (a code obfuscator that is enabled by default in Android Studio) is not an obstacle either since since trigger points are defined over Android API calls. No anti-tamper protection was encountered either, although the repackaging of Pushbullet did affect Google sign-in's functionality. The case study was eventually conducted using the Facebook sign-in option since this functionality was not broken. Yet, this was an eye-opener on the perils of instrumentation. Finally since Pushbullet already requests access to external storage, no further bytecode instrumentation was necessary in this respect.

B. Storage requirements

The storage requirements were calculated on the basis of an estimated average of 33 daily sent text messages[9]. In turn this translates to 33 Dalvik heap dumps and a possible additional 33 native heap dumps per day. Table I shows the storage requirements for Pushbullet (Dalvik heap) and Android's default phone process (native heap) dumps. While this case study does not strictly require the latter they are included to present a more complete picture.

In both cases the figures for both full and SMS-related area dumps are provided. In the case of Pushbullet, the SMS areas are those containing ephemeral messages as per Listing 1. In the case of the phone process a more generic approach was followed by taking into consideration all areas containing UTF-8/16 strings. This is the main reason why native heap dump sizes are significantly larger ($> \times 100$). However, dump size reduction is staggering in both cases. The 0 standard deviation for full native dumps derives from the fact that their size did not change throughout the entire time-frame of sending the text messages. On the other hand the garbage-collected Dalvik heap was more dynamic.

[9]https://www.textrequest.com/blog/many-texts-people-send-per-day

TABLE I
EST. DAILY STORAGE REQUIREMENTS FOR MEMORY DUMPS.

Dump mode	mean (kB)	std. dev. (kB)	sum (kB)
Dalvik heap - Full	11,608	244.752	380,000
Dalvik heap - SMS only	5	1.929	163
Native heap - Full	31,457	0	1,000,000
Native heap - SMS only	505	89.298	16,656

TABLE II
RUNTIME OVERHEADS.

Configuration	mean (s)	std. dev. (s)	Mann-Whitney (p-value=0.93)
`pushbullet.apk`	0.22	0.02	Sum of ranks - 1099
`repackaged.apk`	0.72	0.4	Sum of ranks - 1112
Overheads	227%	-	U - 538

Overall, the 163kB/day required by Dalvik heap dumps compares well to the approximate 3-5MB typically consumed by a selfie with default resolution. However this figure rises sharply to nearly 17MB had the phone to be rooted and native heap dumping enabled.

C. Runtime overheads

From an end-user's point-of-view the runtime overheads incurred by Pushbullet due to bytecode instrumentation are not noticeable. However, even minimal runtime overheads could be a factor from an attacker's point-of-view had they be exploited to detect an ongoing SMS-hijack investigation. Therefore, overheads were measured when sending SMS text messages from Pushbullet's browser interface. In doing so we gained access to the SMS event profiling logs created by `pushbullet.js` inside the javascript console. The relevant log entries are those of `sms_changed` type and examples of which are shown in Listing 3.

Table II shows statistics for the turn-around times, measured between when a text message is sent and the point at which a notification of completion is received asynchronously in a typical Ajax fashion. When computing overheads incurred by the combined Dalvik/native heap dumping instrumentation over an unmodified Pushbullet configuration, the mean overhead for a daily amount of text messages is a considerable 227%. However, when comparing ranks of the two configurations using a Mann-Whitney test the U value is roughly half that of the sum of ranks for both configurations. This indicates that the difference in mean turn-around times between the two is not statistically significant. This outcome indicates that while the turn-around times for the repackaged configuration were higher, other external factors also had an impact. Therefore their difference is not a reliable measure for attackers to detect an ongoing investigation.

```
.method public static a(Ljava/lang/String;Ljava/lang/String;Ljava/lang/String;)V
.locals 12

invoke-static {}, Ljava/lang/System;->currentTimeMillis()J
move-result-wide v8
invoke-static {v8, v9}, Ljava/lang/Long;->valueOf(J)Ljava/lang/Long;
move-result-object v6
invoke-virtual {v6}, Ljava/lang/Long;->toString()Ljava/lang/String;
move-result-object v5
invoke-static {}, Landroid/os/Environment;->getExternalStorageDirectory()Ljava/io/File;
move-result-object v4

:try_start_0
new-instance v3, Ljava/lang/String;
new-instance v7, Ljava/lang/StringBuilder;
invoke-direct {v7}, Ljava/lang/StringBuilder;-><init>()V
invoke-virtual {v4}, Ljava/io/File;->toString()Ljava/lang/String;
move-result-object v8
invoke-virtual {v7, v8}, Ljava/lang/StringBuilder;->append(Ljava/lang/String;)Ljava/lang/StringBuilder;
move-result-object v7
const-string/jumbo v8, "/hdump_hprof_"
invoke-virtual {v7, v8}, Ljava/lang/StringBuilder;->append(Ljava/lang/String;)Ljava/lang/StringBuilder;
move-result-object v7
invoke-virtual {v7, v5}, Ljava/lang/StringBuilder;->append(Ljava/lang/String;)Ljava/lang/StringBuilder;
move-result-object v7
invoke-virtual {v7}, Ljava/lang/StringBuilder;->toString()Ljava/lang/String;
move-result-object v7
invoke-direct {v3, v7}, Ljava/lang/String;-><init>(Ljava/lang/String;)V
invoke-static {v3}, Landroid/os/Debug;->dumpHprofData(Ljava/lang/String;)V
:try_end_0
.catch Ljava/lang/Exception; {:try_start_0 .. :try_end_0} :catch_0

goto :goto_0

:catch_0
move-exception v2
const-string/jumbo v7, "patchgen"
invoke-static {v2}, Landroid/util/Log;->getStackTraceString(Ljava/lang/Throwable;)Ljava/lang/String;
move-result-object v8
invoke-static {v7, v8}, Landroid/util/Log;->e(Ljava/lang/String;Ljava/lang/String;)I

.prologue
:goto_0
... snip ...
```

Listing 2. A snippet of bytecode instrumentation injected into Pushbullet.

```
pushbullet.js:8615 message {"type":"push","targets":["stream","android","ios"],"push":{"type":"sms_changed",
        "source_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","notifications":[]}} 0.042 s
pushbullet.js:8615 message {"type":"push","targets":["stream","android","ios"],"push":{"type":"sms_changed",
        "source_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","notifications":[]}} 0.117 s
pushbullet.js:8615 message {"type":"push","targets":["stream","android","ios"],"push":{"type":"sms_changed",
        "source_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","notifications":[]}} 0.094 s
... snip ...
```

Listing 3. `pushbullet.js` console log entries for SMS event profiling.

```
"_id","thread_id","address","person","date","date_sent","protocol","read","status","type","reply_path_present",
        "subject","body","service_center","locked","sub_id","error_code","creator","seen"
"1","3","123456","","1520866381969","0","","1","-1","2","",
        "","CrimeProxy sms text message 1","","0","1","0","com.pushbullet.android","1"
"2","3","123456","","1520866402042","0","","1","-1","2","",
        "","CrimeProxy sms text message 2","","0","1","0","com.pushbullet.android","1"
"3","3","123456","","1520866420029","0","","1","-1","2","",
        "","CrimeProxy sms text message 3","","0","1","0","com.pushbullet.android","1"
... snip ...
```

Listing 4. Extract from `mmssms.db`.

D. Forensic analysis

Listing 4 shows an extract from an `mmssms.db` export produced after all text messages were sent. Each entry clearly identifies the destination number (`123456`), message content (`CrimeProxy sms text message n`) and the creator app (`com.pushbullet.android`). If this content was present on the device, any SMS backup app with `READ_SMS` permission would have been able to extract this information to solve the SMS-hijack case. However, with a Cloak-and-Dagger malware that deletes the SMS crime-proxy messages, an investigator would have to resort to the volatile memory dumps for evidence. Starting off from the suspicious `123456` destination number extracted from a detailed bill breakdown (step 11 of the investigation process), the subsequent forensic analysis (step 12) retrieved the entries shown in Listing 5. Each entry provides the missing context from the deleted `mmssms.db` entries, namely the message content. Furthermore given that we are dealing with a Pushbullet dump automatically implicates the app in this SMS activity. Moreover the prefix `ujBeKPNHgJMs` is observed to remain constant throughout all the `iden` entries, identifying the utilized Pushbullet account, and therefore also provides the necessary evidence for reporting abuse. Finally, the `created` and `modified` fields store the timestamps related to the text message sending events.

Listing 6 shows the corresponding native heap dump artifacts as retrieved from the phone process. In this case the user account identification is missing, however the message content and creator app are clearly identifiable. Concluding, in both the Dalvik and native heap dump cases, all information that could have gone missing from `mmssms.db` could be reconstructed. At this point with the device owner's assistance the investigator would be able to confirm whether those outgoing messages were related to an SMS-hijack by confirming the user's consent or otherwise. In the latter case, the suspicious app's bytecode from step 1 should be analyzed in order to identify the Cloak-and-Dagger code.

E. Limitations

An alternative to using message turn-around times in order to detect an ongoing investigation, the Cloak-and-Dagger malware could be equipped with checks for the presence of memory dumps inside external storage, suspending its activities if found. While in a way this can be seen as beneficial, this could be problem if the SMS-hijack operation is resumed as soon as the device returns to normal operation. Furthermore, while in the case of Pushbullet no anti-tamper or obfuscation came in the way, the Google sign-in failure is an eye-opener with respect to the difficulties expected during SMSonPC app instrumentation.

V. Related work

Ideally SMS-hijack attacks are thwarted during the app store upload stage using automated malware analysis. Given that a significant part of the attack is actually carried out by a legitimate SMSonPC app, it is rather the identification of the Cloak-and-Dagger malware that should be targeted. Yet, a number of challenges abound. Firstly app obfuscation, e.g. using encryption and runtime class loading, could hide the malware's real intention from static analysis. This issue could be addressed with dynamic analysis [8] where suspicious apps are executed inside a malware sandbox. However this alternative is not without its own limitations, with trigger-based behavior [9], [10] and device emulation detection-based evasion [11] posing major hurdles. The same limitations are encountered whenever malware analysis is carried out for forensics purposes [12], where malware samples are hunted and extracted from within a mobile device for event reconstruction purposes.

In contrast, our proposed digital investigation process differs in scope. It targets those situations where Cloak-and-Dagger malware succeeds in evading app store scanning. Furthermore, the malware's behavior is tracked within its intended runtime environment, with the exception for SMSonPC app repackaging and the resulting dumps. Our work is more akin to related work concerning the digital investigation of mobile devices, for example for SMS text message forensic purposes [13], [14]. Yet, our proposed technique involves a prolonged investigation period, where the device is returned to its owner for continued usage as enhanced with memory dumping instrumentation. Finally, our proposition can also pave the way to thwart the 'Trojan Horse defense' [15], where text messages considered as evidence for a crime investigation are refuted by claims that the device could have actually been compromised to serve as a communication proxy by the actual criminals.

VI. Conclusions

In this paper we considered the problem of Cloak-and-Dagger malware pulling off stealthy SMS-hijacks by abusing legitimate SMSonPC apps. We proposed a solution whereby injected bytecode instrumentation dumps the SMS-relevant areas of volatile memory from the device under investigation at the right triggers. A case study was carried out using Pushbullet as the SMSonPC app abused for setting up an SMS crime-proxy. Results show that the technique can be both effective in collecting the evidence required to solve the SMS-hijack, as well as practical in terms of SMSonPC app instrumentation and storage costs. The runtime overheads incurred were shown to be difficult to exploit by attackers to uncover an ongoing investigation, while at the same time not impacting the device owner.

This case study provided the right setting for initial exploration of the proposed SMS-hijack investigation process, with results showing promise. A similar case study for information leakage is planned. Further experimentation also aims to evaluate the technique at a larger scale using an array of physical smart-phone devices and possibly even involving malware samples captured in the wild. A primary pre-requisite for such an undertaking is the engineering of the investigation tool that also incorporates existing techniques that deal with obfuscated

```
{"active":true,"iden":"ujBeKPNHgJMsjz7aNoLJeK","created":1.5208663683633862E9,"modified":1.520866368366242E9,
"data":{"target_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","addresses":["123456"],"guid":"vfhj9v3t24o2q9544u0frg",
"message":"CrimeProxy sms text message 1"}}!
...snip ...
{"active":true,"iden":"ujBeKPNHgJMsjAsOdablfg","created":1.520866402296782E9,"modified":1.520866401232248E9,
"data":{"target_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","addresses":["123456"],"guid":"ctqvmagsveodhll4hldpv",
"message":"CrimeProxy sms text message 2"}}!
...snip ...
{"active":true,"iden":"ujBeKPNHgJMsjz2mR5H8yy","created":1.520866418990317E9,"modified":1.520866418994791E9,
"data":{"target_device_iden":"ujBeKPNHgJMsjAzp2VNDUW","addresses":["123456"],"guid":"9uv0upvsb1gdjch1qe3f1g",
"message":"CrimeProxy sms text message 3"}}!
...snip ...
```

Listing 5. Dalvik heap dump extracts.

```
123456'\n'CrimeProxy sms text message 1com.pushbullet.android
...snip ...
123456'\n'CrimeProxy sms text message 2com.pushbullet.androidV
...snip ...
123456'\n'CrimeProxy sms text message 3com.pushbullet.androidV
...snip ...
```

Listing 6. Native heap dump extracts.

code and anti-tamper checks. Even more importantly, collaboration with SMSonPC app developers is required to deal with app instrumentation in a cleaner way whenever this breaks functionality in some way. Collaboration is specifically sought on the anti-tampering front.

REFERENCES

[1] Y. Fratantonio, C. Qian, S. P. Chung, and W. Lee, "Cloak and Dagger: from two permissions to complete control of the UI feedback loop," in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017. doi: 10.1109/SP.2017.39 pp. 1041–1057.

[2] Y. Leguesse, C. Sidiropoulos, and L. Palkmets, *Mobile Threats Incident Handling (Part II)*. enisa, 2015.

[3] C. Anglano, "Forensic analysis of WhatsApp Messenger on Android smartphones," *Digital Investigation*, vol. 11, no. 3, pp. 201–213, 2014. doi: 10.1016/j.diin.2014.04.003

[4] J. Sylve, A. Case, L. Marziale, and G. G. Richard, "Acquisition and analysis of volatile memory from Android devices," *Digital Investigation*, vol. 8, no. 3, pp. 175–184, 2012. doi: 10.1016/j.diin.2011.10.003

[5] A. Gargenta, "Deep dive into Android IPC/Binder framework," in *AnDevCon: The Android Developer Conference*, 2012.

[6] A. Singh and A. Bhardwaj, "Android internals and telephony," *Int. J. Emerg. Technol. Adv. Eng*, vol. 4, pp. 51–59, 2014.

[7] M. H. Ligh, A. Case, J. Levy, and A. Walters, *The art of memory forensics: detecting malware and threats in Windows, Linux, and Mac memory*. John Wiley & Sons, 2014.

[8] L. Weichselbaum, M. Neugschwandtner, M. Lindorfer, Y. Fratantonio, V. van der Veen, and C. Platzer, "Andrubis: Android malware under the magnifying glass," *Vienna University of Technology, Tech. Rep. TR-ISECLAB-0414-001*, 2014.

[9] H. Ye, S. Cheng, L. Zhang, and F. Jiang, "Droidfuzzer: Fuzzing the android apps with intent-filter tag," in *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. ACM, 2013. doi: 10.1145/2536853.2536881 p. 68.

[10] S. Pooryousef and M. Amini, "Enhancing accuracy of Android malware detection using intent instrumentation." in *ICISSP*, 2017. doi: https://doi.org/10.5220/0006195803800388 pp. 380–388.

[11] S. Mutti, Y. Fratantonio, A. Bianchi, L. Invernizzi, J. Corbetta, D. Kirat, C. Kruegel, and G. Vigna, "BareDroid: Large-scale analysis of Android apps on real devices," in *Proceedings of the 31st Annual Computer Security Applications Conference*. ACM, 2015. doi: 10.1145/2818000.2818036 pp. 71–80.

[12] J. Li, D. Gu, and Y. Luo, "Android malware forensics: Reconstruction of malicious events," in *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on*. IEEE, 2012. doi: 10.1109/ICDCSW.2012.33 pp. 552–558.

[13] M. I. Husain and R. Sridhar, "iForensics: forensic analysis of instant messaging on smart phones," in *International Conference on Digital Forensics and Cyber Crime*. Springer, 2009. doi: 10.1007/978-3-642-11534-9-2 pp. 9–18.

[14] I. Murynets and R. Piqueras Jover, "Crime scene investigation: SMS spam data analysis," in *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM, 2012. doi: 10.1145/2398776.2398822 pp. 441–452.

[15] C. M. Steel, "Technical soddi defenses: The Trojan Horse defense revisited," *The Journal of Digital Forensics, Security and Law: JDFSL*, vol. 9, no. 4, p. 49, 2014.

2nd Workshop on Internet of Things—Enablers, Challenges and Applications

THE Internet of Things is a technology which is rapidly emerging the world. IoT applications include: smart city initiatives, wearable devices aimed to real-time health monitoring, smart homes and buildings, smart vehicles, environment monitoring, intelligent border protection, logistics support. The Internet of Things is a paradigm that assumes a pervasive presence in the environment of many smart things, including sensors, actuators, embedded systems and other similar devices. Widespread connectivity, getting cheaper smart devices and a great demand for data, testify to that the IoT will continue to grow by leaps and bounds. The business models of various industries are being redesigned on basis of the IoT paradigm. But the successful deployment of the IoT is conditioned by the progress in solving many problems. These issues are as the following:

- The integration of heterogeneous sensors and systems with different technologies taking account environmental constraints, and data confidentiality levels;
- Big challenges on information management for the applications of IoT in different fields (trustworthiness, provenance, privacy);
- Security challenges related to co-existence and interconnection of many IoT networks;
- Challenges related to reliability and dependability, especially when the IoT becomes the mission critical component;
- Zero-configuration or other convenient approaches to simplify the deployment and configuration of IoT and self-healing of IoT networks;
- Knowledge discovery, especially semantic and syntactical discovering of the information from data provided by IoT;

The IoT conference is seeking original, high quality research papers related to such topics. The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. The focus areas will be, but not limited to, the challenges on networking and information management, security and ensuring privacy, logistics, situation awareness, and medical care.

TOPICS

The IoT conference is seeking original, high quality research papers related to following topics:

- Future communication technologies (Future Internet; Wireless Sensor Networks; Web-services, 5G, 4G, LTE, LTE-Advanced; WLAN, WPAN; Small cell Networks. . .) for IoT,
- Intelligent Internet Communication,
- IoT Standards,
- Networking Technologies for IoT,
- Protocols and Algorithms for IoT,
- Self-Organization and Self-Healing of IoT Networks,
- Trust, Identity Management and Object Recognition,
- Object Naming, Security and Privacy in the IoT Environment,
- Security Issues of IoT,
- Integration of Heterogeneous Networks, Sensors and Systems,
- Context Modeling, Reasoning and Context-aware Computing,
- Fault-Tolerant Networking for Content Dissemination,
- Architecture Design, Interoperability and Technologies,
- Data or Power Management for IoT,
- Fog—Cloud Interactions and Enabling Protocols,
- Reliability and Dependability of mission critical IoT,
- Unmanned-Aerial-Vehicles (UAV) Platforms, Swarms and Networking,
- Data Analytics for IoT,
- Artificial Intelligence and IoT,
- Applications of IoT (Healthcare, Military, Logistics, Supply Chains, Agriculture, ...),
- E-commerce and IoT.

The conference will also solicit papers about current implementation efforts, research results, as well as position statements from industry and academia regarding applications of IoT. Focus areas will be, but not limited to above mentioned topics.

EVENT CHAIRS

- **Cao, Ning,** College of Information Engineering, Qingdao Binhai University
- **Furtak, Janusz,** Military University of Technology, Poland
- **Zieliński, Zbigniew,** Military University of Technology, Poland

PROGRAM COMMITTEE

- **Amanowicz, Marek,** Military University of Technology
- **Antkiewicz, Ryszard,** Military University of Technology, Poland
- **Chudzikiewicz, Jan,** Military University of Technology in Warsaw, Poland
- **Cui, Huanqing,** Shandong University of Science and Technology, China

- **Ding, Jianrui,** Harbin Institute of Technology, China
- **Fouchal, Hacene,** University of Reims Champagne-Ardenne, France
- **Fuchs, Christoph,** Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE, Germany
- **Gluhak, Alexander,** Intel Labs Europe
- **Hodoň, Michal,** University of Žilina, Slovakia
- **Johnsen, Frank Trethan,** Norwegian Defence Research Establishment (FFI), Norway
- **Krco, Srdjan,** DunavNET
- **Lenk, Peter,** NATO Communications and Information Agency, Other
- **Li, Guofu,** University of Shanghai for Science and Technology, China
- **Marks, Michał,** NASK - Research and Academic Computer Network, Poland
- **MURAWSKI, Krzysztof,** Military University of Technology, Poland
- **Niewiadomska-Szynkiewicz, Ewa,** Research and Academic Computer Network (NASK), Institute of Control and Computation Engineering, Warsaw University of Technology
- **Paprzycki, Marcin,** Systems Research Institute Polish Academy of Sciences, Poland
- **Sikora, Andrzej,** Research and Academic Computer Network (NASK)
- **Skarmeta, Antonio,** University of Murcia
- **Suri, Niranjan,** Institute of Human and Machine Cognition
- **Te Paske, Bert Jan,** TNO Netherlands
- **Wrona, Konrad,** NATO Communications and Information Agency
- **Xu, Jian,** Northeastern University, China
- **Zhang, Tengfei,** Nanjing University of Post and Telecommunication, China
- **Zhao, Yongbin,** Shijiazhuang Tiedao University, China

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 619–622
DOI: 10.15439/2018F301
ISSN 2300-5963 ACSIS, Vol. 15

Novel Solutions for Smart Cities – Creating Air Pollution Maps Based on Intelligent Sensors

Marzena Banach
Institute of Architecture and Spatial Planning,
Poznań University of Technology,
Nieszawska 13C, 61-021 Poznań, Poland,
E-mail: marzena.banach@erba.com.pl

Tomasz Talaśka and Rafał Długosz
UTP University of Science and Technology,
Faculty of Telecommunication, Computer Science and Electrical Engineering,
ul. Kaliskiego 7, 85-796, Bydgoszcz, Poland,
E-mail: talaska@utp.edu.pl, rafal.dlugosz@gmail.com

***Abstract*—The paper presents novel solutions for systems used to create air pollutions maps in smart cities. Ability to record pollution levels with a function of a short-term prediction of their fluctuations may be useful for cyclists and pedestrians moving through the city. Based on such data they can choose their route through the city in such a way, as to avoid the most polluted areas. Systems of this type are in the range of solutions characteristic for smart cities. Their effectiveness requires a relatively dense wireless sensor network (WSN) composed of miniaturized and cheap intelligent pollution sensors, capable not only of data recording and transmitting, but also of some data processing with the prediction abilities. Sensors of this type require a development of various circuit components that feature small sizes and ultra-low energy consumption. One of the main blocks, in this case, should be an artificial neural network (ANN) implemented at the transistor level. In this work, we present prototype circuits designed by us for the described purposes. The realized blocks include a finite impulse response (FIR) filter, programmable analog-to-digital converters (ADCs) with internal controlling clock generators and main building blocks of a parallel ANN. The specialized chips (ASIC – application specific integrated circuit) with the described components were implemented in the CMOS technology in the full custom style.**

I. Introduction

CITIES are not homogeneous systems. Depending on the area, different may be the terrain as well as urban occupation. The type of buildings includes such parameters as its density and height, width, and location of streets in relation to wind directions, distribution of green areas, etc.

Particular areas of the cities play a different function, which often translates into the intensity of traffic, which is one of the main factors causing pollutions. The described factors affect the natural possibilities of ventilation and the absorption of pollutants, and thus the susceptibility of particular areas of the city to changes in pollution levels. An important factor here is the time of persistence of specific levels of pollutions, even after the expiry of stimuli affecting them.

Air pollution is also affected by the seasons. This is due not only to the fact of a larger emission due to heating but also to natural changes in green areas. The biologically active surface (lawns, parks, green walls, green roofs, trees) has a significant impact on the absorption of pollutants.

In such cities as Kraków in Poland, the terrain does not support the natural ventilation, so the pollution levels are frequently at high levels. On the other hand, when designing new cities, this factor may be taken into account. An example here is the Zenata Eco-city developed in Morocco near the city of Casablanca. One of the examples of the smart approach at the city's design stage is taking into account natural conditions. For example, the wind directions in the area were examined in order to build the city in such a way, to allow for natural ventilation and lowering the temperature by several degrees in the summer [18]. It is also planned to collect rainwater in retention reservoirs and use it to maintain green areas – in plans, 30 % of the area of the city will be covered by greenery (compared to 3 % in Casablanca).

With such a diversity of conditions, relying only on a few or a dozen stations measuring levels of pollutions in a given area is not sufficient. At present, in main cities in Poland, on average, there are only a few to a dozen or so automatic stations measuring the pollution levels [1], [2].

The price of currently offered devices and their sizes do not support dense maps. To solve this problem, it is necessary to significantly reduce the price of a single device. The solution here may be miniaturized integrated sensors realized in the CMOS technology, implemented either as a System-on-Chip or as a System-in-Package. As shown in the next section, examples of the implementation of integrated pollution sensors can be already found in the literature. This is an important step towards advanced small and cheap measuring devices.

Another need is to increase the prediction abilities offered by the monitoring systems, especially for the short time horizon. Such information may be of fundamental importance, for example for cyclists and pedestrians moving around the city. Based on such data, they may consciously choose particular sections of their route through the city. The next step in the development of the monitoring systems may be the implementation of more complex measurement devices that offer an onboard ability to make predictions, fully independently from the main station.

The paper has the following structure. In the next Section, a state-of-the-art study is presented. Pollution monitoring systems cover many different elements that need to be presented separately. This part first discusses the monitoring systems considered as a whole. Subsequently, details relevant to the implementation of the integrated measuring devices are described. Then examples of the use of ANNs in the pollution

prediction problem are being presented. In the following Section, the authors' contribution to the development of integrated intelligent sensors is described, with such circuit components of such devices, as Analog-to-Digital Converters (ADCs), filters and components of the ANN. In the last Section, the conclusions are drawn.

II. State-of-the art study

A. Air pollution monitoring systems

In Poland, Voivodship Environmental Protection Inspectorates provide data from environmental monitoring for particular voivodships [3]. The offered system is based mainly on networks of measurement stations located in sensitive points of voivodships – mainly in large cities. Such stations measure concentrations of, among others, such gases and substances as sulfur dioxide, nitrogen oxides, benzene, carbon monoxide, ozone, suspended dust PM_{10} and $PM_{2.5}$. In 2017, in Poland, air quality measurements were carried out by 1,924 measuring stations, including 1,098 automated stations (57 %).

Airly company develops its own system, which is supposed to be much denser. Currently, the highest density of measuring devices is offered in Kraków. The measurement results offered to the public throughout the company's webpage show, this is justified, as concentrations can change significantly in short periods of time, while large differences may be visible even between areas located close to each other. Selected results from the Airly website, are shown in Figure 1. Diagrams (a) and (b), show results for two following hours (one Saturday in May) for the overall city.

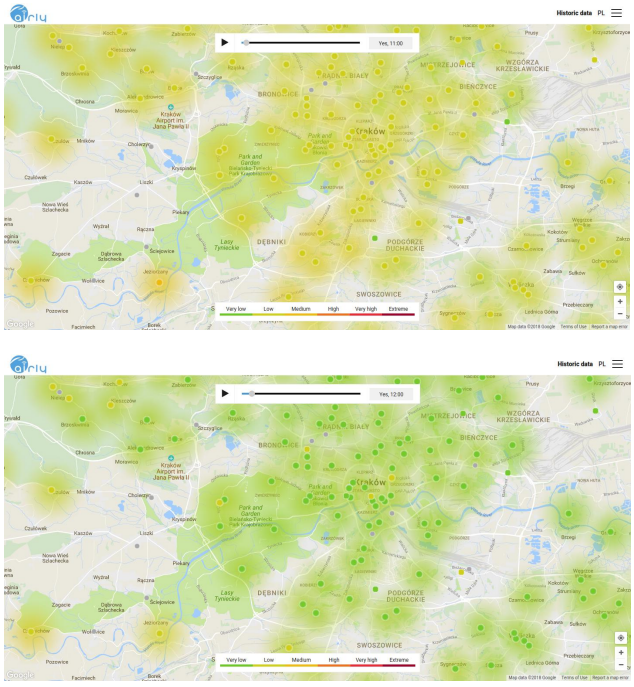


Fig. 1. Selected maps obtained from the Airly company's webpage (source: Airly sp. z o.o., Aleja Pokoju 1a, 31-548, Kraków, Poland, https://map.airly.eu/pl/) for the overall city for two following hours.

At this point, it is worth mentioning wearable pollution sensors already offered on the market. Their usage is limited, as they require a person equipped with such a sensor to be present in a given area of the city. It does not protect this person from the effects of the pollutions. A better solution would be stationary sensors providing data to a central computation station, which would be able to perform a short-term prediction based on the collected data. In this way, city users could plan their route or the mean of transport more consciously in advance. One of the useful options related to the described wearable sensors would be sharing collected data with a central system so that other people could use. Such an approach could support the stationary system, increasing the prediction abilities and the map resolution.

B. Microelectronic sensors of pollution particles

In the literature, one can find several examples of the implementation of particle sensors of various air contaminations. Texas Instruments company proposed an optoelectronic system for the detection and measurement of the $\mathrm{PM}_{2.5}$ and the PM_{10} particles. The system is based on the detection of scattered light by particles suspended in the air [14]. In [15] presented are methods of implementing the $\mathrm{PM}_{2.5}$ and PM_{10} particle micrometer using the zinc oxide based on the Solidly Mounted Resonator (SMD). The authors of [15] present a complete system mounted on a printed board (PCB). On the board, apart from the particle sensor PM_{10} and $\mathrm{PM}_{2.5}$, there is also a specialized ASIC, used by the sensor. The system works correctly, however, it consists of two separate blocks – the sensor and the system managing it. This is a disadvantage, for it increases the dimensions of the device, and what is also important, in such a solution there is a lack of encapsulation of the electronic circuits, which is important taking into account sometimes hard environmental conditions.

Another example is an integrated capacitive sensor designed in the CMOS 0.35 μm technology [16], [17]. The advantage of this system is the fact that it is implemented inside the integrated circuit, which enables, apart from the detection of molecules, also other parallel operations. The system is capable of detecting microscopic particle sizes of pollutants. This work is an important novelty in this area in the world, as the sensor solutions inside the integrated circuit are presented for the first time here.

The availability of the described works is very important. The combination of such solutions with the ANN offered by the authors of this article is important from the point of view of the development of fully intelligent and autonomous sensors.

C. Using ANN to monitor and predict air pollutants

Development of smart cities would not be possible without the use of artificial intelligence (AI) algorithms, including artificial neural networks (ANN), genetic algorithms (AG), fuzzy systems (FS), or expert systems (SE). AI will play a significant role in the development and functioning of such cities. On one hand, this is due to the very large complexity of the system which the city is, and on the other hand, the

lack of the possibility of an accurate mathematical description of the interdependencies between various parameters of this system. Complexity means also a huge amount of data, which is often difficult to express unequivocally with the help of unified indicators.

ANNs are universal tools frequently used in different areas of daily life. They are frequently employed for data classification, prediction, recognition, detection, etc. Neural networks operate more effectively if they receive properly prepared data (e.g. normalized) and when they have the sufficiently large computing power to process them within an acceptable time interval. It is also necessary to match a given type of the ANN to a specific problem. One of the areas in which these tools are more and more frequently used is the problem of air pollution, described in this paper. In this case, they are capable of contamination forecasting [11], [12], [6], [5], [13], [10], [7], [4].

One of the examples here, is the application of the Multi-Layer Perceptron (MLP) and the Radial Basis Function (RBF) neural networks for a long-term prediction of pollutants dust (PM_{10}, $PM_{2.5}$) (especially in cities) [11], [12]. In another work [6] a multilayer perceptron (MLP) has been used to forecast the impact of pollution caused by road traffic (exhaust fumes) on the health of residents. In [13] three different machine learning algorithms are presented, including, among others, a neural network for the monitoring and forecasting of pollutants, such as ground-level ozone O_3, nitrogen dioxide NO_2 and dioxide sulfur SO_2. Another work in this area [10] presents the concept of self-organizing networks [7] used for the classification of data on sulfur dioxide hazards SO_2 is presented.

III. Proposed Contribution to the Development of Intelligent Pollution Sensors

In this Section, we present a model of an integrated system for the detection and prediction of air pollution particles in the form of a specialized chip implemented in the CMOS technology. State-of-the art works in this area, described above, focus on the implementation of either the pollution particles sensors in the form of an integrated circuit or the use of software implementation of neural networks to detect impurities. In this work, we present the solutions that will allow joining those two areas into a single integrated device, capable of working at WSN.

The proposed intelligent sensor consists of several main components, such as: (i) air pollution sensor (a standard solution) used to collect data from the city, (ii) a filter used to remove noises from the signal, (iii) an ADC used to convert measured analog data into a digital signal further processed in, (iv) data processing unit with a hardware realized ANN. The processed data are transferred to the base station using (v) an RF communication block (standard block to be used).

The neural network, that can be used in the sensor, may be either analog or digital. In the second case, a preliminary analog-to-digital conversion with the appropriate resolution is required. Even though, the preferred solution is the use of a digital neural network, due to its robustness to various types of external conditions and the impact of the imperfections of the technological process. Both filters and ADCs were previously designed by us [19]. A prototype ASIC containing these components is shown in Figure 3.

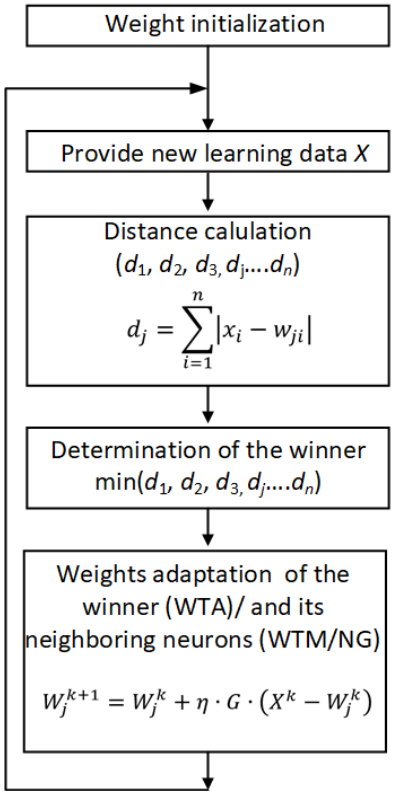


Fig. 2. Self-organizing neural network algorithm.

In this work, we focus on ANN, that will be integrated directly with other components of the sensor. In the proposed approach we use self-organizing networks since such networks are already used with success for the analysis and prediction of the air pollutions. Moreover, such networks offer a relatively simple structure, which is crucial from the point of view of the implementation in the ASIC. Low mathematical complexity allows for a large miniaturization. The structure of such a network is schematically shown in Figure 2. The basic operations performed for each learning pattern X are as follows:

- initialization of the neuron weights,
- providing new learning pattern X from the ADC and indirectly from the pollution sensor,
- calculation of distances (d_1, d_2, d_3, d_j, ..., d_n) between the learning pattern X and weight vectors W of particular neurons. One of the typical distance measures may be used, namely Manhattan or Euclidean one [8],
- determination of the winning neuron. Mathematically it is the min(d_1, d_2, d_3, d_j, ... ,d_n) operation,
- determination of the neighborhood of the winning neuron

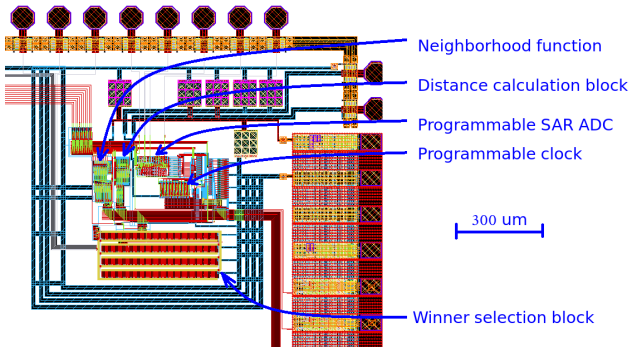


Fig. 3. A prototype ASIC designed by us for the application in the intelligent sensors, containing components of the ANN and a programmable SAR ADC, controlled by internal programmable clock generator.

and the calculation of the learning rate η for particular neighbors,

- adaptation of the winning neuron (in WTA only this neuron is adapted) and its neighbors (in WTM and NG). The adaptation is realized according to a formula:

$$W_j(k+1) = W_k(l) + \eta(k) \cdot G() \cdot [X(k) - W_j(k)] \quad (1)$$

In this formula W_k is the weights vector of a j^{th} neuron. The neurons that belong to the winner's neighborhood, are trained with the intensities determined by the applied neighborhood function $G()$ [8], [9].

The prototype chip, shown in Fig. 3, contains most of the components of the proposed intelligent sensor, including particular building blocks of the self-organizing neural network described earlier. Particular subcircuits of this chip are responsible for the computations performed at particular stages of the learning algorithm of the NN, as described above. The chip also contains an ultra-low power, programmable, 10-bits ADC. All these components were tested by means of laboratory measurements.

IV. Conclusion

The paper presents the concept of a system for the monitoring of air pollution in smart cities. The proposed concept is based on the usage of a dense network of miniature intelligent pollution sensors. The assumption is to build cheap sensors, self-sufficient in energy so that they can be easily deployed on city streets without having access to the power line.

The implementation of such sensors requires the development of particular building components. A special attention should be paid to the sizes of these blocks and their computing power while maintaining their functionality in the comparison with similar typical software solutions.

For this reason, we presented selected hardware solutions designed by us, with particular emphasis on miniature artificial neural networks capable of working in parallel.

In this work, we present intermediate results. The next step will be to assemble the designed blocks into a larger system. It is worth to add, that the proposed solutions can cooperate not only with the pollution sensors. Their usage is universal. A modular structure of the resulting device may be considered, in which the sensor would be an element selected depending on the needs, while the signal processing scheme would be similar.

References

[1] https://map.airly.eu/pl/, (access 2018.09.10).
[2] http://aqicn.org/map/world/, (access 2018.09.10).
[3] http://sojp.wios.warszawa.pl/, (access 2018.09.10).
[4] Moustris K. P., Larissi I. K., Nastos P. T., Koukouletsos K. V., Paliatsos A. G., "Development and Application of Artificial Neural Network Modeling in Forecasting PM10 Levels in a Mediterranean City," *Water Air Soil Pollut*, DOI 10.1007/s11270-013-1634-x, 224:1634, 2013
[5] Sahina U.A., Bayatb C., Uçan O.N., "Application of cellular neural network (CNN) to the prediction of missing air pollutant data," *Atmospheric Research*, Elsevier, Vol.101, s.314-326, 2011.
[6] Fontes T., Silva L.M., Pereira S.R., Coelho M.C., "Application of Artificial Neural Networks to Predict the Impact of Traffic Emissions on Human Health. Progress in Artificial Intelligence," *Springer, Berlin, Heidelberg*, Lecture Notes in Computer Science, Vol.8154, s.21-29, 2013.
[7] Kohonen T., "Self-Organizing Maps (Information Sciences)," trzecie wydanie, Nowy Jork (USA), Springer-Verlag, 2001.
[8] Długosz R., Kolasa M., Pedrycz W., Szulc M., "Parallel programmable asynchronous neighborhood mechanism for Kohonen SOM implemented in CMOS technology," *IEEE Transactions Neural Networks*, Vol. 22, No. 12, s.2091–2104, December 2011.
[9] Talaśka T., Długosz R., "Analog, parallel, sorting circuit for the application in Neural Gas learning algorithm implemented in the CMOS technology," *Applied Mathematics and Computation*, Vol. 319, 2018
[10] J. M. Barrón-Adame and O. G. Ibarra-Manzano and A. Vega-Corona and M. G. Cortina-Januchs and D. Andina, "Air pollution data classification by SOM Neural Network," *World Automation Congress 2012*, pp.1-5, 2012.
[11] W. Kaminski and J. Skrzypski and E. Jach-Szakiel, "Application of Artificial Neural Networks (ANNs) to Predict Air Quality Classes in Big Cities," *International Conference on Systems Engineering*, pp.135-140, 2008.
[12] M. M. Dedovic and S. Avdakovic and I. Turkovic and N. Dautbasic and T. Konjic, "Forecasting PM10 concentrations using neural networks and system for improving air quality," *International Symposium on Telecommunications (BIHTEL)*, pp.1-6, 2016.
[13] K. Bashir Shaban and A. Kadri and E. Rezk, "Urban Air Pollution Monitoring System With Forecasting Models," *IEEE Sensors Journal*, vol. 16, no. 8, pp.2598-2606, 2016.
[14] "PM2.5 and PM10 Particle Sensor Analog Front-End for Air Quality Monitoring Reference Design," *Texas Instruments Incorporated*, http://www.ti.com/tool/TIDA-00378.
[15] S. Thomas and F. H. Villa-López and J. Theunis and J. Peters and M. Cole and J. W. Gardner, "Particle Sensor Using Solidly Mounted Resonators," *IEEE Sensors Journal*, vol. 16, no. 8, pp.2282-2289, 2016.
[16] P. Ciccarella and M. Carminati and M. Sampietro and G. Ferrari, "Multichannel 65 zF rms Resolution CMOS Monolithic Capacitive Sensor for Counting Single Micrometer-Sized Airborne Particles on Chip," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 11, pp.2545-2553, 2016.
[17] M. Carminati and P. Ciccarella and M. Sampietro and G. Ferrari, "Single-Chip CMOS Capacitive Sensor for Ubiquitous Dust Detection and Granulometry with Sub-micrometric Resolution," *Springer International Publishing AG 2018*, Sensors, Lecture Notes in Electrical Engineering, pp.8-18, 2018.
[18] Christin Roby, "An inside look into Africa's first eco-city: Zenata, Morocco," devex, 13 March 2017, https://www.devex.com/news/an-inside-look-into-africa-s-first-eco-city-zenata-morocco-89741
[19] Długosz R. and Fischer G. "Low chip area, low power dissipation, programmable, current mode, 10-bits, SAR ADC implemented in the CMOS 130nm technology," *International Conference Mixed Design of Integrated Circuits and Systems (MIXDES)*, Poland, 2015.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 623–627
DOI: 10.15439/2018F290
ISSN 2300-5963 ACSIS, Vol. 15

Raspberry Pi as an Inexpensive Platform for Real-Time Traffic Jam Analysis on the Road

Robert Baumgartl, Dirk Müller
Faculty of Informatics/Mathematics
Dresden University of Applied Sciences
Friedrich-List-Platz 1, D-01069 Dresden, Germany
Email: [firstname].[secondname]@htw-dresden.de

***Abstract*—Using mobile phones for accessing the Internet has become a standard use case of such devices, nowadays even more important than the good old phone call. WiFi at home or public ones allow for a low-cost or even unpaid access to the virtual world of the Internet. But, as we will show, this is only true to some degree in terms of monetary cost. One thing we're paying a lot with is the loss of our privacy. In this paper, we will show how easily and cheap potential attackers can track your mobile phone and, thus, you via data it sends all the time, so-called probe requests. Additionally we show by experimental data how this tracking can be used for traffic jam analysis on roads.**

***Index Terms*—Raspberry Pi Real-Time Traffic Security Privacy**

I. Introduction

DURING the last years, the use of mobile devices like mobile phones and tablets for accessing the Internet has celebrated a breakthrough due to technological advances and social changes. Mobile access has overtaken stationary one from desktop computers and statically used laptop computers.

As a result of this development, end users can access the huge library of the Internet from many places all over the world including situations when traveling by car, by train or even by airplane. From one point of view, this is great news since data can be retrieved easily, and this new level of information can be used for the good. On the other hand, data transmitted via the Internet often turns out to be spam, noise or just jokes. But at least we get the potential to do more useful things with the almost ubiquitous mobile Internet.

A matter of particular interest for huge market penetration is the pricing of the goods and services. A cost-efficient semi-mobile Internet access is typically provided by the use of a WiFi connection according to IEEE 802.11 standard. Its typical range is some $100\,\mathrm{m}$ in the field and $20\,\mathrm{m}$ or $30\,\mathrm{m}$ in buildings depending upon the material of the walls. Several mobile devices can be associated to one and the same access point in parallel, and all of them can be freely moved. Hence, in order to guarantee a stable connection, there needs to be a key for identifying every device. The established key having been used for enabling a target-oriented delivery of packets is the *Media Access Control* or *MAC* address.

A serious problem with the technically well-motivated MAC address approach is the significant decrease of privacy for the end user carrying a mobile device. A static 1-to-1-link between a device and an identifier perfectly allows at least for tracking, and, by some reverse or social engineering finally to uncover the identity of the person who carries a particular device. Combining both mechanisms by data merging ultimately allows for tracking everyone all over the world, a scenario completely violating all privacy requirements. Note that such a kind of tracking is by far not only an academic issue, but can happen and happens on a grand scale [1].

A heavily promoted counter-action of mobile device sellers fighting this privacy issue was *MAC address randomization* as implemented by major companies, cf. [2], starting from 2014. Unfortunately, recent publications [3] [4] show clearly that attacking privacy has only become a little bit more difficult, but by far not impossible as initially claimed by mobile devices companies.

In this article, we will show by some experiments how such a tracking can be performed with a little bit of knowledge and some inexpensive equipment. Finally, we will present and discuss the results of our most advanced setup for performing a traffic jam analysis via a so-called *section control*[1].

Here, individual cars' average velocities are calculated via the measurement of their time passing a fixed-length (some kilometers) section of a road. Of course, exceeding the speed limit in terms of the average speed implies also an illegal speeding in terms of peak speed whatever the actual velocity profile looks like. On the other hand, on the majority part [5] of the German autobahn, there is no speed limit at all. While the first *section control* was installed in Austria more than 10 years ago, there are only plans to apply it in Germany as well. In 2011 in Poland, an experimental section control on a $16\,\mathrm{km}$ section close to the city of Gdańsk revealed 28 drivers driving at average velocities of more than $200\,\mathrm{km\,h^{-1}}$ while $140\,\mathrm{km\,h^{-1}}$ was the allowed top speed [6]. Conventional *section control* is based on automatic number plate recognition. We discuss here an alternative mobile-device-based approach.

The remainder of this paper is structured as follows: In section II we discuss projects and publications related to our findings. Next, we give a short overview of our experimental hard- and software in section III. In section IV we describe a series of experiments of increasing complexity we conducted. We discuss the setup and summarize the main results. The

[1]This is actually a pseudo-anglicism like *handy* for a mobile phone or *beamer* for a video/digital projector. The term used in UK is SPECS for *Speed Check Services*, see also http://www.jenoptik.co.uk/product/specs.

paper ends with a summary of our main findings and an outlook to open questions.

II. Related Work

Vehicular traffic monitoring is a very popular field of research [7][8]. Conventional sensor technologies use inductive loop, piezoelectric, magnetometer, pressure switch, video camera, microwave radar, ultrasonic, optical, and laser radar data [8]. None of them will be used in our experiments. Instead, our data will be MAC addresses extracted from probe requests.

The general topic of tracking mobile devices and finally end users via their MAC addresses passively via probe requests is a common topic in the literature.

Many authors are aware of the privacy issue of the approach. Demir [9] proposed a multiple hashing of MAC addresses. Fuxjäger *et al.* [10] show that brute-force attacks on just hashed MAC addresses are quite simple, and, thus suggest a truncated and hashed MAC address approach with a higher level of privacy. Finally, Martin *et al.* [4] recently showed that even the more advanced technique of MAC address randomization can be attacked with a $100\,\%$ success ratio.

Chilipirea *et al.* [3] performed experiments on WiFi tracking of pedestrians. They could improve the quality of the data sets by various data filters.

Fuxjäger *et al.* [10] report on traffic jam analysis experiments on Austrian roads, but they used a more expensive equipment with external antennae.

A comprehensive study of WiFi probe requests for tracking and monitoring was given by Freudiger [11]. He managed to recognize several phone and OS types via profiling. But—compared to us—he used as well a more expensive monitoring equipment.

III. Experimental Platform

As cheap and ubiquitous hardware platform we used the Raspberry Pi Version 3 which offers an integrated WiFi chipset (Broadcom bcm43438). As mass storage medium we utilized cheap microSD cards of 32 GiB size. To ensure a maximum of autonomous operability, the systems were powered by external power-banks with a capacity of 20.000 mAh, which appeared to be somewhat over-sized. Hardware cost for one system amount to 50$. We utilized off-the-shelf Raspbian[2] Linux Version 8 as operating system base which provides a tailored Linux kernel version 4.4.50-v7+. Both systems were configured and used in headless mode.

The Raspbian standard firmware for the WiFi chip is not able to switch to monitor mode, therefore we installed the alternative firmware `nexmon`[3], version 7_45_41_26. The received data frames were captured using `dumpcap`, version 1.12.1, which is part of the well-known `wireshark` tool suite. By means of a capture filter, only probe requests were logged to persistent memory.

[2]http://www.raspbian.org

[3]https://github.com/seemoo-lab/nexmon

The resulting dumps were transferred to an external computer and converted to text records using `tcpdump`. Afterwards, we eliminated all irrelevant information except sender MAC addresses and accompanying timestamps within the measurement interval with the help of standard UNIX tools.

IV. Experiments

A. Receiving Probe Requests while Driving on the Autobahn

As a first attempt, we wanted to find out whether the Raspberry Pi is able to capture probe requests when moving fast. We placed the board under the windshield just like a dashcam and captured while driving.

On 04/11/2017, we entered the German autobahn A17 at access no. 3 "Dresden-Südvorstadt" at 15:45, headed for Dresden, changed to the A4 heading to Erfurt and left it at 16:33 at A4 exit no. 66 "Wüstenbrand". The distance was $83\,\mathrm{km}$ the average speed amounted to $104\,\mathrm{km\,h^{-1}}$.

During these 48 minutes, we captured 3379 MAC addresses, 609 of them were unique. It seemed that we were able to receive probe requests not only from cars driving in the same but also in the opposite direction, especially when both were using the leftmost lane.

This and the result of the next experiment were encouraging and proved that the board is very well capable of capturing a large number of probe requests while moving.

B. Receiving Probe Requests on a Train

In this experiment, we took the regional train *RE 26984* departing from *Dresden Hbf* to *Plauen(Vogtl) ob Bf* on 3rd May 2017. Only the section *Dresden Hbf* to *Chemnitz Hbf* corresponding to a scheduled travel from 15:52 to 16:54 was part of this experiment. Due to the recording of approximately one hour, we hoped for many probe requests with a lot of various MAC addresses.

We recorded as many as 6752 probe requests, i.e., on the average almost 2 per second. Among them, there could be 219 different sender MAC addresses of broadcast probe requests extracted. This number gives us a raw estimation of the order of magnitude of the number of travelers in this part of the train.

C. Receiving Probe Requests at the Road

Description: To receive probe requests from passing vehicles on a multi-lane highway, two principal positions could be used: a) on a bridge above the middle lane of one travel direction or b) by the right side of the road. Position a) seems favorable due to its elevation (and probably better receiving conditions) but requires constructions which cross the highway such as bridges. Position b) seems better suited in terms of cost and convenience (the system could easily be attached to some post or crash barriers).

In contrast to the scenario described in section IV-A we statically positioned the receiver a) on a bridge three meters above the middle lane (it is the same as measurement point B in section IV-D) and b) ten meters to the right of the rightmost

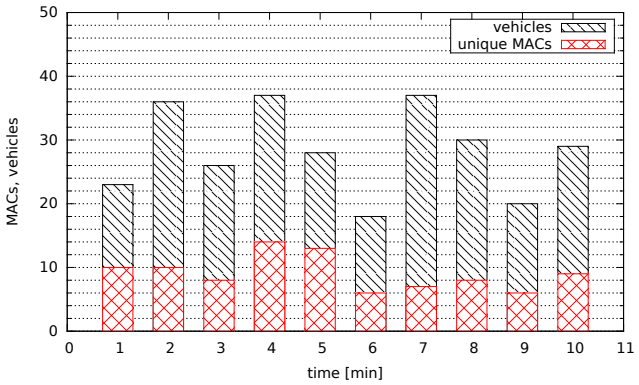


Fig. 1. Capturing on a bridge

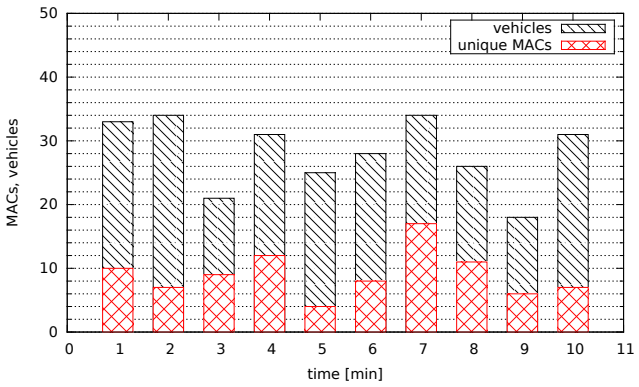


Fig. 2. Capturing at the side of the road

lane of the German autobahn A4[4] at a height of $1.50\,\mathrm{m}$ above ground (the GPS position was N 50.821650°, E 012.761600°). Due to restriction fences we were not able to get any closer to the lane.

Results: During the 10 minute capture interval, a total of 284 vehicles passed the position on the bridge and 281 vehicles were observed at the side of the road. During that interval, we were able to capture 91 unique MAC addresses resulting from probe requests at both positions which is almost a third. Figures 1 and 2 compare the number of cars and received unique MAC addresses on a minute-per-minute basis for both positions.

The ratio between passing cars and unique MAC addresses varies from $16\,\%$ to $50\,\%$ (both extrema were observed at the side of the road) with an average of $32\,\%$. Variance seems also a bit higher when capturing for the position at the side of the road but the short measurement interval prohibits deeper analysis.

Of course, there is no 1:1 relation between vehicles and MAC addresses for several reasons. First, some drivers may have switched off the WiFi functionality or could even have no smartphone at all. Second, other vehicles could carry more than one smart device, especially all kind of buses. Receiving more than one MAC from the same vehicle is redundant when trying to estimate vehicle speed (cf. section IV-D), but increases the chance of receiving two probe requests at different locations. Third, received MAC addresses from outside the context (passing bicyclists, vehicles from the opposite lane) could deteriorate our perceived numbers.

Nevertheless we can conclude, that a reasonable fraction of the passing vehicles sends probe requests such that our receiver hardware is able to capture them. Further, both logging positions seemed equally suitable.

D. Estimating Vehicle Speed

Description: In the final experiment, we tried to measure (or at least estimate) the average velocity of vehicles cruising in one direction for a certain section of the German autobahn. To this aim, we positioned two Raspberry Pis at a height of 3 to 4 meters above the middle lane of the A4 in direction of traffic Erfurt on two crossing bridges (The GPS coordinates are N 50.833591°, E 012.792370° for Point A, and N 50.819305°, E 012.745936° for Point B, respectively). Between A and B, the track runs almost straight. Figure 3 depicts the relevant topographical aspects. The distance between both points amounts to $5.03\,\mathrm{km}$ according to *openrouteservice.org*.

One motorway access is located between A and B therefore the numbers of passing vehicles may not be identical for both positions. During the time of our experiments, no explicit speed limit was mandated, visibility was very good.

Beforehand, the system clocks were synchronized manually with a Δ of ca. one second. Both systems logged all received probe requests for a fixed time interval of 15 minutes starting at 17:12 on 09/05/2017, a normal workday. Additionally, we manually recorded the number of passing vehicles per minute.

Because most MAC addresses were broadcast in short bursts we eliminated all but the first occurrence of a new unique MAC address. Then we searched for MAC addresses occurring in both log files (at different times) representing one and the same vehicle passing sequentially both measurement positions. We then determined the temporal difference t of the respective time stamps rounded to full seconds. Using the equation $v = s/t$ and the driving distance $s = 5.03\,\mathrm{km}$ between both points A and B, we finally computed the average speed of the vehicles.

Results: During the measurement interval of 15 minutes, a total of 453 vehicles passed point A. During that time, we observed a total of 115 unique MAC addresses. That 25 percent fraction seems to be somewhat optimistic, because a certain number of probe requests might also result from the opposite driving direction (see below). Figure 4 illustrates the number of passing cars and the number of received unique MAC addresses on a minute-per-minute basis. Nevertheless, we consider the number of unique MAC addresses surprisingly high given the cheap hardware platform and the high velocity of the passing vehicles which results in a visibility interval of a few seconds at the most.

[4]It is part of the longest European route, E40 from France to Kazakhstan. The majority of the A4 in Germany is a 3-lane-per-direction highway, including the part considered here.

Fig. 3. Measurement points for the estimation of vehicle speed

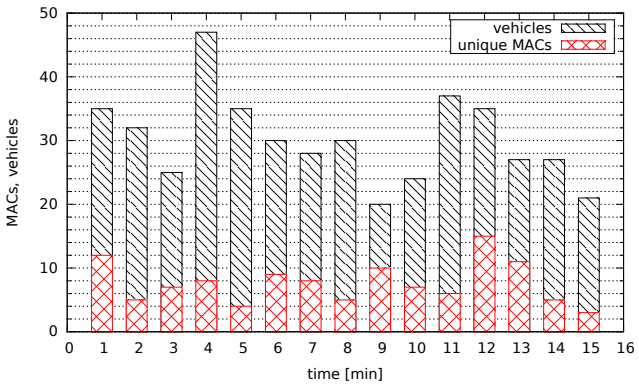


Fig. 4. Numbers of vehicles and unique MAC addresses per minute

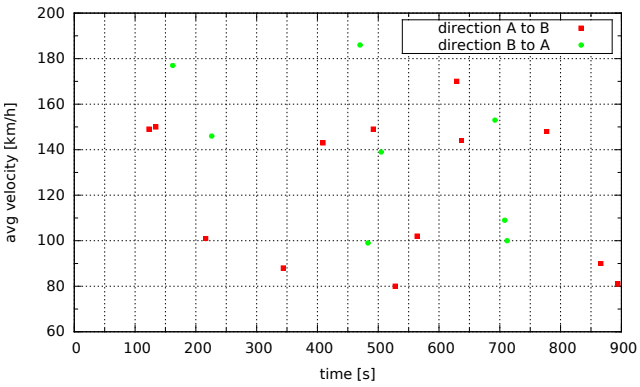


Fig. 5. Vehicle velocities measured over an interval of 15 minutes

Out of 115 unique MAC addresses recorded at point A and 128 unique MAC addresses recorded at B, we obtained 21 MAC addresses occurring at *both* positions. The slightly higher number of addresses at point B could result from a $1\,\mathrm{m}$ to $1.5\,\mathrm{m}$ lower position relative to the road surface. Figure 5 illustrates the individual velocities during our 15 minutes measurement interval.

Two of the address pairs had exactly the same time difference (121 seconds). Obviously, we monitored two different smart devices residing in the same car and sending their probe requests at the same instant. Further, it is interesting to note that we monitored 8 out of the 21 address pairs stemming from vehicles driving in the opposite direction. Third, a clear distinction between slow-driving trucks ($v < 120\,\mathrm{km\,h^{-1}}$) and faster passenger cars ($v > 140\,\mathrm{km\,h^{-1}}$) can be made. This reflects our empirical perception of the traffic situation and very good driving conditions. All obtained velocity values are plausible.

We can conclude that our setup allows to reliably estimate average velocity of vehicles for a given section on the autobahn and one driving direction. Further, it seems effortlessly possible to cover both driving directions by placing the Raspberries above the middle of the highway.

V. Conclusion and Outlook

Our considerations and experiments have shown that a tracking of mobile devices based on MAC addresses is feasible even with low-end equipment like a *Raspberry Pi Version 3* without any external antennae. Recently added features of modern smartphones like MAC address randomization render a tracking of such devices more difficult. But there is still a high market share of older mobile phones and such ones where the feature is not (yet?) implemented due to compatibility issues, cf. [4].

Use cases of such a tracking could be traffic jam analysis on a road. Here, we don't need to track many cars, some representative data is perfectly sufficient. Such real-time data can be used immediately for congestion alerts in navigation systems or in the good old FM radio.

Second, a preliminary assessment of road sections in terms of the percentage of speeding cases and their respective severity can be done. Our setting is very inexpensive and works anonymously. The recorded MAC addresses serve only as a matcher between two measurement points and can even be deleted directly after raw data processing. Our configuration is much cheaper than a conventional section control system and, thus, can be used more extensively for finding out where to place speed control points or sections.

Our results raise some interesting questions. First of all, we want to provide bounds for the accuracy of our velocity values. To that aim, two influence factors must be studied: one has to know the receiving range of the built-in antenna of the Raspberry Pi which depends on the position of the system itself and the period and sending pattern of probe requests has to be known which depends, among other, from the particular device type, its operating system version and operating mode.

We think that our approach is robust against heterogeneous and lane-less vehicular traffic being typical for countries like India due to lacking lane discipline or lacking lane infrastructure [8]. But this hypothesis needs to be verified by some additional experiments.

Further, because for our use case the Raspberry Pi has quite some unnecessary peripheral components, it seems promising to try even cheaper platforms, such as the Raspberry Pi Zero or the Espressif ESP8266 and ESP32. To drastically increase the duration of our measurements or even let the platform work autonomously, some means for solar powering the Pi should be investigated.

Finally, cost of an individual system approaching five dollars or even less would permit to equip a longer section of the highway with systems who are able to connect to each other via WiFi and propagate traffic data accordingly. Such a *smart highway* will probably lead to new interesting use cases.

References

[1] S. Khandelwal, "Spying agencies tracking your location by capturing mac address of your devices," *The Hacker News - Security in a serious way*, 2014, http://thehackernews.com/2014/01/spying-agencies-tracking-your-location_31.html.

[2] A. Mamiit, "Apple implements random mac address on ios 8. goodbye, marketers," *Tech Times*, 2014, http://www.techtimes.com/articles/8233/20140612/apple-implements-random-mac-address-on-ios-8-goodbye-marketers.htm.

[3] C. Chilipirea, A. C. Petre, C. Dobre, and M. van Steen, "Presumably simple: Monitoring crowds using wifi," in *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, vol. 1, June 2016, pp. 220–225.

[4] J. Martin, T. Mayberry, C. Donahue, L. Foppe, L. Brown, C. Riggins, E. C. Rye, and D. Brown, "A study of MAC address randomization in mobile devices and when it fails," *CoRR*, vol. abs/1703.02874, 2017. [Online]. Available: http://arxiv.org/abs/1703.02874

[5] S. Anker, "Freie Fahrt – Wo geht das noch in Deutschland?" *PS-Welt*, 2013, https://www.welt.de/motor/article121455433/Freie-Fahrt-Wo-geht-das-noch-in-Deutschland.html.

[6] WP moto, "Odcinkowy pomiar prędkości na a1 działał tydzień," *WP moto*, 2011, http://moto.wp.pl/odcinkowy-pomiar-predkosci-na-a1-dzialal-tydzien-6068745844532353a.

[7] P. Bellavista, F. Caselli, A. Corradi, and L. Foschini, "Cooperative Vehicular Traffic Monitoring in Realistic Low Penetration Scenarios: The COLOMBO Experience," *Sensors (Basel, Switzerland)*, vol. 18, no. 3, March 2018. [Online]. Available: http://europepmc.org/articles/PMC5876597

[8] N. K. Singh, L. Vanajakashi, and A. K. Tangirala, "Segmentation of vehicle signatures from inductive loop detector (ILD) data for real-time traffic monitoring," in *2018 10th International Conference on Communication Systems Networks (COMSNETS)*, Jan 2018, pp. 601–606.

[9] L. Demir, "Wi-Fi tracking : what about privacy," Master's thesis, M2 SCCI Security, Cryptologyand Coding of Information - UFR IMAG, Sep. 2013. [Online]. Available: https://hal.inria.fr/hal-00859013

[10] P. Fuxjaeger, S. Ruehrup, T. Paulin, and B. Rainer, "Towards privacy-preserving wi-fi monitoring for road traffic analysis," *IEEE Intelligent Transportation Systems Magazine*, vol. 8, no. 3, pp. 63–74, 2016.

[11] J. Freudiger, "How talkative is your mobile device?: An experimental study of wi-fi probe requests," in *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '15. New York, NY, USA: ACM, 2015, pp. 8:1–8:6. [Online]. Available: http://doi.acm.org/10.1145/2766498.2766517

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 629–636
DOI: 10.15439/2018F68
ISSN 2300-5963 ACSIS, Vol. 15

Rapid Embedded Systems Prototyping – an effective approach to embedded systems development

Robert Brzoza-Woch, Łukasz Gurdek, Tomasz Szydło
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Krakow, Poland
Email: robert.brzoza@agh.edu.pl

***Abstract*—In this paper we introduce the Rapid Embedded Systems Prototyping (RESP) approach aimed at accelerating the development of novel, experimental, and proof-of-concept implementations of embedded devices based on microcontrollers and Field Programmable Gate Array (FPGA) chips. It is intended to be used in the fast-paced business environment in which an early working prototype is required. The RESP approach can be useful for remote developing and temporary monitoring of various embedded devices: primarily for resource-constrained IoT platforms, microcontroller-based sensor nodes, and customized ad hoc systems. The RESP-compliant system uses a central server and one or multiple Remote Reconfiguration and Monitoring (RRM) modules. Each RRM allows the software developers to manage reprogramming and monitoring of multiple target embedded devices. It can be applied to a device that needs to be remotely reconfigured, tested, or reprogrammed in its target environment without implementing a reliable bootloader. The RRM described in this paper has been successfully implemented and its functionality and performance have been tested.**

I. Introduction

IN COURSE of research projects and at an early stage of embedded systems development cycle there is often a need to quickly develop a working prototype of an embedded device which will operate in its target environment as an IoT node or as an innovative solution for control, acquisition or monitoring purposes. In the start-up business environment, a client or an investor may wish to see a demonstration of a working proof-of-concept prototype of an application-specific embedded system. In a traditional approach, an embedded device is handed to a customer after the device's software is developed to a point in which it has full functionality and, usually, its bootloader program is working to update the device's program memory or its configuration. Developing a fully functional embedded device with a reliable operating bootloader may be a tedious and time-consuming task.

When utilizing an adaptive and evolutionary approach to software development, the functionality of the final product may not be precisely specified at an early development stage. It can be beneficial to start developing the application-specific embedded hardware and software for demonstration purposes and then to continue the product development when initial results are evaluated in practice and client's expectations are more specific. In this article we discuss various aspects of this strategy which we call the Rapid Embedded Systems Prototyping (RESP). The approach is based on fast prototyping iterations with an ability to remotely reconfigure or reprogram a hardware platform using the Remote Reconfiguration and Monitoring (RRM) module intended to support the RESP development. The idea of RESP is to deliver a working prototype of an embedded device as fast as possible to a client and demonstrate its functionality. Then, the product is evaluated in its target environment by the client or the investor. The product's operation can be monitored and its software functionality may be easily tested or changed using the remote reconfiguration and monitoring capabilities. Thus, following the RESP approach can lead to more competitive time of proof-of-concept prototype development which, in turn, may result in gaining swifter funding for a project and better time-to-market.

In the domain of extremely resource-constrained, cost-sensitive, and tentative ad hoc devices with short-term support, the RESP combined with RRM can also be useful during the development in a target environment instead of a bootloader. We hypothesize that in those classes of embedded devices the bootloader itself could even be omitted and the development can be done using the RESP approach with RRM.

The described RRM subsystem is primarily intended to be utilized for embedded software development purposes when the device under development is placed in its target environment, but a reliable remote program memory upload or reconfiguration with a bootloader is not available or is not yet developed. After the successful software development with RESP approach, the RRM module can be extracted from the target platform and the device may work as a stand-alone unit. Alternatively, depending on a specific use case, the RRM can be utilized for the embedded system long-term testing by monitoring its operation with a set of sensors and digital interfaces. Then the RRM will operate as a remote sensing node.

The concepts described in this paper can be applied to multiple classes of embedded systems including microcontroller-based IoT platforms. The RESP approach is applicable for embedded platforms (a) based on microcontroller units (MCUs) in which program memory can be reprogrammed using common in-system programmers or debugging interfaces or (b) platforms based on Field Programmable Gate Arrays (FPGAs). When using FPGA hardware platforms, the presented RRM

The research presented in this paper was partially supported by the National Centre for Research and Development (NCBiR) under Grant No. LIDER/15/0144/L-7/15/NCBR/2016.

device can be used for convenient remote development of both programmable hardware and embedded software. Also, if the RESP augmented with RRM is utilized for programming or reconfiguration of a remote device under development, the Internet connection will be required. Alternatively, the system can be utilized in a local network and in that case the Internet connection is not necessary.

The physical hardware development is not in the scope of this paper as present-today ad hoc prototypes can be based on ready-to-use computing platforms such as well-known Arduino boards family or inexpensive evaluation boards offered by many semiconductor manufacturers, for example STM32 Discovery or Nucleo series from ST Microelectronics or Freedom FRDM from NXP. Those boards can be enhanced with a wide selection of sensor or actuator modules equipped with common integrated extension circuits available at low cost. In the case of the FPGA-based platforms, we assume the use of an already developed physical hardware. The RRM can then be utilized not only to modify the embedded software but also the programmable part of the hardware project.

The rest of this paper is organized as follows. In Section II we summarize recent related research in the area of effective approaches to embedded systems development and remote reconfiguration. In Section III we present the RESP method of developing embedded device prototypes in time-critical manner and in fast-paced business competition environment. Section IV provides a general idea on how to implement a RESP-compliant system. Section V describes the construction methodology and a sample implementation of the RESP with RRM device. Finally, in Section VI we summarize our work.

II. Related research

The embedded hardware platforms are characterized by their diversity when compared to the general-purpose computing platforms. Some of the differences between embedded software and computer application development are the results of the fact that the embedded software must tightly cooperate with usually non-standard, specialized hardware platform and a set of peripheral devices. An embedded software developer needs an access to either a good simulation environment or to the real hardware platform. At the very early development stages, the product requirements are not yet fully specified and they may change multiple times. Engineers can then utilize a general-purpose solution from a wide portfolio of ready-to-use sensing, data acquisition, and actuation devices controlled with e.g. National Instruments hardware and software solutions. However, a prototype can be more enticing for the investor if it could be backed up with a demonstration of a custom working hardware and software even at an initial development stage.

In commercial and industrial MCU firmware development practice, the bootloader is one of the most specialized software parts. There are multiple MCU platforms that provide a dedicated bootloader without any additional installation, but that solution usually relies on a predefined interface and protocol. Changing the default settings requires either to re-implement the bootloader or to modify it. The Nordic Semiconductor's nRF52832 is a good example of integrated circuits (ICs) which provide a very well developed Bluetooth Low Energy (BLE) bootloader. However, not all ICs can utilize BLE to update firmware and not all manufacturers provide such a convenient firmware update functionality. Usually, a bootloader is a very application-specific part of embedded software and it needs to be either developed solely for a given platform and interface or ported from another project. Embedded systems based on an application microprocessor (with Memory Management Unit and running e.g. Linux) often use U-Boot as a first stages' bootloader. For the MCU-based embedded devices it is difficult to point out a most common bootloader solution – different platforms offer different solutions, such as the STM32 Bootloader [1]. An example of a custom bootloader is described in [2]. The bootloader program must be well designed and tested to avoid firmware corruption eventually resulting in an inability to reprogram the device with the provided bootloader. Another common problem in writing a bootloader is to make it insensitive to transmission errors and complete transmission interruption. Preventing those situations require much time, engineering effort, and some design redundancy (additional memory, correcting errors in software, etc.). All the problems mentioned here can be solved, but it usually costs additional development time.

The ability to remotely reprogram, reconfigure, and supervise an embedded system is especially useful in the domain of programmable logic, mainly FPGA. We should also be aware that the software and hardware development flows may proceed in parallel, depending on a design (e.g. in [3]). The reconfiguration allows developers and maintenance staff not only to remotely update firmware, but also to change the functionality of the system [4]. For example the FPGA-accelerated smart camera described in [5] is able to run multiple configurations which can be substituted depending on a higher level adaptation policy. Enhancing an FPGA-based device with the remote reconfiguration feature costs design issues due to losing the programmable logic functionality during the reconfiguration process [6]. In that case the partial reconfiguration feature [7], [8] could be helpful, but it tends to complicate the hardware-software design flow hence it may be ineffective for time-critical project. Less sophisticated, but much more convenient methods include using remote programmers, such as the Intel FPGA Ethernet Cable (formerly the Altera EthernetBlaster II) as described for example in [9]. Those devices offer only limited computing capabilities at the target side. In our solution we greatly increased the computing power of the remote programmer by utilizing a single-board computer (SBC). It allowed us not only to easily implement modern communication protocols, but also to gain much more flexibility compared to other solutions.

The idea of remote programming can be extended to the remote firmware management. It is also a well-known topic, and some aspects of networked systems performing such tasks are patented e.g. in [10], [11]. Currently the remote reconfiguration and management are, however, typically per-

formed by using similar architecture and by utilizing e.g. OMA Lightweight M2M (LWM2M) protocol. The LWM2M is an increasingly popular remote management protocol for intelligent connected and IoT devices [12], [13]. It has low transmission overhead and its implementation can be relatively easily ported to many connected platforms. LWM2M also supports a framework for a remote firmware update which was especially desirable in the solution described in this paper.

Despite the fact that the embedded and general computing hardware platforms are different, embedded software development can be based on similar principles as the computer software development. For example, agile development model has been adapted to embedded systems [14], [15]. Other approaches to embedded systems development, such as the *V-Model* descibed in e.g. [16], can also be applied to fast development of embedded software.

As presented in this section, there are multiple systems, methods and approaches to fast development of embedded software. There are also multiple solutions for remote reconfiguration, programming, and management of embedded systems. Those methods can be applied to ad-hoc embedded systems firmware and software development process. Based on the presented state of the art we propose the following solutions. First, we suggest, that the remote programmers and firmware management systems can be improved. To prove that statement we propose the practical implementation of the RRM described further in this paper. Moreover, the RRM can be utilized to implement the RESP approach. We state and prove that utilizing the proposed RESP approach can reduce time-to-market and allow developers to deliver a working prototype of the resource-constrained embedded system faster compared to traditional approaches.

III. Proposed development method

To explain and justify the proposed RESP development method, we introduce a simplified model of the experimental embedded system software development. The model reflects practical experiences while cooperating on an innovative IoT solution with actual business representatives. We assume that the model is applicable when the MCU-based embedded hardware platform is developed and it is ready for initial firmware implementation.

The simplified development flow is following. We assume that the development time can be represented by a number of *iterations*. In this case the *iteration* can be perceived e.g. as a *time interval* or as a *programming task*. In our considerations each iteration represents an amount of work required for one developer or for a team to complete a given task. Many developers can work on the project concurrently, but in the simplified model we just count the total number of iterations as if the project was developed in a fully sequential manner. A goal is specified for each iteration. Reaching Alpha development stage requires at least two iterations: the initial development stage and the actual Alpha development-testing stage.

In this model the device is ready to be shipped to the potential customer or an investor for further review if the following two conditions are met: 1) the application has at least minimum experimental functionality with basic Alpha stage tests done, and 2) the device firmware can be reliably and remotely updated or changed to allow developers and the client to collaborate on the final firmware version and further features. The consequence of the latter prerequisite is that the bootloader should be developed up to the final release version unless the RESP approach is used. In this model the embedded device is passed to the potential customer or investor after its development reaches the Alpha stage. Then the customer initially evaluates the product. If the customer accepts the initial results, the developers shall continue to work on the product's software (*success*). Otherwise the development of the product in the current form shall be ceased (*failure*). That situation can happen e.g. when the product is unable to meet the requirements or it needs a major redesign. In the *failure* case, most of the recent developing effort is wasted because the project or the idea has been rejected by the investor or the client.

Figure 1 shows a graphical representation of a sample embedded system development time line using the presented model. The goals of each iteration are denoted inside a box representing that iteration. Two cases are analyzed.

The first case, which is shown in Figure 1 (a), represents a scenario with a classic approach applied. In order to reach the product development stage at which the prototype can be passed to the client, the developers require six iterations: four iterations for the bootloader development and two for the initial application development. In case of success, only the application needs to be further developed, but in case of

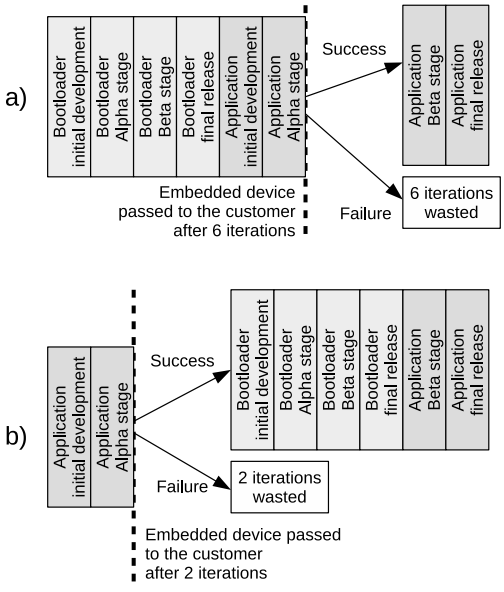


Fig. 1. Sample development time lines for the described model and: typical approach (a), utilizing RESP approach with RRM module (b).

failure, the six iterations are wasted.

In the second scenario shown in Figure 1 (b), the developers utilize the RESP approach with the dedicated RRM module. The device can be shipped to the client even at an early development stage, after just two iterations and the bootloader development can be postponed for later stages. In case of the project failure, only two iterations are wasted.

IV. Implementation concept and overview

In this section we present a general ideas which concern realization of RESP approach with the RRM module.

To implement the RESP development approach the developers should utilize the *RESP-compliant system*. The RESP-compliant system consists of a management unit and the RRM hardware with a dedicated software. One RRM can be connected to one or multiple instances of the device under development. The number of devices under development connected to one RRM depends on hardware interface capabilities of the utilized RRM embedded computer. Moreover, it is possible to manage multiple RRMs using a single server with a remote application. Those features allow software developers to manage hundreds of devices with a single server. The sample set-up of a multi-node reconfiguration-debugging system using RRMs is presented in Figure 2.

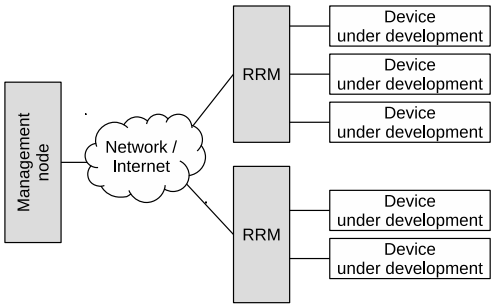


Fig. 2. Sample RESP-compliant hardware architecture with multiple RRMs capable of managing several devices under development.

After the embedded software development is complete, the RRM can either be disconnected from the device under development, or it can be left connected to monitor the operation of the embedded system by utilizing e.g. external sensors, such as temperature, voltage or current.

To provide flexibility and advanced functionality, the implementation of RRM is based on an SBC. The use of SBC allows developers to implement much more advanced features compared to typical programmers with network interface.

The choice of generic hardware platforms combined with free and open source software solutions is one of the key concepts in the implementation of RESP-compliant system. The configuration needs to be performed with the use of scripting languages. Combining those features allows the RESP-compliant system to be very flexible and easily adapted to new target hardware platforms and specialized use cases. The SBC-based RRM can potentially log, filter, and transmit the debug messages printed on e.g. a serial port of the device under development. The messages then can be transmitted over Internet to developers and testers in a remote location. The functionality of RRM can be extended even further. As an example, the RRM after extending its software may allow developers to obtain even a live view of the system under development. That feature obviously cannot be utilized to reliably monitor the operation of a safety-critical equipment, however it can be utilized to determine overall environment conditions for which sensors were not included in the initial design or to detect some obvious reasons of malfunction. For example the camera can facilitate determining if the light intensity in general is low or high, if the device has been covered, or if it has been moved.

V. Practical implementation of the RESP-compliant system

This section contains technical description of the sample RESP-compliant reconfiguration and monitoring architecture developed according to the information in previous sections of this paper.

A. RRM hardware design

Currently there are many different SBCs available at very low cost. In our sample implementation we have chosen Raspberry Pi Zero to implement the RRM. The Raspberry Pi Zero has multiple advantages as a choice for RRM. That computer is characterized by its very low cost and compact size. Another advantage of it is a General-Purpose Input-Output (GPIO) interface presence and its 3.3 V logic levels compatibility, which makes it well suited to implement versatile digital interfaces, including programming interfaces.

The network interface is provided with a generic Wi-Fi dongle with Universal Serial Bus (USB) interface. Depending on the SBC used, the network interface could also be implemented using a built-in peripheral as in e.g. full-sized Raspberry Pi computers.

Devices under development are connected to RRMs directly using Joint Test Action Group (JTAG) or Serial Wire Debug (SWD) interface for programming and reconfiguration purposes. Those interfaces are implemented using GPIO hardware of the SBC. Devices equipped with a built-in programmer with USB interface, can connect to the SBC with that interface. In the proof-of-concept implementation no additional protection circuits were added, but they should be considered in the RRM hardware.

Another important aspect of the RRM implementation is the possibility to measure various physical quantities: the operation parameters of the device under development (e.g. input-output voltages, temperature, logic states, power supply current) and the general parameters of the environment (e.g. temperature, humidity). We have chosen Inter-Integrated Circuit (I^2C) as a typical interface for RRM external sensors because it is supported in hardware and software of many SBC platforms or it can be relatively easily implemented using GPIO. There is also a wide selection of compatible sensors

with I^{2}C interface and, what is very convenient, multiple sensors can be connected to a single bus.

As a sensor for the first implementation of the RRM we have chosen a current sensor intended to monitor a power supply current of the target device under development. The reason was to provide a sensor which is commonly needed during the process of developing embedded software. In our practice, the power management of an embedded system is one of the vital parameters that needs to be monitored and we often need embedded systems to be optimized for energy efficiency. The choice of INA219 current sensor appeared to be reasonable because it is equipped with digital I^{2}C interface and it is able to measure current at the power supply rail (high-side).

B. RESP-compliant system logic and software architecture

The internal structure of RRM and its sample connections to multiple devices under development are shown in Figure 3.

We intended to chose a free, open source, and highly configurable solution for interfacing management software with hardware reconfiguration-programming interfaces of target devices under development. An actively developed project that seems a very good choice for that purpose is OpenOCD. It is a powerful and flexible debugging and memory programming tool which can be configured with TCL scripts and which provides multiple convenient control interfaces.

As the management interface needs to easily cooperate with standard software solutions, we decided to utilize the more and more popular OMA LWM2M protocol. The management node is implemented as a server for the OMA LWM2M protocol. Eclipse Leshan LWM2M server demo was chosen for the project implementation purposes. It provides basic unified network interface and Representational State Transfer (REST) application programming interface (API). Custom LWM2M object definitions have been added to the server in order to properly recognize custom resources which are specific to the project. The LWM2M API utilizes Firmware Update object from the specification and three custom LWM2M objects: OpenOCD LWM2M RPC, Firmware Target Selector and INA219 current sensors interface.

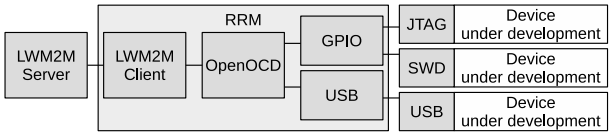


Fig. 3. General block diagram of the RRM and an example of connecting devices under development.

The internal software architecture of the RRM is shown in Figure 4. The software operates as multiple separate LWM2M client instances, one for each device connected to the RRM hardware. The LWM2M allows software developers to reprogram the target device and fetch sensors readings. It holds an OpenOCD instance for each device.

The software is written in Java with Eclipse Leshan libraries. For every device instance it creates a separate LWM2M client instance and runs OpenOCD. In our implementation the OpenOCD uses remote procedure call (RPC) interface for clients to issue TCL commands and obtain results from TCL engine. The commands are generated according to information derived from configuration files passed as arguments. The configuration files are described in detail further in this paper (please refer to Section V-C).

When using Eclipse Leshan the binary image for the reconfiguration purposes of the device under development can be transferred from the management server. Alternatively, the reconfiguration can be initiated from the server along with providing an Unified Resource Identifier (URI) to a location in which the binary image is available. We have implemented the latter option, because in practice it proved to be more flexible and convenient for the developer who manages the process of reconfiguration or reprogramming. In current implementation, two protocols are supported for transferring firmware images: Hypertext Transfer Protocol (HTTP) and HTTP Secure (HTTPS).

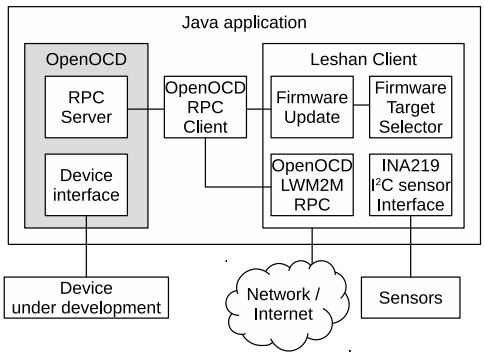


Fig. 4. Software architecture (gray block means external process managed by parent process).

C. Configuration

The configuration method was devised especially for RRMs. YAML was chosen as a file format for its simplicity and readability. SnakeYAML library was chosen to load YAML files into configuration objects. It also provides runtime validation of configuration syntax.

There are two types of configuration: the first for each device class and the second for device instances connected to RRM. Device class configuration is stored in files separate for each device class (such as STMF4DISCOVERY, ZYBO). It contains OpenOCD initialization and flash commands. Instance configuration specifies communication interface and sensors for each of the actual devices connected to RRM (such as stm-prod, zybo1).

Listing 1. Sample instances configuration file for two identical devices connected to the RRM hardware.

```
instances:
- name: stm-by-gpio
  deviceConfigPath: stmf4discovery/config.yml
  interface:
    bcm2835gpio:
```

```
      swdNum1: 25
      swdNum2: 24
      trstNum: 7
      srstNum: 18
      transport: swd
- name: stm-with-current-sensors
  deviceConfigPath: stmf4discovery/config.yml
  interface:
    custom:
      initCommands:
       - source [find interface/stlink-v2.cfg]
       - transport select hla_swd
  sensors:
    - ina219:
        i2cAddr: 0x40
        i2cBus: 1
        shuntResistance: 0.1
```

A sample configuration is shown in Listing 1. It is a configuration file for two identical device instances connected to the RRM hardware. The first instance is connected by JTAG directly to the GPIO ports of the SBC, whilst the second instance is using board built-in USB debugger and also has current sensor attached. It was prepared for BCM2835 chip present in Raspberry Pi boards. It eliminates the necessity of writing OpenOCD commands, yet allowing to specify custom GPIO port numbers and transport.

An example of the device configuration file is shown in Listing 2. It allows system developers to define device class specific OpenOCD initialization commands as well as commands that are executed as a result of executing Update resource on Firmware Update object.

Listing 2. Device configuration file

```
initCommands:
 - source [find target/stm32f4x.cfg]
 - reset_config srst_only
firmwareTargets:
 - name: mcu
   flashCommands:
    - reset init
    - flash write_image {{ image }}
    - reset
```

D. Implemented sensor support

In the sample implementation, the RRM software supports multiple sensors connected using I^2C bus supported by the physical interface of the utilized SBC. In the presented software implementation, each RRM instance supports zero, one or many INA219 current sensors as a sample implementation of that functionality. The RRM software fetches data from sensors using Pi4J library and provides on-demand access with multi-instance INA219 sensors LWM2M object (*urn:oma:lwm2m:ext:3403*). For convenience, current, voltage, and power values are provided.

E. Reconfiguration flow

Reconfiguration flow is presented as a sequence diagram in Figure 5. The reconfiguration process consists of two dependent stages. In the first stage, the developer provides a URI for a new binary image to be uploaded to the target embedded system using Eclipse Leshan and LWM2M protocol. In the second stage, the execution of the reconfiguration itself takes place – the RRM performs the reconfiguration with an instance of OpenOCD.

F. Basic security considerations

The RRM is primarily intended to be applied only temporarily during the embedded system's development stage, but the security is still an important issue. The basic transport-level security using Datagram Transport Layer Security (DTLS) is supported by default for LWM2M. RRM software supports HTTPS to enable secure path of fetching images. User might want to add another layer of security such as an encrypted virtual private network (VPN) or Secure Shell (SSH) tunnel.

G. Achieved prototype functionality and practical verification

We have successfully developed a working prototype of RRM device according to the ideas described in previous sections. The presented example of the RESP-compliant system provides LWM2M-based remote firmware upgrade API for a wide variety of embedded systems. Thanks to the fact that we have chosen OpenOCD as the software for reconfiguration control, the RRM supports virtually any target platform that can be reconfigured or reprogrammed with OpenOCD – the main requirement is that the system administrator provides adequate configuration scripts. We have designed and implemented a specialized configuration scheme using YAML scripts. The utilization of free and open source software and common versatile off-the-shelf hardware allows developers to easily modify, extend, and tailor the functionality of the RRM for a particular use case. The implemented and presented sensor extension for measuring a target's power consumption may be extremely useful in remote debugging and development of energy constrained embedded systems.

We have measured reconfiguration times achieved with RRM and the module's average power consumption. The reconfiguration time results are summarized in Table I and visualized in Figure 6. Each reconfiguration time is an average computed from 10 sample transmissions. The results were obtained during reprogramming or reconfiguring MCU and FPGA on development boards. The reprogrammed MCU was STM32F407VGT6 and the programming interface was SWD implemented with GPIO of Raspberry Pi Zero. The reconfigured FPGA was Xilinx Zynq-7000 on Digilent Zybo ARM/FPGA SoC Trainer Board with the JTAG programming interface connected to SBC using USB interface. The SBC was connected to a local network using a generic Wi-Fi card with USB interface. We have measured power consumption of the RRM hardware. It averages to 1.3 W when idle (Wi-Fi connectivity enabled, LWM2M server is during registration) and to 1.4 W when performing firmware update. To set-up a Wide-Area Network (WAN) we used a virtual machine (VM) located in New York which was hosting LWM2M server and binary images. The LWM2M client was located in Krakow, Poland. The measured WAN Round Trip delay Time (RTT) was 120 ms at maximum transfer rate of 12 Mb/s. The

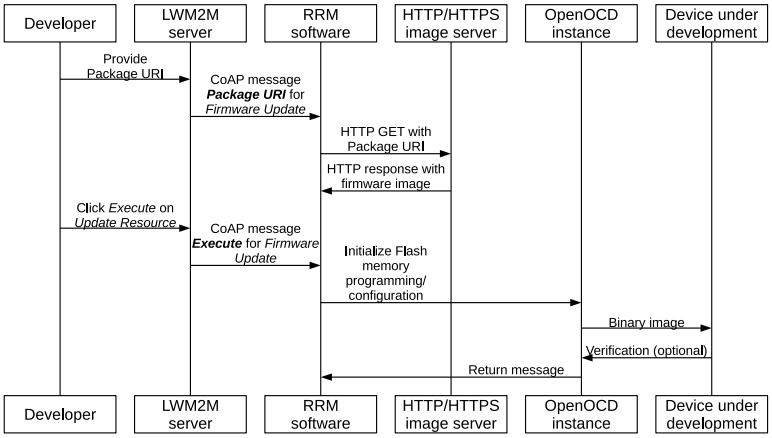


Fig. 5. Reconfiguration flow

TABLE I
RESULTS OF PRACTICAL EXPERIMENTS

Target type	Firmware/ configuration file size (KiB)	Network for LWM2M	Image server protocol	DTLS for LWM2M	Total reconfiguration time (s)	Binary upload time (s)
MCU	21.8	LAN	HTTP	disabled	1.9	1.3
MCU	21.8	WAN	HTTPS	enabled	3.3	1.3
MCU	295.3	LAN	HTTP	disabled	10.5	9.1
MCU	295.3	WAN	HTTPS	enabled	12.1	9.1
MCU	978.9	LAN	HTTP	disabled	26.0	23.9
MCU	978.9	WAN	HTTPS	enabled	28.3	23.9
FPGA	4185	LAN	HTTP	disabled	12.2	9.1
FPGA	4185	WAN	HTTPS	enabled	14.7	9.1

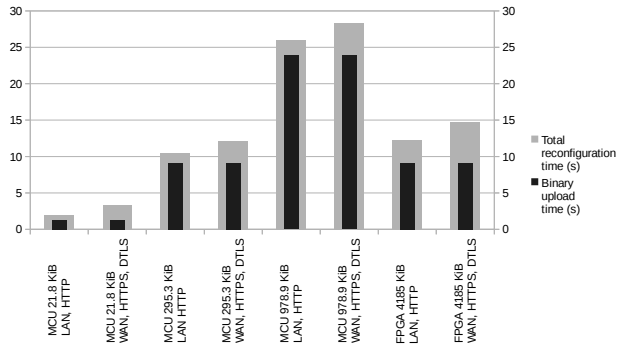


Fig. 6. Remote reconfiguration and programming times comparison.

reconfiguration time consisted of two components: an actual binary image upload time (target memory programming) and a communication overhead. According to the results the actual binary image upload was always taking an overwhelming amount of time.

VI. CONCLUSION AND FUTURE WORK

In this article we present the RESP approach for fast embedded systems prototyping. The presented RESP approach may be one of the future directions in embedded software development, especially for experimental and ad hoc systems, because it is aimed at effectiveness, low development cost, short time-to-market, and minimizing implications of a project failure. We propose a theoretical justification for the proposed solution and a sample practical implementation of the compliant system that supports the described approach. The RESP approach does not limit the use of common software development techniques and approaches, such as agile, but it may extend their possibilities and simplify their application in practice.

We also propose a new approach to remote programming and reconfiguration of microcontrollers and FPGAs by implementing much more advanced functionality in the RRM than in common off-the-shelf programmers. The use of the RRM can speed-up the embedded software development before the final version of an application-specific bootloader is developed. We present a sample working implementation of the RRM. It supports modern and promising network protocols, mainly OMA LWM2M, and flexible monitoring features. Thanks to the application of OpenOCD software along with the described implementation of scripting configuration, the RRM can be adapted for a wide variety of embedded hardware platforms, MCU-based IoT nodes, and their respective memory programming interfaces. It can be a versatile remote reconfiguration and development hardware-software tool. Thanks to the use of compact but full-fledged computer platform, the developed RRM system has far superior versatility compared to commercially available remote programmers-debuggers. The comparison between the RESP-RRM approach and solutions reviewed in Section II is summarized in Table II.

The RRM can be used not only for RESP development, but also as a multi-purpose remote reconfiguration and management extension as well as long-term operation monitor or

TABLE II
COMPARISON BETWEEN THE REVIEWED AND THE PROPOSED SOLUTIONS.

Reviewed solution	Aspects which can be improved	Aspects improved by utilizing the RESP approach and/or the RRM
Bootloaders for embedded MCU-based devices [1], [2]	Bootloaders require much development time and effort.	Bootloader is not required during initial development, and can be added at later development stages.
Remote FPGA reconfiguration, programmable logic partial reconfiguration with traditional design flow approach [5], [6], [7], [8]	Complex design which includes a reconfiguration subsystem and more elaborate design flow when considering partial reconfiguration.	RESP can improve time-to-market by allowing for remote development of both programmable hardware and software. The RRM substitutes additional fixed hardware and logic resources for remote reconfiguration. Later, the RRM can be disconnected when not needed.
Remote programmers for FPGA designs as utilized for example in [9]	Single-purpose hardware, no advanced management features, not customizable.	High flexibility of the hardware and software. Possibility to implement advanced firmware and configuration management.
General approach to remote firmware management [10], [11], [12], [13]	More versatile and general-purpose approach with standardized protocols can be considered.	The proposed solution can be easily customized to various applications.
Common approaches to embedded software development [14], [15], [16]	Usually considered for software development with hardware available locally.	RESP does not interfere with a selected embedded software development approach, but it allows developers to achieve a working prototype faster and without a direct access to a hardware platform.

logger for various embedded and connected devices, including IoT nodes. We plan to implement some of that features during further development. An example of a very useful extension is an integration of a camera module for basic visual inspection of the device under development. The RRM could also be considered as a tool which allows remote access to specialized embedded systems for educational and training purposes.

REFERENCES

[1] "Stm32 bootloader," https://github.com/akospasztor/stm32-bootloader, accessed: 2017-12-29.

[2] R. J. Landeo Márquez, "Can bus bootloader for the stm32f407vg," Master's thesis, Universitat Politècnica de Catalunya, 2017.

[3] A. V. Parkhomenko, O. Gladkova, E. Ivanov, A. Sokolyanskii, and S. Kurson, "Development and application of remote laboratory for embedded systems design," *International Journal of Online Engineering (iJOE)*, vol. 11, no. 3, pp. 27–31, 2015. doi: 10.1109/REV.2015.7087265

[4] M. D. V. Pena, J. J. Rodriguez-Andina, and M. Manic, "The internet of things: The role of reconfigurable platforms," *IEEE Industrial Electronics Magazine*, vol. 11, no. 3, pp. 6–19, 2017. doi: 10.1109/MIE.2017.2724579

[5] R. Brzoza-Woch, A. Ruta, and K. Zieliński, "Remotely reconfigurable hardware–software platform with web service interface for automated video surveillance," *Journal of Systems Architecture*, vol. 59, no. 7, pp. 376–388, 2013. doi: https://doi.org/10.1016/j.sysarc.2013.05.007

[6] R. Brzoza-Woch and P. Nawrocki, "Fpga-based web services–infinite potential or a road to nowhere?" *IEEE Internet Computing*, vol. 20, no. 1, pp. 44–51, 2016. doi: 10.1109/MIC.2015.23

[7] R. Hymel, A. D. George, and H. Lam, "Evaluating partial reconfiguration for embedded fpga applications," in *Proceedings of High-Performance Embedded Computing Workshop (HPEC'07)*, 2007, pp. 1–2.

[8] C. Conger, R. Hymel, M. Rewak, A. D. George, and H. Lam, "Fpga design framework for dynamic partial reconfiguration," in *Proceedings of Reconfigurable Architectures Workshop (RAW)*, 2008.

[9] J. Belleman, D. Belohrad, L. Jensen, M. Krupa, and A. Topaloudis, "The lhc fast beam current change monitor," *WEPF29, IBIC*, 2013.

[10] M. Ogura, "Remote management system, intermediary apparatus therefor, and method of updating software in the intermedary apparatus," U.S. Patent US7 516 450B2, 2003.

[11] R. Pathak, "Remote firmware management for electronic devices," U.S. Patent US9 112 891B2, 2007.

[12] S. Rao, D. Chendanda, C. Deshpande, and V. Lakkundi, "Implementing lwm2m in constrained iot devices," in *Wireless Sensors (ICWiSe), 2015 IEEE Conference on*. IEEE, 2015. doi: 10.1109/ICWISE.2015.7380353 pp. 52–57.

[13] J. Prado, "Oma lighweight m2m resource model," in *IAB IoT Semantic Interoperability Workshop*, 2016.

[14] J. Grenning, "Agile embedded software development," *ESC Boston*, 2011.

[15] D. Dahlby, "Applying agile methods to embedded systems development," *Embedded Software Design Resources*, vol. 41, p. 1014123, 2004.

[16] "Embedded System development Process Reference guide," Information-technology Promotion Agency, Reference Guide, 2012.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 637–643
DOI: 10.15439/2018F278
ISSN 2300-5963 ACSIS, Vol. 15

FetchIoT: Efficient Resource Fetching for the Internet of Things

Badis Djamaa, Mohamed Amine Kouda, Ali Yachir, and Tayeb Kenaza
Ecole Militaire Polytehnique,
BP 17, Algiers, Algeria
{badis.djamaa, kouda.amine, ali.yachir, ken.tayeb}@gmail.com

***Abstract*—Finding the right resource at the right time and space is a key enabler for a wide adoption and spread of the Internet of Things (IoT). The Constrained Application Protocol (CoAP) and related standards are among the most prominent efforts working towards such a goal. Indeed, CoAP-related standards provide interesting mechanisms for resource discovery in both centralized and distributed architectures based on the CoAP's GET method. In this paper, we, first, highlight the limitations of GET-based discovery mechanisms. The paper, then proposes a new solution using the recently standardized FETCH method and develops its specifications, rules and semantics. The proposed solution is implemented in the recently released, secure and reliable OpenThread platform and compared with GET-based approaches in different home automation scenarios. Obtained results demonstrate the performance of FETCH-based discovery in achieving fine-grained, time-efficient and reliable discovery while preserving network resources.**

I. Introduction

WITH the growing number of sensors, actuators, devices, smartphones and embedded chips, along with the (r)evolution of computer and network technologies, the world is talking more and more about the Internet of Things (IoT); a term designating the extension of the Internet to everyday objects. When interconnected, these objects can form a network for measuring, storing, transferring, processing, and exchanging data between physical and virtual worlds. Today, such smart objects are able to discover, detect and exchange messages across the Internet thanks to the newly introduced protocols such as 6LoWPAN [1], RPL [2] and CoAP [3]. In fact, with the provided features, smart object networks can be built spontaneously and can be doted with capabilities of self-configuring, self-regulating and self-healing.

In IoT, the vision is that a significant number of new devices including refrigerators, clothes, cars, and traffic-lights will be dynamically connected to the Web for communication, command and control of the surrounding environment. This trend creates the so-called Web of Things (WoT) with the introduction of new constrained servers that have different features from traditional web servers and users. This pushes the WoT to face many challenges related mainly to the heterogeneous nature of networks constituted by these objects and their very limited capacity in terms of computing resources, communication capabilities, memory and energy. To overcome such challenges, the offered functionalities in the WoT are encapsulated as autonomous constrained REST (Representational State Transfer) resources [4] that are accessible from other objects or traditional Web services. This simplifies transparent integration of the physical world with the virtual one. However, to do so, there must be mechanisms to discover available resources and their capabilities with the minimum of human intervention. Thus, resource discovery becomes a fundamental requirement for the success of any IoT solution.

One of the main protocols implementing the REST-based mechanism for resource description and discovery in the web of things is CoAP [3]. Indeed, besides being the de-facto standard for data exchange in the WoT, CoAP provides distributed and centralized solutions for achieving resource discovery. It does so by employing the GET method for the sake of finding available resources in an IoT environment. For instance, a device searching for available temperature sensors in its environment issues a GET request to a *well-known* URI (Uniform Resource Identifier) asking for all sensors offering resources of type temperature. The queerer will get responses with the description of such resource and chooses the ones that best meets its needs.

This GET based mechanism is limited in many aspects that will be discussed and detailed in this paper. To overcome such issues, we introduce a new usage of the FETCH method [5] for the sake of efficient resource discovery in the IoT. The specification of such a usage along with the definition of rules allowing to achieve rich, expressive and compact resource discovery in the web of things are the main contributions of this paper.

Finally, it should be noted that to the best of our knowledge, this is the first paper introducing the use of the newly standardized FETCH method for resource discovery in the IoT. The paper also adds a resource discovery layer to the secure and reliable openThread networking protocol making the proposed approach ready for commercial deployments.

The remainder of this paper is organized as follows: Section II reviews related resource discovery work over CoAP. This will be followed by identifying the main issues of such approaches before proposing our approach and detailing its mechanisms in section III. section IV is devoted to implementing and evaluating the performance of the proposed approach in multiple IoT scenarios over the commercially-proven, secure Thread platform. The paper ends in Section V with a conclusion and directions for future work.

II. Related Work

Resource discovery is a well-investigated topic in traditional Networks with a plethora of solutions proposed in the literature. IoT objects have radically different features than traditional Web servers. As a result, traditional discovery approaches can not be applied directly and will not produce accurate and efficient results in many IoT scenarios. Consequently, new solutions are emerging for the IoT. Such solutions follow two main architectures, centralized and distributed, with standards being proposed for both architectures. The most promising ones are based on CoAP and/or DNS-SD [6]. While DNS-SD is starting to get intention, currently, CoAP-based discovery is the main standard solution in today's IoT applications.

CoAP-based resource discovery can be achieved via three main mechanisms, namely: CoAP Resource Directory (RD) [7], CoAP Distributed Resource Directory (DRD) [8], and CoAP Resource Discovery [9]. The main purpose of these mechanisms is to provide URIs, also called links, for the resources available within a server, as well as the attributes that describe them [9].

With the RD, all the resources offered by the servers are saved in a single directory so that clients can discover any required resource by looking up the RD. For instance, once the RD has been successfully discovered, a server can register its resources in the RD by performing a POST request to the path indicated by the RD. When a client wants to search the RD, they must issue a GET request to the RD. For this, the client uses a specific request to obtain the results that correspond to its interest. The use of GET imposes many constraints on the expressiveness of the request since the parameters must be "bundled up in some unspecified way into the URI" [5]. It should be noted that Following the success of RD, many solutions including [10] have been proposed. All, however, suffer from the same problems related to the use of GET for resource discovery.

In DRD-based discovery [8], before an object can register its resources, it must find an Entry Point (EP) at the DRD. The initial EP can be any object connected to the DRD. In order to find an EP, one method is to use a multicast address */.well-known*, where the object sends a POST request to that */.well-known* address to obtain the DRD information. Other means include searching for the nearest EP or DRD using dynamic discovery [8]. To improve this approach, work based on hierarchical repertoires has been proposed. Indeed, authors of [11], have introduced a usage of the REsource LOcation And Discovery (RELOAD) protocol [12] to discover CoAP resources. RELOAD forms an overlay network to provide storage and messaging services in a peer-to-peer (P2P) environment and allows applications to define specific use cases. For instance, [11] authors describe how to use CoAP with RELOAD to discover interconnected CoAP resources across a large geographic area. However, as with the RD, the discovery is based on GET, which in addition to the above limitations, only supports discovery in the CoRE Link Format (CLF) [9].

Direct approaches rely on IP multicast to achieve discovery [9]. Similarly to the above approaches, it uses the GET method to diffuse a request to all nodes in an IP domain targeting the well-known URI (*/.well-known/core?search**) of all nodes members of the group's multicast address. Unlike RD and DRD, here, not being able to filter the request may generate a huge number of irrelevant responses that consume network resources and slow its operations. Thus, direct resource discovery must include the *search** filter in its requests, which is still insufficient for fine-grained efficient discovery. Finally, a hybrid approach that tries to take advantages of both direct discovery and RD is proposed in [13]. However, similarly to the above, it also relies on the GET method to formulate the requests.

To the best of our knowledge, all CoAP-based discovery approaches deploy the GET method that limits their capabilities into achieving rich and concise discovery in IoT. Such limitations will be discussed in the following section before introducing our FETCH-based resource discovery for the IoT approach.

III. FETCH-based resource discovery for the IoT

Before introducing the FETCH-based resource discovery for the IoT, the following subsection identifies and discusses the main issues with the GET-based approach.

A. Issues with the GET-based resource discovery

As in HTTP, the GET method is used to obtain the complete representation of a resource, which can be refined according to the additional parameters conveyed in the request. However, using GET, a user/device can only allow the specification of a URI and the query parameters in CoAP options [3] as can be seen in Figure 1. Indeed, the GET method does not support the transmission of a payload detailing the request, which generates verbose replies (Figure 1) that consume energy and throughput. These restrictions have caused some applications to use the POST method for the sake of formulating queries with semantic alteration and standard compatibility violation [5].

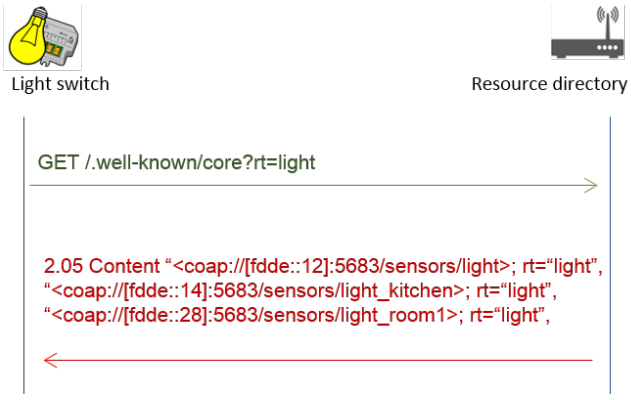


Fig. 1. An example of GET-based resource discovery

The following points summarize and discuss some of the major limitations and gaps related to using GET for the sake of resource discovery:

- Query parameters of GET-based requests are limited by the form and the size of the URI, which restricts their capacities to convey user/device requirements. This constraint imposes strict limitations on the expressiveness of requests, which may result in generating verbose replies that will, at the end, be discarded by the client.
- Despite some research efforts aiming to increase the number of filters included in a GET request, until now, CLF [9] limits this number to one by query. Indeed, a GET discovery request follows the scheme: (*/.well-known/core ?search**), where *search* represents a single parameter. For example: *GET /.well-known/core?rt=light*.
- The GET method does not support filtering based on logical operations (AND, OR, <,>,> = ...) between query parameters. These types of filtering are necessary to refine queries to receive only the most relevant answers.
- With GET, we can not include and/or exclude nodes/resources based on cached descriptions or any information known by a user or device. The use of these features is important for resource discovery in the IoT. Indeed, with such features, the user/device will have the possibility to specify the known resource descriptions in order not to include them in the answer. More importantly, having such an option will provide the possibility of specifying particular nodes to avoid or include, as well as the specification of the particular requirements in terms of security, reliability and the required quality of service.
- With GET, we can not specify/limit the search domain, the location of the searched resources, the maximum number of hops a query can reach, and so on. Such information is of paramount importance for a more fine-grained, effective and efficient discovery. For example, a client looking for temperature sensors in his immediate environment is not interested in having answers from all the temperature sensors available in the network.
- Finally, GET-based resource discovery only supports discovery operations in the CoRE link format [9], which limits its applicability when resources are described in other formats such as JSON, CBOR, EXI, etc.

From the above, and knowing that in IoT most objects are very constrained in terms of resources, a substitute mechanism that offers an explicit, compact and comprehensive expression of user requirements is required. Indeed, it is very important for an alternative approach to minimize congestion, excessive use of resources, energy consumption and latency, while offering more relevant results to better satisfy user requests. Such an approach will be the subject of the following sections.

B. The CoAP FETCH method

The FETCH method has been proposed in draft-ietf-core-etch-04 [5], which has just become RFC 8132 [14]. FETCH tries to provide a solution that covers the gap between the use of GET and POST. In the same way as POST, the parameters of FETCH are transmitted in the payload of the request rather than in the context of its URI. However, unlike POST, the semantics of FETCH is more specifically defined to ensure tasks similar to those of GET.

As defined in RFC 8132 [14], the FETCH method of CoAP is used to get a representation of a resource, providing a number of query parameters. Unlike GET, which requires a server to return a representation of the resource identified by the request URI (as defined by RFC 7252 [3]), the FETCH method is used by a client to request the server to produce a representation as described by the query parameters (including query options and payload) based on the resource specified by the effective URI. As a result, the payload returned in response to a FETCH request can not be assumed to be a complete representation of the resource identified by the effective request URI.

Using the FETCH method for the sake of resource discovery in IoT can remedy GET-related problems identified in the previous section. It is also extremely useful when efficient result filtering that preserves network resources is desired. For instance, if a client is only interested in the types of resources available at a server, it formulates a FETCH request asking of that part of the representation as can be seen in Figure 2. The server, then, replies only with the required information, which saves throughput and energy.

Furthermore, if several resources of similar types are provided by different objects and the client knows beforehand the existence of certain resources not meeting its requirements, it can indicate them in the request payload to avoid undesired replies. Thus, a client can exclude non-needed resources, nodes, and parameters by specifying them in the body of its request. In the same way, a client can specify the desired parameters, nodes, content-formats, etc. to be included. Moreover, a client/device can combine all their requirements and knowledge in a single request to further filter returned responses.

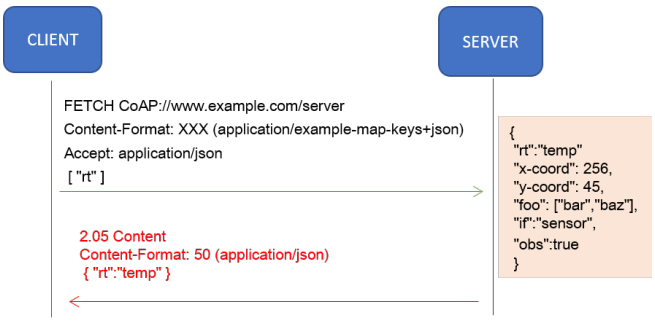


Fig. 2. A FETCH request in JSON

From the above, it is clear that the new FETCH method opens up promising prospects for successfully filtering the returned results, which can significantly reduce network traffic, congestion, collision problems and power consumption when performing resource discovery in constrained networks. In view of these advantages, the question, which will be ad-

dressed in the following section, is how to design an effective use of FETCH for resource discovery in the IoT.

C. FETCH-based Resource discovery for the IoT

Having discussed the limits of GET-based resource discovery and presented the opportunities provided by FETCH, this section introduces the proposed design of an effective scheme of FETCH-based resource discovery.

To do so, we took inspiration from the IGMPv3 protocol for the use of resource filtering based on the INCLUDE and EXCLUDE modes; EXINC for short. These modes are considered very adaptable to the context of resource discovery in the IoT. Indeed, it is useful to have a mechanism that allows to EXCLUDE the resources we do not need. At the same time, providing a technique allowing to INCLUDE the desired features is crucial for an fine-grained resource discovery. Indeed, with these two modes, several combinations of user/device discovery requirements can be formulated in the same request in a very compact format. Building on this, we have developed a scheme allowing to provide rich combinations of desired parameters to be included along with resources, nodes, and parameters to be excluded when replying. Such a scheme is presented in Figure 3, where a EBNF representation is given for the case of CoRE link format. It should be noted, however, that our scheme is not tight to CLF and can be adopted to any other description format including JSON, EXI, CBOR to allow resource discovery in formats other than CLF.

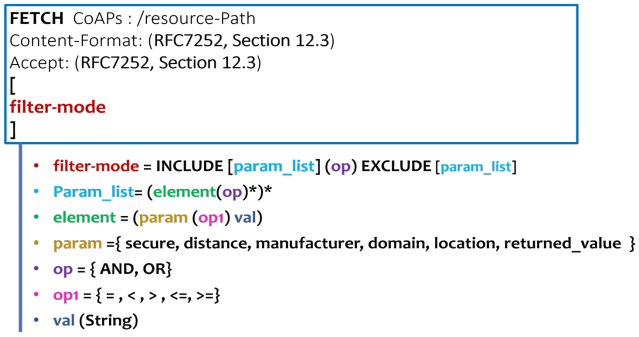


Fig. 3. Semantics of the FETCH-based resource discovery

In addition to the EXINC filtering technique, we propose to format the payload of the FETCH request in such a way as to take logical operators into account when specifying parameter lists to INCLUDE or EXCLUDE. For instance, a client may specify the parameters to be included AND/OR those to be excluded. It can also formulate the request to only get a specific part of the description in a specific media type. In this context, implementations can formulate a request payload of any media type that is compatible with the semantics of FETCH-based resource discovery, detailed in Figure 3. Finally, Figure 4 presents an example of a FETCH discovery request formulated in accordance of the proposed scheme.

To further highlight the importance of the proposed FETCH-based resource discovery, we employ it in a home automation scenario illustrated in Figure 5. In this scenario, the smart air-conditioner must get the average temperature of the house from the thermostats located in different rooms. However, this air conditioner only trusts in Nest thermostats that are secured with Thread and are located less than five hops away. The realization of such a scenario is not possible with GET-based discovery as defined in [9]. On the other hand, with FETCH, we can filter the results by using the proposed EXINC scheme. The FETCH request for this scenario along with the resolution process and returned results are shown in Figure 6.

Finally, the proposed FETCH-based discovery opens up promising prospects for significantly reducing network traffic, congestion, collisions, and power consumption during resource discovery. Note that the proposed technique is expandable to take into account more parameters and more combinations between these parameters in the same query. It is also designed to be adaptable for future uses. In addition, it is as valid with direct (distributed) discovery as with the centralized approaches and any other CoAP-based discovery approach. Indeed, it only requires changing the formulation of the queries by using FETCH with the semantics given above.

IV. Performance Evaluation

This section details the performance evaluation of the proposed method when compared with the GET-based approach adopted by CoAP. It starts by detailing the methodology and used tools, before presenting the evaluation scenarios and measured metrics along with discussing the obtained results.

A. Evaluation tools and Implementations

All our evaluations were carried out in the recently released Thread platform targeting IoT applications in home automation and similar environments. The choice of Thread is motivated by the fact that is a proven secure and reliable solution. It is, also, already implemented in many commercial products on sale for several years now. Moreover, this secure and open network protocol is built on a collection of existing standards similarly to environments such as Contiki [15]. Finally, a user can employ a smartphone, an application, and/or any device to communicate, directly or via the Cloud, with the Thread network.

Based on the open-source implementation of Thread, dubbed OpenThread [16], and inspired by Contiki, we have added a resource discovery layer over CoAP as specified in Figure 7. This layer contains three main components; namely: the request engine implementing the semantics of both GET and FETCH look-ups; the publication engine that is responsible on resource registration; and the resource description component. Subsequently, we developed prototypes of nodes that implement client and/or CoAP server along with the RD.

B. Evaluation protocol and scenarios

To evaluate our solution, we created different home automation and similar network configurations. Each configuration

```
FETCH CoAPS: ./well-known/
Content-Format: application/link-format
Accept: application/link-format
[
 [INCLUDE]: {secure=[Sval_1],Distance[=;<;>;<=;>=]Dval_1,location=[Nodes_List;],Domain=[domain],Manufacturer=[name]}
 [EXCLUDE]:
{secure=[Sval_2],Distance[=;<;>;<=;>=]Dval_2,location=[Nodes_List;],Domain=[domain],Manufacturer=[name]}
]
```

Fig. 4. Example of the proposed scheme for FETCH

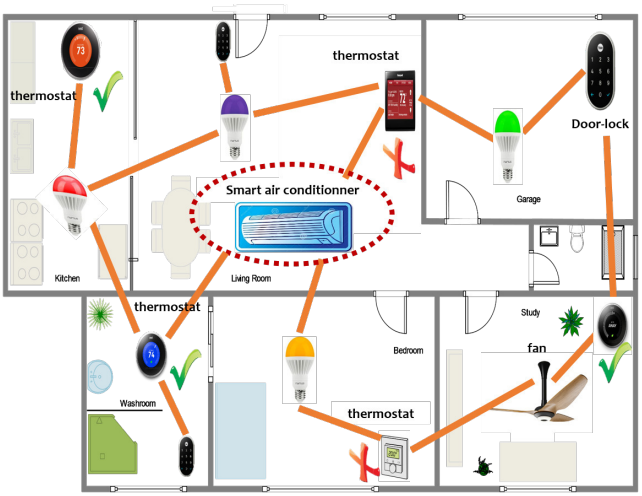


Fig. 5. A use-case of FETCH-based resource discovery

contains 13 nodes over which we run several single-hope and multi-hop scenarios for both centralized and distributed cases. Below are the details of multi-hop scenarios.

- Centralized discovery (CoAP-RD) in multi-hop networks: in this test, we designed a scenario where three servers will register their resources in the RD, then two concurrent clients consult the RD to obtain the descriptions of registered resources. To do so, clients send their requests in unicast, using the Thread unicast routing protocol, to the RD. This latter will respond with unicast messages containing the requested parts.
- Distributed discovery in multi-hop networks: in this test, one client sends a multicast request to discover required resources. This request will be propagated in the network by the multicast routing protocol MPL [17]. Once this request arrives to a node having the desired resources, it will respond directly to the client with a unicast response.

The above test cases were performed by simulations using virtual instances of Thread objects. Hence, environment-related parameters such as signal propagation, influence of obstacles and interference with other wireless signals are not taken into account. Also, the estimation of parameters, such as the consumed energy will not be possible. For each test case, we set the simulation time to 600 seconds and varied the request frequency. To put the results in context, we compared our FETCH-based discovery approach with the standard GET-based approach under the following performance metrics.

- Average discovery time: this parameter is defined as the average waiting time between the transmission of a request and the reception of the first response averaged over several requests. This metric is used to evaluate the efficiency of our solution in terms of latency.
- Discovery success rate: measures the number of received responses to all sent queries. This metric is used to evaluate the reliability of the proposed technique.
- Size of request/reply messages: accumulate the size of request/reply messages. It is defined as the ratio between the sum of request/reply sizes for each node over the total number of nodes.

By comparing our approach with that of GET under different networks and discovery scenarios, we aim to encompass most of the performance indicators of the proposed approach. Simulation parameters are summarized in Table I and obtained results are discussed in the following subsection.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Duration of one simulation	600 seconds
Number of iterations	5
Number of nodes	13
type of nodes	Thread CLi Instances
Message payload	Variable size
Network layer	IPv6 over 6LoWPAN
MAC layer	802.15.4 with enabled MAC security

C. Results and discussions

This section discusses the obtained results in both centralized and distributed scenarios.

1) Centralized approach: In this section, we will discuss the results obtained for the centralized approach of our FETCH-based solution (CoAP-RD-FETCH) and that of GET (CoAP-RD-GET). Obtained results are depicted in Figure 8.

Figure 8.(a) presents the average time of discovery of both approaches when varying the request frequency. As can be seen from this figure, discovery with FETCH achieved a lower discovery time compared to that of the GET method. This can be explained by the fact that GET generates more traffic since the response message contains descriptions of all available resources in the RD. It takes a longer time to reach the client because of its size and the congestion created in the network. However, the FETCH method performs efficient filtering so

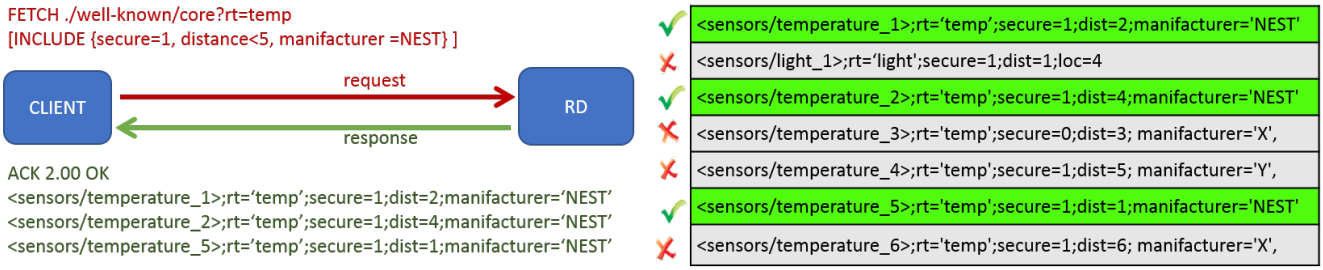


Fig. 6. Details of a FETCH-based resource discovery process

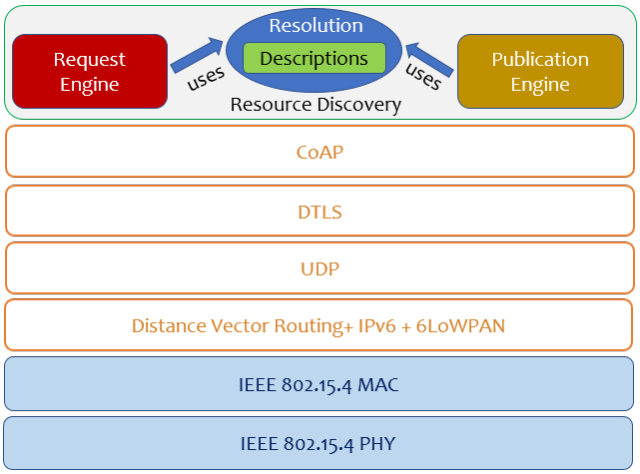


Fig. 7. Resource discovery implementation in OpenThread

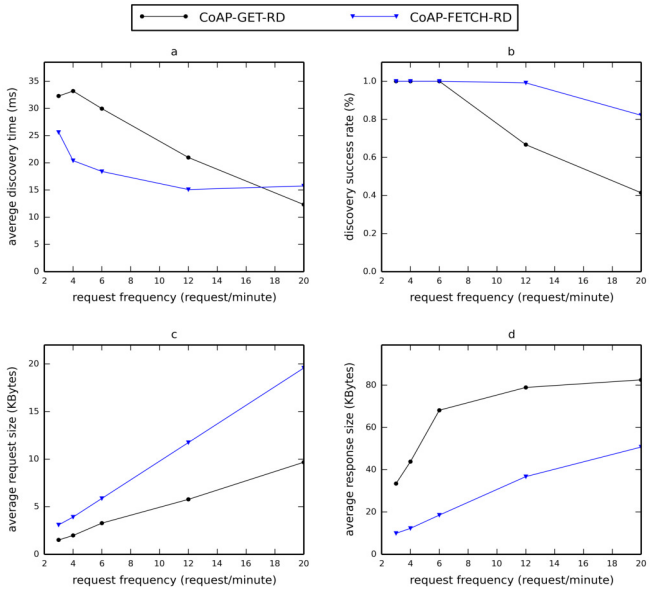


Fig. 8. Evaluation of proposed mechanisms in a unicast discovery scenario

that the RD only returns the desired resource descriptions, which significantly reduces the size of the response messages and also allows for faster routing (less congestion).

Concerning the average discovery success rate, Figure 8.(b) shows that both GET and FETCH registered a high success rate approaching 100 % in low request frequencies. This is due to the reliable routing protocol deployed in Thread along with the fact that the simulation environment is considered perfect. With a high query frequency, however, the discovery success rate becomes very low, approaching 40% for the GET method. This could be explained by the very high congestion of the network caused by the high frequency of query generation and the size of GET responses that may even lead to the elimination of responses at the transmission buffer. However, with the FETCH method, we notice that the rate does not decrease too much and approaches 80% at the highest frequency thanks to the minimized size of returned responses.

With regards to the size of requests, it is clear from Figure 8.(c) that the queries generated by FETCH are slightly larger compared to those generated by GET. This is due to the extra data that FETCH needs in its payload to specify the filter that will be used during the discovery process. This has the advantage of minimizing response size by only returning the desired resources. Knowing that a single query can match several large responses, this surplus provides an acceptable compromise. Indeed, it improves the relevance of responses along with the discovery time and network utilization as can be seen from Figure 8.(d). For instance, this figure shows that the difference in size, between the responses generated by GET and those generated by FETCH, is very important. This is due to the fact that, with FETCH, the RD only returned description parts that exactly matched user's specifications. This ensures both user satisfaction concerning discovery relevance and time as well as network fluidity with regards to congestion and resource utilization.

2) Distributed approach: This subsection discusses the results obtained on the distributed resource discovery scenario with the two discovery approaches: GET (CoAP-GET-multicast) and FETCH (CoAP-FETCH-multicast). Obtained results are depicted in Figure 9.

As can be seen from Figure 9.(a) FETCH-based resource discovery achieved noticeably lower discovery time compared to that achieved by GET-based discovery. This difference reaching up to 130 ms, shows the power of FETCH especially for multicast traffic, which is slow and expensive in terms of time and energy. With regards to the average discovery success

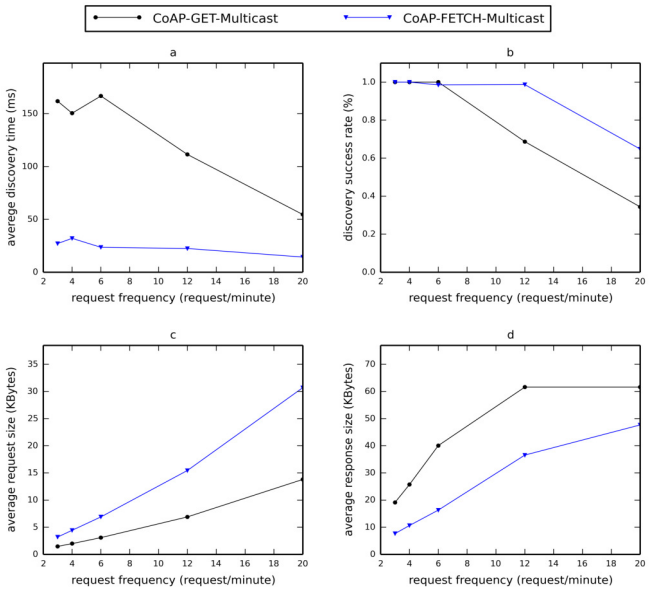


Fig. 9. Evaluation of the proposed mechanisms in a multicast discovery scenario

rate, Figure 9.(b) shows that both methods are equally reliable in both centralized and distributed approaches through the use of the MPL, which ensures the efficient routing of messages.

Finally, and concerning the size of generated messages, it is clear from Figure 9.(c) that the size of the multicast requests sent by FETCH is slightly larger than those of the GET. These results are similar to those of the centralized approach because it is the same client that sends the same requests. On the other hand, the size of returned responses is noticeably smaller in FETCH-based discovery in comparison with GET as can be seen from Figure 9.(d). Similarly to the centralized case, this is due to the fact that with FETCH the servers only returned the descriptions of the adequate resources.

By analyzing these results, we can confirm that the use of FETCH method for the sake of resource discovery is very important and outperforms GET in many aspects regarding the granularity, efficiency, and relevance of discovery along with resource utilization. Such performance is equally efficient and effective for both approaches (centralized and distributed). Therefore, an in-depth elaboration of the FETCH-based specification will open up more advanced and more efficient prospects for resource discovery in the IoT.

V. Conclusion

In this paper, a new CoAP-based discovery mechanism was proposed building on the newly standardized FETCH method. Obtained results demonstrated the capacities of FETCH to achieve fine-grained, expressive and efficient discovery when compared with the GET-based discovery adopted by CoAP. These achievements open up new horizons to formulate a compact and expressive resource discovery framework for the IoT. Our future work ports on the generalization of FETCH-based resource discovery to encompass a wide-range of IoT look-ups in view of proposing a specification of IoT resource discovery based on FETCH.

References

[1] G. Montenegro, N. Kushalnagar, J. Hui, and D. Culler, "Transmission of ipv6 packets over ieee 802.15.4 networks," RFC 4944, RFC Editor, September 2007. http://www.rfc-editor.org/rfc/rfc4944.txt.

[2] P. Thubert, A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, and R. Alexander, "Rpl: Ipv6 routing protocol for low power and lossy networks," *RFC 6550*, 2012.

[3] Z. Shelby, K. Hartke, and C. Bormann, "The constrained application protocol (coap)," RFC 7252, RFC Editor, June 2014. http://www.rfc-editor.org/rfc/rfc7252.txt.

[4] R. T. Fielding, *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, Irvine, 2000.

[5] P. Stok, C. Bormann, and A. Sehgal, "Patch and fetch methods for constrained application protocol (coap)," Internet-Draft draft-ietf-core-etch-04, IETF Secretariat, November 2016. http://www.ietf.org/internet-drafts/draft-ietf-core-etch-04.txt.

[6] S. Cheshire and M. Krochmal, "Dns-based service discovery," RFC 6763, RFC Editor, February 2013. http://www.rfc-editor.org/rfc/rfc6763.txt.

[7] Z. Shelby, M. Koster, C. Bormann, and P. V. der Stok, "Core resource directory," Internet-Draft draft-ietf-core-resource-directory-10, IETF Secretariat, March 2017. http://www.ietf.org/internet-drafts/draft-ietf-core-resource-directory-10.txt.

[8] M. Liu, T. Leppanen, E. Harjula, Z. Ou, A. Ramalingam, M. Ylianttila, and T. Ojala, "Distributed resource directory architecture in machine-to-machine communications," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2013 IEEE 9th International Conference on*, pp. 319–324, IEEE, 2013.

[9] Z. Shelby, "Constrained restful environments (core) link format," RFC 6690, RFC Editor, August 2012. http://www.rfc-editor.org/rfc/rfc6690.txt.

[10] T. A. Butt, I. Phillips, L. Guan, and G. Oikonomou, "TRENDY: An adaptive and context-aware service discovery protocol for 6lowpans," in *Proceedings of the third international workshop on the web of things*, p. 2, ACM, 2012.

[11] J. Maenpaa, J. J. Bolonio, and S. Loreto, "Using RELOAD and CoAP for wide area sensor and actuator networking," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 121, 2012.

[12] C. Jennings, B. Lowekamp, E. Rescorla, S. Baset, and H. Schulzrinne, "Resource location and discovery (reload) base protocol," RFC 6940, RFC Editor, January 2014.

[13] B. Djamaa, A. Yachir, and M. Richardson, "Hybrid CoAP-based resource discovery for the Internet of Things," *Journal of Ambient Intelligence and Humanized Computing*, Feb. 2017.

[14] P. van der Stok, C. Bormann, and A. Sehgal, "Patch and fetch methods for the constrained application protocol (coap)," RFC 8132, RFC Editor, April 2017.

[15] "The official git repository for contiki." [Online] Available: https://github.com/contiki-os/contiki.

[16] "Openthread github website." [Online] Available: https://github.com/openthread/openthread.

[17] J. Hui and R. Kelsey, "Multicast protocol for low-power and lossy networks (mpl)," RFC 7731, RFC Editor, February 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 645–649
DOI: 10.15439/2018F232
ISSN 2300-5963 ACSIS, Vol. 15

Using Publish/Subscribe for Short-lived IoT Data

Frank T. Johnsen
Norwegian Defence Research Establishment (FFI)
P.O. Box 25
2027 Kjeller, Norway

***Abstract*—Efficient distribution of IoT sensor data requires one-to-many communication, for which publish/subscribe is a better communication approach than request/response. In this paper, the goal is to identify the/those publish/subscribe protocol(s) that are best suited for IoT data. The premise is that data should be as fresh as possible. Hence, the metric is end-to-end delay and the recommended approach is the solution that yields the lowest delay under the test conditions. Raspberry Pi 3 was used as the testbed, since it is representative as an IoT platform. The protocols evaluated are: AMQP, MQTT, MQTT-SN, STOMP, WSN, and XMPP, as well as using a mediation service to translate between them.**

I. Introduction

The term Internet of Tings (IoT) can be traced back as early as 1999 when Kevin Ashton used it to describe a network that linked physical "stuff" to the Internet. Nevertheless, it would be a few years before "IoT" became an active research area and the buzzword it is today. There are many different interpretations of what IoT is, but the core idea stems from Ashton. A more recent and elaborate definition of IoT is as follows:

> The Internet of Things (IoT) describes the revolution already under way that is seeing a growing number of internet enabled devices that can network and communicate with each other and with other web-enabled gadgets. IoT refers to a state where Things (e.g. objects, environments, vehicles and clothing) will have more and more information associated with them and may have the ability to sense, communicate, network and produce new information, becoming an integral part of the Internet. A widespread Internet of Things has the potential to transform how we live in our cities, how we move, how we develop sustainably, how we age, and more. – From [1]

The reason why IoT has become commonplace over the last five years is that a number of factors that can be considered mandatory precursors to this phenomenon have come into place:

- Cheap sensors (easily accessible from eBay, deal extreme, etc).
- Cloud computing (serves as a backend for IoT systems and can handle big data)
- Powerful smartphones (often used as a consumer's control panel in the IoT context)

Given these precursors, there have been many business ideas in the healthcare sector, logistics and other areas that give rise to a number of applications that fuel the current IoT trend. IoT as a concept is definitely relevant in a defense context. An example of this is the pioneer work described in [2], which deals with the use of sensor networks and lightweight processing platforms that require low power. IoT includes several disciplines, as one needs networking, embedded hardware, software architectures, sensors, information management, data analysis and visualization to fully leverage the concept. A key component within IoT is the use of distributed online devices that communicate using Internet protocols. A "thing" in IoT may be any device that is able to communicate, gather data or offer some kind of control. With this wide interpretation of "things", IoT may include, but is not limited to: Vehicles, appliances, medical equipment, power grids, transport infrastructure and production equipment.

Military organizations can exploit IoT deployed in battlefields and operational theaters to improve situational awareness, mission performance and achieve information superiority [3]. Within NATO, the Research Task Group (RTG) IST-147 "Military Application of Internet of Things" is investigating how to best employ IoT in a coalition force, particularly in the context of augmenting situational awareness in military operations in smart cities [4].

Today there is a great focus on using Commercial off-the-shelf (COTS) products where possible because it is considered a cost-effective way of acquiring a capability. This idea is well rooted in NATO, and has been considered foundational for an effective Network Enabled Capability (NEC) as identified in the NATO NEC Feasibility Study [5]. In this study it was also pointed out that the principles of service orientation must be taken into account when building distributed systems. These observations can be continued within the IoT venture, as there will also be a need to build large, efficient and interoperable systems.

NATO has identified a set of *Core Services*, which provide common communication functionality that other services (e.g., C2 services) depend upon. An example of a Core Service is messaging, which includes both request/response and publish/subscribe services. In this paper, the focus is publish/subscribe services as applied to support IoT. Here, the focus is on short-lived IoT data, in other words data that comes fresh from a sensor and needs to be delivered as soon as possible. Publish/subscribe is considered the most efficient communication paradigm for this type of data, in contrast to long-lived data, where the new approach of Information-Centric Networking offers some desirable properties [6].

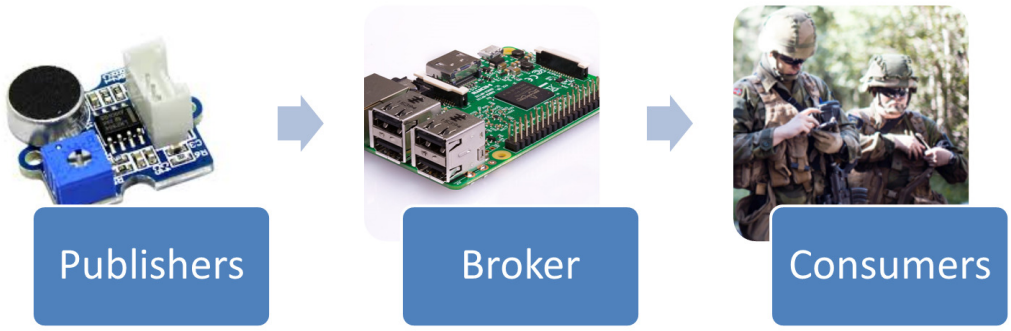


Fig. 1. Publish/subscribe information flow. It is assumed that subscriptions have been set up prior to this occurring.

II. Publish subscribe protocols

The publish/subscribe pattern implies that a consumer explicitly signals its interest in a given type of data by registering a subscription. The most common approach to signal such an interest is through a *topic*, i.e., a string that is used to identify the data a consumer is after. When new data is available on a certain topic, all consumers that have expressed interest in that topic receive it. This pattern is particularly well suited for IoT, as many sensors produce information more or less periodically, and thus, a push-pattern can reduce network activity considerably when compared to using a request/response or pull-pattern. A *broker* is used between the producer and consumer. Its tasks include subscription management and message dissemination according to topics, so that the producer only has to send new data to the broker, which then handles all further dissemination. For an illustration of the publish/subscribe pattern, see Fig. 1.

A number of publish/subscribe standards exist. This paper considers several of the most commonly available standards today. A short overview of the protocols follows.

A. Message Queue Telemetry Transport (MQTT)

MQTT is a popular publish/subscribe protocol for IoT, standardized as ISO/IEC PRF 20922 [7]. It provides publish/-subscribe messaging for resource-constrained devices: Low processing power, low memory, as well as network constraints. MQTT is designed to function well over unreliable networks by providing three levels of Quality of Service (QoS): Level 0, “at most once”-semantics – messages are delivered on a “best effort” basis. As MQTT is based on TCP, this is usually enough to ensure delivery. However, if the TCP connection is broken there will be no retransmission later on reconnection with this QoS level. So, though not likely, message loss can occur. Level 1 provides “at least once”-semantics, where messages are assured to arrive but duplicates can occur, hence systems must be able to handle duplicate packets. Level 2 gives “exactly once”-semantics, so messages are assured to arrive exactly once. This latter method requires an exchange of four packets, and decreases performance of the broker.

B. MQTT for Sensor Networks (MQTT-SN)

MQTT-SN is, in short, a version of MQTT that is optimized specifically for sensor networks [8]. The major difference is that it uses UDP as the underlying transport protocol rather than TCP.

C. Advanced Message Queuing Protocol (AMQP)

AMQP [9] is a binary wire protocol, which was designed as a reliable and interoperable open replacement for existing proprietary messaging middleware. As the name implies, it provides a wide range of features related to messaging, including reliable queuing, topic-based publish-and-subscribe messaging, flexible routing, transactions, and security. AMQP has been shown to be scalable and reliable, and is much used for civilian applications, notably for supporting financial transactions and also as a backbone in cloud computing clusters.

D. Web Services Notification (WSN)

NATO has chosen WSN [10] for publish/subscribe in its SOA baseline [11]. WSN is a part of the family of SOAP Web services standards. SOAP services promote interoperability, but being based on XML the cost is increased overhead compared to other protocols.

E. Simple/Streaming Text Oriented Messaging Protocol (STOMP)

STOMP [12] is text-based, making it somewhat similar to how HTTP operates. The main design principle was to create something simple to use and understand. However, the STOMP flavor of topics (called a *destination* in STOMP) is not mandated in the protocol specification, meaning that different brokers may support it differently. This, in turn, lowers interoperability across brokers, since publishers and consumers that function well with one broker implementation may not work with another. If you don’t encounter portability issues, then STOMP is simple, lightweight, and offers a wide range of language bindings.

Fig. 2. The Raspberry Pi 3B single board computer.

F. Extensible Messaging and Presence Protocol (XMPP)

Just like WSN, XMPP [13] is an XML-based protocol. Unlike WSN, XMPP does not use SOAP, so it does away with the extra abstraction layer (and thus extra overhead). XMPP is most known as a chat (instant messaging) and presence protocol, however it does offer additional features as well, like SIP-compatible multimedia signaling for voice, video, file transfer, and other applications as well as publish/subscribe functionality. XMPP aims to be the main competitor to MQTT for civilian IoT applications, and is, as such, interesting to compare with MQTT to see which protocol is "best".

III. Testbed setup

The testbed was put together of two Raspberry Pi 3B single board computers. The main motivation for using this as the testbed was that the Raspberry Pi 3B is a somewhat capable yet cheap computer that is representative for IoT development. The board is shown in Fig. 2. The technical specifications of the board are as follows [14]:

- Broadcom BCM2837 64bit ARMv7 Quad Core Processor powered Single Board Computer running at 1.2GHz
- 1GB RAM
- BCM43143 WiFi
- Bluetooth Low Energy
- 40pin extended GPIO
- 4x USB 2 ports
- 4 pole Stereo output and Composite video port
- Full size HDMI
- CSI camera port for connecting the Raspberry Pi camera
- DSI display port for connecting the Raspberry Pi touch screen display
- Micro SD port for loading your operating system and storing data

One Raspberry Pi 3B functioned as the client: That is, it set up subscriptions to pre-determined topics up front, published messages to said topics, and ran the consumers that received the messages. This was done so that time measurements across publishers and consumers (via the broker) would be accurate, since timestamps would originate from one and the same node rather than several, where clock skew could become an issue. The second Raspberry Pi 3B offered the protocol brokers and the mediation service to translate between protocols. All publishers and consumers were implemented using Java, and the respective protocols' native Java libraries. The brokers and mediation service (let us call this component a *multi-protocol broker*) was also Java software. In fact, the multi-protocol broker was an extended version of the federation mechanism described here [18], enhanced for the purpose of this paper to support all the protocols discussed above. For networking, 100 Mbps Ethernet was used, and both Raspberry Pi 3B's were connected to a switch. This was done to ensure the best possible networking conditions during the tests, so that local disturbances and interference should not affect the results, which could have been an issue if using e.g., WiFi.

Tests were executed as follows:

1) The multi-protocol broker was started.
2) A subscription to a topic was set up for a particular protocol α.
3) A publisher was initiated to fire off a burst of 100 messages over protocol β.
4) Having received 100 messages, the consumer terminated its subscription to protocol α.
5) The duration of steps 3-4 above was measured.
6) Steps 1-5 above were repeated for all α, β of protocol permutations.

IV. Results

Tab I shows the results when transmitting 100 messages. This table shows when the publisher sends 100 consecutive messages via the multi-protocol broker. The consumer receives the messages, and terminates the subscription after message number 100. The time (in seconds) of this entire burst of messages was measured here.

We see that WSN is the overall loser when considering our protocol delay metric. Consistently WSN shows the highest delays here. We see that having WSN as either the publisher or subscriber results in a high delay, with an even higher delay exhibited (just over 20s – the highest in the test) when both publisher and subscriber used WSN. This can be attributed to the SOAP layer used in the protocol; having this extra abstraction layer on top of HTTP does have an impact performance wise.

Of the other protocols, the performance difference is not so large when considering publisher/subscriber pairs of the same protocol (no translation is involved – indicated in **bold** in the table). Here, we see that AMQP is the "worst" (just above 4.5s) and STOMP is the "best" (just above 2.5s). This is understandable when thinking about the fact that AMQP provides a reliable message queue. Messages are ensured delivery and acknowledged in the queue. STOMP does not perform this added value service. MQTT achieves slightly better results than MQTT-SN, which again is slightly more efficient than XMPP (just over 3s). At first glance this may seem strange, since MQTT-SN is based on UDP which inherently has lower overhead than TCP, which is the underlying transport in

TABLE I
PROTOCOL DELAY PUBLISHING AND RECEIVING 100 MESSAGES.

	AMQP Sub.	**MQTT Sub.**	**MQTT-SN Sub.**	**STOMP Sub.**	**WSN Sub.**	**XMPP Sub.**
AMQP Publisher	**4.577313**	5.054525	5.065977	3.529224	16.309429	4.443499
MQTT Publisher	3.022697	**2.655891**	2.303168	2.448346	14.992253	4.470155
MQTT-SN Publisher	3.438891	2.716955	**2.895711**	2.466580	14.462532	4.340074
STOMP Publisher	3.017157	3.215914	4.876303	**2.256608**	15.458013	4.347493
WSN Publisher	12.518343	11.745184	12.116381	11.534597	**20.769935**	13.644167
XMPP Publisher	7.165085	6.414229	6.492992	6.086224	16.023309	**3.001673**

MQTT. The reason why MQTT-SN has slightly higher delays here, is that it is actually implemented as a gateway that uses plain MQTT in the backend. Hence, MQTT-SN gets a slight performance impact going through this gateway which moves it from UDP to TCP and vice versa locally on the broker node (UDP is used between client and broker across the network).

The remaining rows in the table show the different protocols' delay where the impact of translating between them is also included. We see that XMPP and WSN here show the cost of translating to/from XML based protocols. WSN has the impact of using both XML and SOAP, whereas XMPP only uses XML. An interesting observation is the above mentioned performance of the XMPP publisher/subscriber pair, which shows that the XML-based XMPP does not perform too badly when no translation is involved.

V. RELATED WORK

The NATO IST-090 RTG has demonstrated the use of WSN at the tactical level. WSN has the benefit of being a NATO recommended standard for information exchange in a coalition environment. However, it is a resource heavy protocol and its application at the tactical level requires applying proprietary optimizations [15].

More recently, the IST-118 RTG conducted initial experiments comparing different publish/subscribe approaches on tactical broadband radios. Namely, WSN, MQTT and AMQP were investigated in a preliminary small-scale study [16]. Here, MQTT was found to be a very lightweight alternative to the other two protocols when applied in the tactical network. Currently, the IST-150 RTG is continuing work where IST-118 left off, and is considering MQTT specifically for use in soldier systems on the tactical level [17].

In [18], the authors provided a solution for federating between different publish/subscribe protocols, i.e., WSN, MQTT, and AMQP. The work in this paper uses an extended version of that open source implementation with support for additional protocols as the broker and mediation service (aka multi-protocol broker) in the testbed.

VI. CONCLUSION

If you need to use UDP from your sensor to the mediation service, then your choice is MQTT-SN, which is the only one based on UDP of the protocols tested.

If you need interoperability with NATO (i.e., you need WSN subscribers), then the most efficient protocols to use when going through the mediation service are MQTT-SN (UDP) and MQTT (TCP). You definitely don't want your sensors to deliver data using WSN directly, as that is the protocol with the overall highest delay.

If you are free to choose any protocol you want for the entire network, then using STOMP as both publisher and receiver was marginally quicker than using any of the other protocols. However, if you need advanced capabilities like multi broker meshing, then MQTT or AMQP can offer that, though they had slightly larger delays than STOMP. Though never best, XMPP shows overall favorable results given that it is based on XML. It goes to show that it is possible to implement a somewhat efficient XML based protocol, in comparison with WSN, which has very high delays. The reason for this is that WSN, being based on SOAP, adds an extra abstraction layer in the protocol that none of the other protocols have.

The overall recommendations are summarized in Tab. II.

TABLE II
RECOMMENDATIONS FOR PUBLISH/SUBSCRIBE COMBINATIONS.

GOAL	**Publisher**	**Subscriber**
Lowest overall delay	STOMP	STOMP
UDP necessary	MQTT-SN	MQTT-SN
NATO Interoperable	MQTT or MQTT-SN	WSN
Meshable brokers	MQTT or AMQP	MQTT or AMQP

The table gives an overview of which protocol combinations to use to achieve a specific goal. Note that where the lowest delay is not the primary goal, it is considered the secondary goal when giving the recommendations.

VII. FUTURE WORK

The testbed consisted of a publisher and subscriber running on one Raspberry Pi and the broker on another. Hence, it is only a small scale test of how the protocols perform. One of the challenges of IoT is the scale, so for future work it would be interesting to make similar tests of a larger scale setup.

Also, it would be beneficial to test the protocol performance over a typical IoT networking technology, such as LoRa. Other relevant networking options would be 4G and WiFi, to name a couple.

REFERENCES

[1] IoT Special Interest Group. Technology Strategy Board. 2013.

[2] Wind River Systems. The Internet Of Things For Defense. White Paper, 2015.
[3] Niranjan Suri et al. Analyzing the Applicability of Internet of Things to the Battlefield Environment. IEEE ICMCIS 2016, Brussels, Belgium, May 2016.
[4] Frank T. Johnsen, Zbigniew Zielinski, Konrad Wrona, Niranjan Suri, Christoph Fuchs, Manas Pradhan, Janusz Furtak, Bogdan Vasilache, Vincenzo Pellegrini, Michal Dyk, Michal Marks, and Mateusz Krzyszton. Application of IoT in Military Operations in a Smart City. IEEE ICMCIS 2018, Warsaw, Poland, 22nd – 23rd May 2018.
[5] P. Bartolomasi, T. Buckman, A. Campbell, J. Grainger, J. Mahaffey, R. Marchand, O. Kruidhof, C. Shawcross, and K. Veum. NATO network enabled capability feasibility study. Version 2.0, October 2005.
[6] A. Carzaniga, M. Papalini, and A. Wolf. Content-based Publish/Subscribe Networking and Information-centric Networking. Proceedings of the ACM SIGCOMM workshop on Information-centric networking, ACM, 2011.
[7] ISO/IEC 20922:2016. Information technology – Message Queuing Telemetry Transport (MQTT) v3.1.1. ISO/IEC JTC 1 Information technology. Publication date: June-2016. https://www.iso.org/standard/69466.html
[8] Andy Stanford-Clark and Hong Linh Truong. MQTT For Sensor Networks (MQTT-SN) Protocol Specification. Version 1.2. November 14, 2013. http://mqtt.org/new/wp-content/uploads/2009/06/MQTT-SN_spec_v1.2.pdf
[9] RabbitMQ. AMQP 0.9.1 protocol specification. https://www.rabbitmq.com/resources/specs/amqp0-9-1.pdf
[10] OASIS. WSN specifications. https://www.oasis-open.org/committees/wsn/
[11] Consultation, Command and Control Board (C3B). CORE ENTERPRISE SERVICES STANDARDS RECOMMENDATIONS: THE SOA BASELINE PROFILE VERSION 1.7. Enclosure 1 to AC/322-N(2011)0205, NATO Unclassified releasable to EAPC/PFP, 11 November 2011.
[12] STOMP Protocol Specification, Version 1.2 http://stomp.github.io/stomp-specification-1.2.html
[13] XMPP is the open standard for messaging and presence. https://xmpp.org/
[14] Farnell.com. Raspberry Pi 3 Model B. http://www.farnell.com/datasheets/2020826.pdf
[15] IST-090. SOA Challenges for Real-Time and Disadvantaged Grids, Final Report of IST-090. AC/323(IST-090)TP/520. NATO. Published April 2014
[16] IST-118. IST-118 SOA recommendations for Disadvantaged Grids: Tactical SOA Profile, Metrics and the Demonstrator Development Spiral. Paper presented at the SCI-254 Symposium on “Architecture Assessment for NEC”. 14-15 May, 2013 in ESTONIA.
[17] Marco Manso, Frank T. Johnsen, Ketil Lund, and Kevin Chan. Using MQTT to Support Mobile Tactical Force Situational Awareness. IEEE ICMCIS 2018, Warsaw, Poland, 22nd – 23rd May 2018.
[18] Eirik Bertelsen et al. Federated publish/subscribe services. 9th IFIP International Conference on New Technologies, Mobility & Security 26 to 28 February 2018. Paris, France.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 651–658
DOI: 10.15439/2018F194
ISSN 2300-5963 ACSIS, Vol. 15

Identifying Hidden Influences of Traffic Incidents' Effect in Smart Cities

Attila M. Nagy
Department of Networked Systems and Services
Budapest University of Technology and Economics
Magyar Tudósok krt 2., Budapest, Hungary
Email: anagy@hit.bme.hu

Vilmos Simon
Department of Networked Systems and Services
Budapest University of Technology and Economics
Magyar Tudósok krt 2., Budapest, Hungary
Email: svilmos@hit.bme.hu

***Abstract*—The road network of big cities is a complex and hardly analyzable system in which the accurate quantification of interactions between nonadjacent road segments is a serious challenge. In this paper we would like to present a novel method able to determine the effects (the time delay and the level of the correlation) of distinct road segments on each other of a smart city's road network. To reveal these relationships, we are investigating unexpected events such as traffic jams or accidents. This novel analysis can give a significant insight to improve the operation of currently widespread traffic prediction algorithms.**

I. Introduction

NOWADAYS smart city services are becoming more widespread than ever as cities are growing and becoming more and more crowded as a result of urbanization and growth of the world population. The rapid progress of urbanization improved life of many people, but also brought remarkable challenges, like traffic congestions that can lead to increased energy/fuel consumption and enormous emission of pollutants [1]. These phenomena heavily and directly impact the health, the life quality and expectancy of city dwellers. According to [2], laboratory studies indicated that transport-related air pollution may increase the risk of developing allergies and can exacerbate symptoms, particularly in susceptible subgroups, while [3] showed that traffic jams increase the risk of heart attacks.

Intelligent management systems, such as Advanced Traffic Management System (ATMS) and Intelligent Transportation System (ITS) can help overcome or significantly reduce the impact of such negative effects on city dwellers. Forecasts of these systems can support traffic control centers in managing the road network and allocating resources systematically, for example opening/closing lanes, pricing dynamically parking places or adapting the traffic lights to the current traffic trends.

In vehicle navigation the knowledge of traffic forecasts for different routes during the route planning is advantageous, as the devices will be able to calculate more efficient routes and reduce travel time. Insight into vehicular flows of smart cities could make searching for parking spaces much easier and faster. It could also provide added value for emerging V2X based traffic control systems, which can play an important role in route planning of self-driving cars.

In the literature, there are numerous prediction models utilized for traffic flow prediction, however the road network of big cities is a complex and hardly analyzable system in which unexpected events could significantly decrease the correctness of the result of the prediction models.

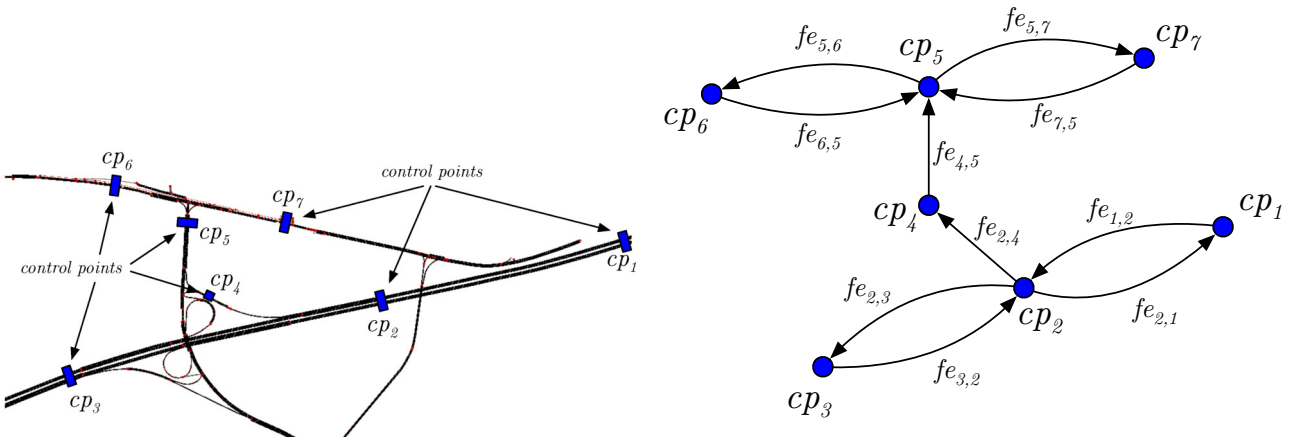
Every predicted value is composed of a predictable component and an error [4], which includes both prediction error and unpredictability of uncertainty. Thus the predictable value is derived from the deterministic part and the predictable part of uncertainty. Therefore it follows, that the predictability of the traffic flow depends on whether the model is able to predict the uncertainty part or not. Fortunately, lots of prediction models can be prepared for handling the uncertainty part by integrating different external data sources. The uncertainty is influenced by many factors, like weather condition, mass events, road constructions, road closures, accidents etc. By considering these external environmental factors, the error of prediction models can be decreased.

In this paper, a new method will be presented, which aims to reduce the previously mentioned uncertainties. The algorithm focuses on unexpected events, such as traffic accidents, which can have a negative impact on the traffic prediction. By investigating the effect of these phenomena, the algorithm is able to:

- identify neighboring road segments that could be affected by the event
- to determine the level of correlation between the road segment affected by the accident and its neigboring road segments,
- to quantify the time delay: the time needed for the effects of the accident to propagate to neighboring road segments.

By fusing this information with real-time traffic information, the prediction models will be able to provide more accurate predictions even in the case of an unexpected event happening nearby.

The paper is organized as follows. In Section II., various prediction techniques are introduced aimed at reducing the error of prediction models. This section also contains a short summary of usable data sources. In Section III, the algorithm is presented in detail, which is followed by a case study simulation in Section IV. Section V concludes the paper and points out the current weaknesses and future improvements of the algorithm.



(a) A road network with control points

(b) The flow graph model of the road network

Fig. 1: An example of the traffic graph model interpretation. On Figure 1a, the original road network is depicted with control points, while Figure 1b shows the road network's graph representation.

II. Traffic Prediction Methods in Smart Cities

A. Prediction Techniques

There are several proposed prediction models [5], which perform well for regular conditions, however an unexpected event could significantly decrease the quality of their traffic forecasts. The reason is that the predictability of the traffic flow depends on whether the model is able to predict the uncertainty part of the traffic flow with the required precision. The uncertainty is influenced by many factors, like weather, events, road constructions, lighting conditions etc. Incorporating external environmental factors and fused data [6] to the model is crucial to decrease the error of the prediction and increase the predictable part of uncertainty.

Traffic is predictable in the sense that it does not vary significantly during weekdays and during most months of the year [7]. In [8] similar results were obtained as well as they have found a relatively high daily predictability of traffic conditions despite the absence of any apriori knowledge of drivers' origins and destinations and the quite different travel patterns between weekdays and weekends.

A neurowavelet prediction algorithm was proposed [9] to forecast the hourly traffic flow considering the effect of rainfall. In the article, the authors use a stationary wavelet transform to reveal correlation between different weather conditions and changes in the traffic flow. An examination was carried out [10] whether or not road usage on a particular location determines the impacts of various weather conditions. The study showed that the precipitation, cloudiness and wind speed reduce traffic intensity, while high temperatures and hail significantly increase traffic intensity.

Other papers contentrated on the spatio-temporal property of the traffic flow, by using different Autoregressive Integrated Moving Average (ARIMA) model variants [11]–[13], applying K-Nearest Neighbors (KNN) models [14], [15] while others employed Convolutional Neural Network (CNN) for this purpose [16]–[18]. Deep learning based prediction model was also presented for spatio-temporal data [17]. The prediction model uses spatial and temporal relations and integrates global information (such as day of week, meteorological conditions, etc.) to decrease the uncertainty.

Relation between traffic predictability and prediction time horizon was investigated [4] by examining spatio-temporal traffic relationship using Cross-correlation function (CFF). The time lag calculated by CFF can be used to determine prediction time horizon, and the cross-correlation coefficient can be utilized to identify the spatial relations that can be used in prediction.

Solutions enumerated in this section aim to handle the previously listed uncertainties of traffic flows. However, we have not found significant work aimed at increasing the predictability of traffic flow through the investigation of unexpected events like accidents or traffic congestions of uncontrolled traffic flow. In this paper, we will present a novel method, which can exploit these events to measure the time delay and calculate the level of influence between distinct road segments.

B. Data sources

In the first generation of ATMSs and ITSs [19] the utilized data sources were various presence sensors in fixed positions, able to detect the presence of nearby vehicles. Initially inductive loop detectors were the most popular, but nowadays a wide variety of sensors became available [20].

Recently, the advent of GPS equipped smartphones and vehicles has given rise to a relatively new type of data source that could supplement presence type sensors to gather more detailed information or get data about roads, which have not been covered with presence sensors yet.

Our method can leverage both types of data sources, but in the case of GPS traces, a preprocess step is needed to be inserted before the data model building.

III. Methodology

The road network of big cities is a complex and hard to analyze system in which the accurate quantification of interactions between nonadjacent road segments is almost an insolvable objective, because of the unique decisions of thousands of drivers which makes the interactions invisible. However an unexpected event could be used to reveal hidden connections between road segments, because they always appear as an outlier in the timeseries of the investigated data type (traffic speed, traffic flow count, travel time, etc.). It follows, that if the emergence of an unexpected event is known, the ripple of that event can be observed through the network, thus the hidden correlations will become observable. As an analogy, we suppose that one can think of a road network as a huge black box system, which has one input and numerous outputs. If the system is fed with an unusual input, the inner behavior of the system can be inferred through the outputs.

In this section, the Algorithm for Identifying Hidden Influences (AIHI) will be introduced in detail, which targets to exploit unexpected traffic events to reveal the previously mentioned connections. In Subsection III-A, the graph based data model will be presented which is suitable for the analysis of unexpected events, then the distinct steps of the algorithm will be explained in Subsection III-B.

A. Data model

In our solution, the road network is modeled with a $FG = (CP, FE)$ directed flow graph. Each $cp \in CP$ node represents a control point, which is a special point of the road network, where the traffic measurement is feasible by different type of traffic sensors (such as inductive loop detectors, radar sensors, audio sensors, etc.). A cp control point itself does not store any traffic data, they just measure the traffic flow at their position.

The $fe \in FE$ directed flow edges represent a link between two adjacent control points (cp_{src}, cp_{dst}), where there is at least one lane between the two control points in the spreading direction of the traffic. To every fe directed flow edge, a ts time series is assigned, which stores historical traffic flow data. The ts time series of fe flow edge will contain those measurements, which are provided by the cp_{dst} destination control point of fe. For instance, if there is a directed flow edge between cp_1 (source) and cp_2 (destination) control points denoted by $fe_{1,2}$, then the $fe_{1,2}$ edge will contain the measurements of cp_2 control point.

Besides the data from fixed position sensors, the data model is also able to utilize GPS traces, if virtual control points are defined and trace data is aggregated in these points.

On Figure 1, an example of the traffic graph interpretation is depicted, it shows how the data model have to be interpreted on a simple road structure. Figure 1a illustrates a simple road network with control points, which are marked by cp_i identifiers. On Figure 1b, the directed flow graph of the previous road network is visualized. The cp_i identifiers on this figure are identical with the identifiers of the ones on Figure 1a, and there are $fe_{x,y}$ directed flow edges in the graph, if cp_x and cp_y control points are connected in the road network in the spreading direction of the traffic.

We also have to deal with the timeliness of our model. The different fixed position or GPS sensors sample the measured data type with different frequences based on their settings. Consequently the traffic analysis requires a homogeneous sampling frequency. Different aggregating time intervals are used in the literature for this purpose, which mainly depend on the task at hand.

Generally, narrow intervals, for example 10 seconds, are meaningless and really noisy. We have found that the most common time intervals are in minutes (5-10 minutes) [21], [22], but there are also many papers claiming that longer time intervals would be more effective, like quarter or half an hour [16], [23]. For our model, we chose a 30 seconds time interval because too long time intervals could hide important features of an unexpected event and the noisiness of the 30 seconds scale does not affect the correctness of the AIHI.

B. The steps of the algorithm

1) Initialization: Let us denote a time series of $fe_{src,dst}$ flow edge by $ts_{src,dst}$ in which data can be described as an ordered sequence of discrete measurements:

$$ts_{src,dst} = \{m_t\} \qquad t = 1, 2, \dots T \tag{1}$$

where m_t is the measured traffic count value at time t at the cp_{dst} control point.

The entry point of the algorithm will be an fe directed flow edge, for which the associated ts time series shows unexpected event between an arbitrary (t_{start}, t_{end}) time interval. The (t_{start}, t_{end}) time interval contains an unexpected event, if a statistically significant change can be detected in the behavior/shape of the ts time series between t_{start} and t_{end} compared to the ts_{hist} historical average, or in other words, ts time series contains an anomalous part compared to the ts_{hist} historical average.

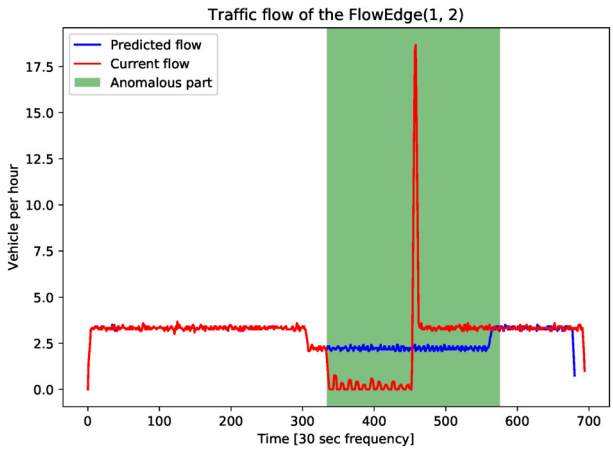


Fig. 2: Visualization of anomaly behavior compared to Predicted flow

Input :
- FG: Directed flow graph of the road network
- fe: The investigated directed edge
- t_{start}: The start time of the unexpected event

Output: The effects of the event organized in a tree structure

```
1  job_pool ← [];
2  visited_edges ← [];
3  ts_an = findAnomaly(fe.ts, t_start);
4  addJob({fe, ts_an, t_start, forward});
5  while job_pool is not empty do
6      j = getNext(job_pool);
7      addVisitedEdge(j.fe);
8      (b, i) = bestFitLag(j.ts_an, j.fe.ts, j.t_start,
         j.dir);
9      is_anomaly, next_ts_an =
         findAnomaly(j.ts, nj.t_start + b);
10     if is_anomaly then
11         s_neighs = getEdges(j.fe.cp_src);
12         foreach edge in s_neighs do
13             if j.fe.cp_src is edge.cp_dst and edge not
                 in visited_edges then
14                 addJob({edge, next_ts_an,
                     t_start + b, backward});
15             end
16         end
17         d_neighs = getEdges(j.fe.cp_dst);
18         foreach edge in d_neighs do
19             if j.fe.cp_dst is edge.cp_src and edge not
                 in visited_edges then
20                 addJob({edge, next_ts_an,
                     t_start + b, forward});
21             end
22         end
23     end
24 end
```

Algorithm 1: AIHI algorithm

There are two methods to discover such time intervals. The easiest way when we have apriori knowledge about the occured unexpected events like accidents or road closures as these can be used directly as the input of the AIHI algorithm. The other possible approach is to execute an extensive search in the raw historical dataset for unexpected events. This approach requires a classification model that is able to decide whether a $ts_{src,dst}$ time series contains an unexpected event or not.

The AIHI algorithm needs the following three inputs:

- The FG flow graph of the road network
- The fe flow edge, which was the source of the unexpected event
- The t_{start} time of the emergence of the unexpected event on fe

Utilizing these three inputs, the algorithm will follow the effect of the unexpected event through the traffic flow graph and determining those flow edges that could be affected by the input event.

2) Processing of a job: AIHI algorithm 1 uses a job pool based approach in which the whole investigation task is separated to smaller independent subtasks. In this case, a subtask is responsible for the investigation of whether the currently examined fe flow edge's ts time series is affected by the source unexpected event or not. To run a job, the following elements are necessary:

- An fe flow edge
- The ts_{an} time series, which contains the whole anomaly part identified by the source job
- The t_{start} emergence time of the anomaly in the source job
- The dir direction of the spread of the event (forward or backward), because the unexpected events of the road network can have an effect in both directions.

To start the algorithm, the first job will be created from the entry point. The entry point contains all necessary job input parameters, except the ts_{an} time series, thus by using the findAnomaly function (in Section III-B4), the anomalous part of the ts time series of fe flow edge have to be calculated. After that, the execution of the algorithm can be started and the processing of jobs will be continued until the pool has been emptied.

A job will execute these steps:

1) Find the best fitting $best\_lag$ time lag between the ts time series of fe flow edge and ts_{an} anomalous time series from t_start in the chosen dir direction (see Algorithm 2)
2) Check that an anomalous event can be identified from $best\_lag$ or not (see Algorithm 3)
3) If an anomalous event is detected during the second step, then the adjacent flow edges of the investigated fe flow edge are put into the job pool as new jobs in which the source job will be the current job

3) Find best lag: The BestFitLag function is responsible for determining the start of an anomalous event in the ts time series of the current $job.fe$ flow edge.

We can assume that the shape of the anomalous time series ts_{an} is quite similar to the anomaly observed in the actual ts time series of $job.fe$. To measure this similarity we defined a new distance function, called shape_dist (Equation 2). Contrary to other distance function like Manhattan, Euclidean or DynamicTimeWarping [24], it measures the similarity the time series' shape by differentiating changes in the different time series:

$$shape\_dist(x, y) = \sqrt{\sum_{i=1}^{n}((x_i - x_{i-1}) - (y_i - y_{i-1}))^2} \quad (2)$$

The shape_dist function is calculated with increasing $lag = 0, 1, 2, ...$ between ts time series and ts_{an} time series. It can

Input :
- ts_{an}: The time series containing the whole anomaly part
- ts: The time series of the fe flow edge
- t_{start}: The emergence time of the anomaly in the time series
- dir: The direction of the spread of the event (forward or backward)

Output:
- $bestlag$: The best fitting time lag
- $influence$: The level of influence

$last \leftarrow$ `dist` $(ts_{an},$ `shift` $(ts, t_{start}))$;
if dir *is* $forward$ **then**
  $i \leftarrow 1$;
end
if dir *is* $backward$ **then**
  $i \leftarrow -1$;
end
$lastlag \leftarrow 0$;
while `dist` $(ts_{anom},$ `shift` $(ts, t_{start} + i)) \leq last$ **do**
  $lastlag \leftarrow lastlag + i$;
end
$bestlag = i$;
$length = len(ts_{an})$;
$influence = \frac{1}{1+exp(\frac{1}{length} * last)}$;

Algorithm 2: BestFitLag

be assumed, that while the delay is increasing, the distance between the two time series will decrease until the best fit is reached. Thus if the calculated distance values start to increase, the possible best delay has been reached, because the adjacent road segments show high correlation in general. However sometimes the calculated delay can be just a local minimum, thus a simulate annealing is applied to find the global optimum.

Furthermore, the best distance value can be used to express the influence between the two flow edges, however a transformation is required, converting the shape_dist function's $[0, \inf)$ domain to $[1, 0]$ domain. Higher values mean a stronger influence, while lower values mean a weaker influence. Equation 4 is designed for this purpose:

$$distance = shape\_dist(ts_{an}, shift(ts, lastlag)) \quad (3)$$

$$influence = \frac{1}{1 + exp(\frac{1}{length} * distance)} \quad (4)$$

, where the $length$ parameter equals with the length of ts_{an}.

4) Find anomaly: The findAnomaly function (Algorithm 3) is responsible for determining whether an anomalous part starting from t_{start} can be identified. If an anomalous part is found, the function also returns its length.

The findAnomaly function exploits the observation, that the error of a prediction model significantly increases when

Input :
- ts: The time series of the fe flow edge
- ts_{hist}: The historical average time series of the fe flow edge
- t_{start}: The emergence time of the anomaly in the time series

Output:
- $is\_anomaly$: The input ts time series contains anomaly from t_{start} or not
- $next\_ts_{an}$: The anomaly part in the input ts time series if it exists

$model \leftarrow$ `exp_smooth` (ts_{hist});
$error \leftarrow$ `measure_error` $(model, ts_{hist})$;
$error\_dist \leftarrow$ `norm_dist` $(error)$
$length \leftarrow 0$;
while $ts[t_{start} + length]$ *from* $error\_dist$ **do**
  $length \leftarrow length + 1$;
end
if $length$ *is* 0 **then**
  $is\_anomaly \leftarrow false$;
else
  $is\_anomaly \leftarrow true$;
$next\_ts_{an} \leftarrow ts[t_{start}, t_{start} + length]$

Algorithm 3: findAnomaly

anomalous behavior can be observed. The error of prediction was analyzed and we could conclude, that the prediction error has a normal distribution in case of normal behavior (depicted on Figure 3), while the distribution of the error significantly differs in case of anomalous behavior. Thus the operating principle of findAnomaly is:

1) Fitting a prediction model on ts_{hist}
2) Running until the prediction error of the model returns to a normal distribution

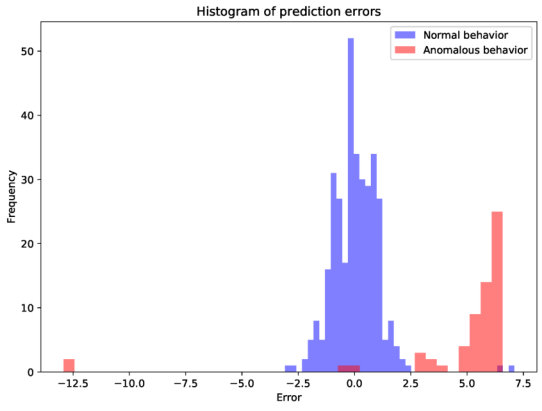


Fig. 3: Histogram of different prediction errors

Therefore for findAnomaly, a prediction model is constructed by applying a simple moving average on ts_{hist}, then

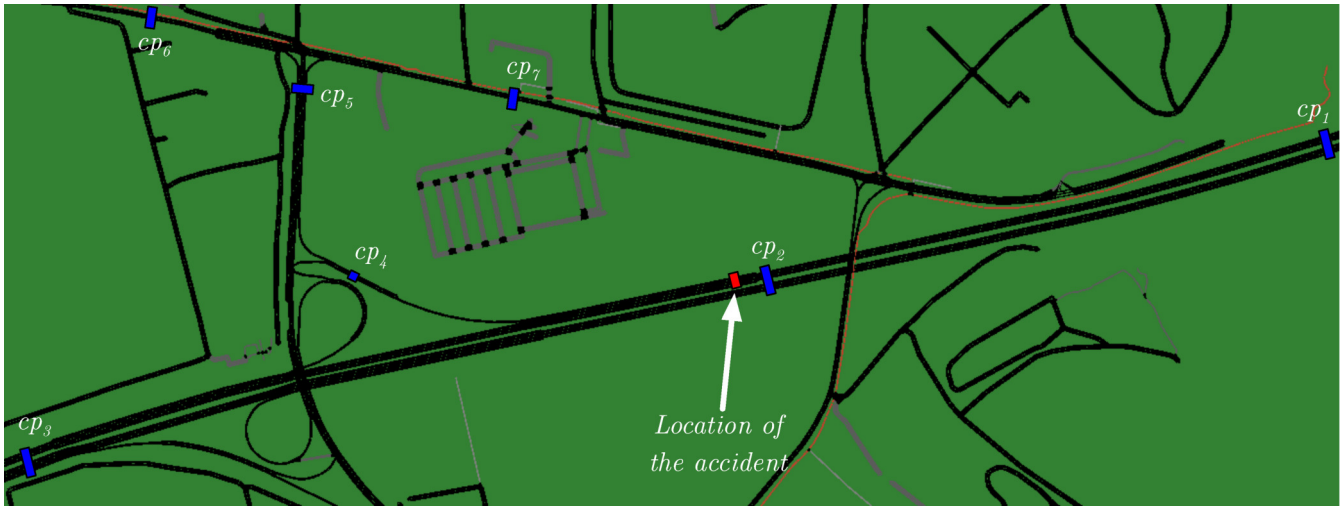


Fig. 4: The simulation map with location of traffic accident and control points

based on that model the distribution of the prediction error is determined. The algorithm starts from t_{start} and runs until the prediction error of the model returns to a normal distribution. This point in the time series is $length$ apart from t_{start}, therefore the length of the anomaly can be calculated. If the $length$ equals with zero, it means that ts does not contain an anomaly.

IV. Case study

After the implementation of the AIHI, in this section we will demonstrate its ability to follow the spread of the effect of a traffic accident through a traffic flow graph.

At first, we searched for datasets containing traffic flow data and traffic accidents as well, but unfortunately there was no such publicly available traffic flow dataset at the time of writing. Because of this, a simulation framework was used for the evaluation.

As the simulation framework, we chose Simulation of Urban Mobility (SUMO), a free and open traffic simulation suite, which is available since 2001. SUMO allows modeling of intermodal traffic systems, including roads, vehicles, public transport and pedestrians. Included with SUMO is a wealth of supporting tools, which handle tasks such as route finding, visualization, road network import from open street maps and emission calculation.

In our scenario, besides the real traffic flow data, a traffic accident had to be generated. The authors of [25] used traffic lights for this purpose, thus this approach had been applied in our simulation as well. The accident is simulated by opening only one of the four available lanes on a road.

In the simulation, high rank roads of Budapest's suburb are examined. Seven control points (using inductive loop detectors) were placed on the map as depicted on Figure 4. We simulated five hours of traffic flow, which was similar to afternoon rush hours. The vehicles are simulated with $speedFactor = normc(1, 0.1, 0.2, 2)$, which means that 95% of the vehicles drove between 80% and 120% of the legal speed limit. In the beginning we simulated normal traffic flow, then an one hour long traffic accident had been inserted after two and a half hours near to cp_2, so the effect of the accident could be identified first on $fe_{1,2}$ flow edge (Figure 2).

As mentioned in Subsection III-B1, the entry point of AIHI was $fe_{1,2}$ flow edge with t_{start} start time of the simulated traffic accident. The result of AIHI, is displayed on Figure 5. The spreading tree of the accident shows that the farther control points were, the bigger the detected time lags became, while the influences were decreasing as expected. The change of the pattern of the anomalous part was also visualised between cp_4 and cp_5 on Figure 6.

V. Conclusion

The road network of big cities is a complex and hard to analyze system in which the accurate quantification of interactions between nonadjacent road segments is almost an insolvable objective, because of the unique decisions of thousands of drivers which makes the interactions invisible. In this paper a novel algorithm (AIHI) has been presented, that is able to exploit unexpected traffic events to reveal the hidden connections between nonadjacent road segments and provides the following information:

- identify nearby road segments that could be affected by the event
- if a road segment is affected, the exact level of influence between the affected and accident road segments,
- the time delay: the time between the event and its detection on the affected road segments

Combining these new information sources with real-time traffic information, the accuracy of prediction models able to integrate external environmental variables can be increased.

We have demonstrated the capabilities of AIHI with simulations on a real road network which results can be depicted as spreading tree of the accident.

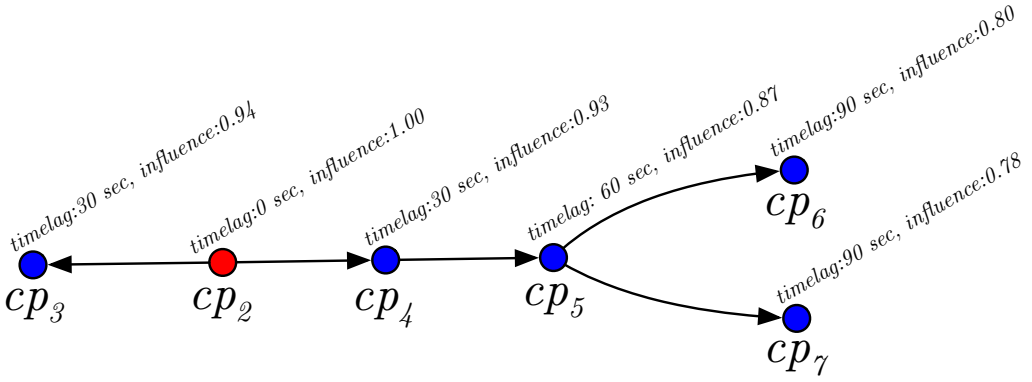
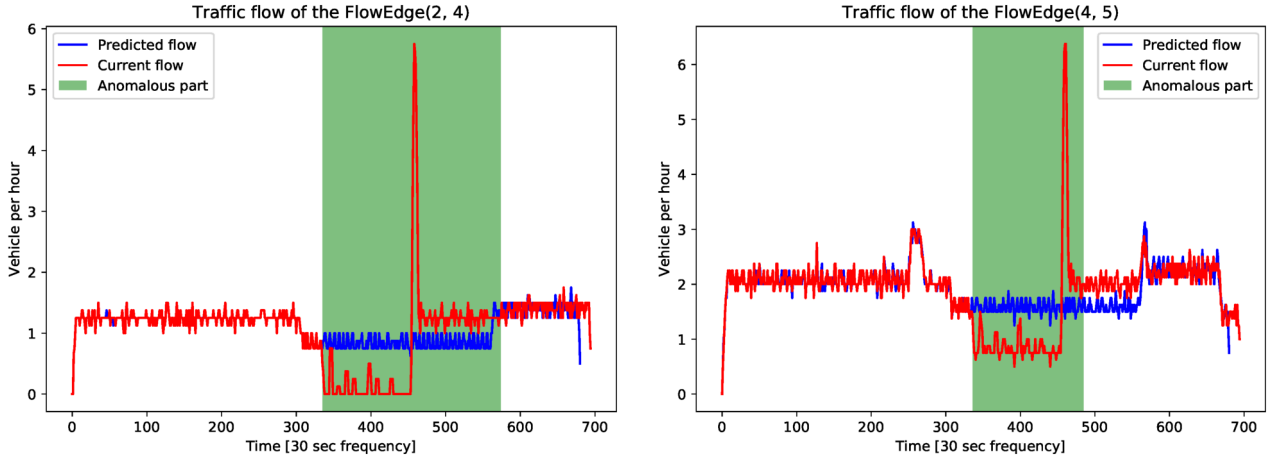


Fig. 5: Visualization of the AIHI's result. The examined control point was cp_2 (red dot), where the time lag is zero and the influence is 1.0. Other influenced control points in the road network identified by AIHI are marked with blue dots.



(a) Traffic flow of FlowEdge(2, 4)

(b) Traffic flow of FlowEdge(4, 5)

Fig. 6: On Figure 6a and Figure 6b the measured traffic flows are depicted at cp_4 and cp_5, respectively. On Figure 6b the effect of other connected roads can be seen, where the anomalous part starts to differ compared to the anomalous part of Figure 6a

.

In the future, we are planning to extend prediction models like Neural Networks (NN) and KNN to utilize the result of AIHI. We also want to develop a classification model able to decide whether a time series contains an unexpected event or not, because the input of parameters of AIHI are currently determined manually. Applying the classification model, the determination of AIHI's input parameters will be automated.

VI. Acknowledgment

The research reported in this paper was supported by the BME- Artificial Intelligence FIKP grant of EMMI (BME FIKP-MI/SC).

References

[1] J. He, S. Gong, Y. Yu, L. Yu, L. Wu, H. Mao, C. Song, S. Zhao, H. Liu, X. Li, et al., Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major chinese cities, Environmental pollution 223 (2017) 484–496.

[2] M. Krzyżanowski, B. Kuna-Dibbert, J. Schneider, Health effects of transport-related air pollution, WHO Regional Office Europe, 2005.

[3] A. Peters, S. Von Klot, M. Heier, I. Trentinaglia, A. Hörmann, H. E. Wichmann, H. Löwel, Exposure to traffic and the onset of myocardial infarction, New England Journal of Medicine 351 (17) (2004) 1721–1730.

[4] Y. Yue, A. G. Yeh, Y. Zhuang, Prediction time horizon and effectiveness of real-time data on short-term traffic predictability, in: Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, IEEE, 2007, pp. 962–967.

[5] A. Ermagun, D. Levinson, Spatiotemporal traffic forecasting: review and proposed directions, Transport Reviews (2018) 1–29.

[6] N.-E. El Faouzi, H. Leung, A. Kurian, Data fusion in intelligent transportation systems: Progress and challenges–a survey, Information Fusion 12 (1) (2011) 4–10.

[7] A. Stathopoulos, M. Karlaftis, Temporal and spatial variations of real-time traffic data in urban areas, Transportation Research Record: Journal of the Transportation Research Board (1768) (2001) 135–140.

[8] J. Wang, Y. Mao, J. Li, Z. Xiong, W.-X. Wang, Predictability of road traffic and congestion in urban areas, PloS one 10 (4) (2015) e0121825.

[9] S. Dunne, B. Ghosh, Weather adaptive traffic prediction using neu-

rowavelet models, IEEE Transactions on Intelligent Transportation Systems 14 (1) (2013) 370–379.
[10] M. Cools, E. Moons, G. Wets, Assessing the impact of weather on traffic intensity, Weather, Climate, and Society 2 (1) (2010) 60–68.
[11] Q. T. Tran, Z. Ma, H. Li, L. Hao, Q. K. Trinh, A multiplicative seasonal arima/garch model in evn traffic prediction, International Journal of Communications, Network and System Sciences 8 (04) (2015) 43.
[12] W. Min, L. Wynter, Real-time road traffic prediction with spatio-temporal correlations, Transportation Research Part C: Emerging Technologies 19 (4) (2011) 606–616.
[13] S. V. Kumar, L. Vanajakshi, Short-term traffic flow prediction using seasonal arima model with limited input data, European Transport Research Review 7 (3) (2015) 21.
[14] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, J. Sun, A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting, Transportation Research Part C: Emerging Technologies 62 (2016) 21–34.
[15] B. Yu, X. Song, F. Guan, Z. Yang, B. Yao, k-nearest neighbor model for multiple-time-step prediction of short-term traffic condition, Journal of Transportation Engineering 142 (6) (2016) 04016018.
[16] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, Y. Wang, Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction, Sensors 17 (4) (2017) 818.
[17] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, Dnn-based prediction model for spatio-temporal data, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2016, p. 92.
[18] Y. Li, R. Yu, C. Shahabi, Y. Liu, Graph convolutional recurrent neural network: Data-driven traffic forecasting, arXiv preprint arXiv:1707.01926.
[19] B. Singh, A. Gupta, Recent trends in intelligent transportation systems: a review, Journal of Transport Literature 9 (2) (2015) 30–34.
[20] Y. Yue, Traffic sensors, International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology (2016) 1–7.
[21] R. Fu, Z. Zhang, L. Li, Using lstm and gru neural network methods for traffic flow prediction, in: Chinese Association of Automation (YAC), Youth Academic Annual Conference of, IEEE, 2016, pp. 324–328.
[22] D. Xia, B. Wang, H. Li, Y. Li, Z. Zhang, A distributed spatial–temporal weighted model on mapreduce for short-term traffic flow forecasting, Neurocomputing 179 (2016) 246–263.
[23] Y. Tian, L. Pan, Predicting short-term traffic flow by long short-term memory recurrent neural network, in: Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on, IEEE, 2015, pp. 153–158.
[24] M. Müller, Dynamic time warping, Information retrieval for music and motion (2007) 69–84.
[25] J. Q. Shi, L. Cheng, Simulation and analysis of highway traffic accident based on vissim, in: Applied Mechanics and Materials, Vol. 253, Trans Tech Publ, 2013, pp. 1682–1685.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 659–663
DOI: 10.15439/2018F25
ISSN 2300-5963 ACSIS, Vol. 15

Performance Analysis of Slotted ALOHA Systems with Energy Harvesting Nodes and Retry Limit Using DTMC Model

Katsumi Sakakibara Yoji Nakata Kento Takabayashi
Department of Information and Communication Engineering
Okayama Prefectural University
111, Kuboki, Soja, Japan
Email: {sakaki, cd29029y, kent.hf}@c.oka-pu.ac.jp

***Abstract*—We analyze performance of slotted ALOHA systems with energy harvesting nodes and retry limit. We assume that the capacities of data and energy buffer at a node are one packet and E packets, respectively, and that one data packet transmission consumes one energy packet. The data and the energy packet arrival processes are modeled by independent Bernoulli processes. Under these assumptions, we develop a node-centric two-dimensional discrete time Markov chain model. Based on the equilibrium point analysis, we derive the fixed point equation with respect to the ratio of nodes transmitting a data packet. The accuracy of numerical results derived from the fixed point equation is verified by computer simulation. The numerical results indicate that throughput, the offered traffic and the discard probability roughly depend on the minimum of the data packet generation probability and the energy packet generation probability.**

I. Introduction

ENERGY harvesting techniques have been attracting researchers' interest in minimization of nodes by removing batteries. For example, in Wireless Sensor Networks (WSNs), a huge number of such battery-less tiny nodes may be dispersed in a wide area. Each node may harvest their energy from environment. When a huge number of nodes with bursty traffic contend with one another for a common communication channel such as WSNs, a Medium Access Control (MAC) protocol plays an important role which can greatly influence performance of networks with or without the use of energy harvesting techniques [1].

Performance of MAC protocols with energy harvesting nodes has been extensively investigated in the literature. Moradian and Ashtiani [2] analyzed the maximum stable throughput of slotted ALOHA systems consisting of finite number of energy harvesting nodes. They constructed a node-centric two-dimensional Discrete-Time Markov Chain (DTMC) model with (j, x), where j is the number of energy packets in the energy buffer and x is the elapsed time for the next retransmission. Foss et al. [3] discussed stability conditions of slotted ALOHA systems with infinite population of energy harvesting nodes from the system-centric viewpoint of a queueing network. The system is described by a two-tuple (q, v), where q and v are the total number of data and energy packets in the system, respectively. Bae [4] analyzed the delivery ratio of slotted ALOHA systems with energy harvesting nodes under delay constraints. A node-centric two-dimensional DTMC model (W, E) was constructed, where W is the elapsed sojourn time of the leading data packet in the data buffer and E is the number of energy packets in the energy buffer. Notice here that no retry limit of unsuccessful data packet is considered in the above literature [2], [3].

In this paper, we analyze performance of slotted ALOHA systems consisting of energy harvesting nodes with retry limit [5] using a node-centric two-dimensional DTMC model. Then, the performance is analyzed in terms of throughput, average transmission delay and discard probability of data packet due to excessive retransmission trials.

II. System Model

Consider a slotted ALOHA system consisting of N energy harvesting nodes contending for a common channel. Each node is equipped with not only a single data packet buffer but also an energy packet buffer of E-packet capacity. From the single-buffered assumption each node can store only one data packet. The length of data packet to be transmitted is assumed to be fixed to the unit length. The time axis is divided into slots which suffices for a single data packet transmission. The propagation delay between nodes and the common receiver is negligible. A data packet arrives independently at each node with probability λ in a slot. A node harvests an energy packet with probability ε in a slot. A node can transmit a data packet with probability p, if its energy buffer is not empty. One energy packet is consumed when a node transmits a data packet. We consider neither capture effects nor channel noise, so that a data packet transmission succeeds, only if no other data packets are transmitted simultaneously. All data packets involved in collision are to be retransmitted, until the number of retransmission trials including the first transmission reaches to the retry limit L. As an example, a system model with two energy harvesting nodes is depicted in Fig. 1.

This work was partly supported by Japan Society for the Promotion of Science under Grant-in-Aid for Scientific Research (C) (KAKENHI No. 25420379).

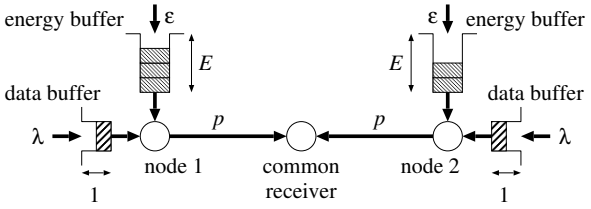


Fig. 1. System model of two energy harvesting nodes with single data buffer and with energy buffer of E-packet capacity.

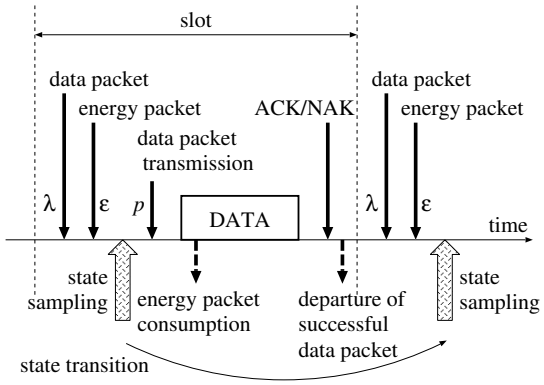


Fig. 2. Timing chart for state sampling.

III. Discrete Time Markov Chain Model

In order to accurately analyze the steady-state performance of the system, it is required to construct and solve a system-centric DTMC model. However, it demands a considerably high dimensional DTMC model, which is complex to solve. Here, we take advantage of an equilibrium point analysis (EPA) [6], [7], which approximately evaluate the steady-state performance by using a node-centric DTMC model with an assumption that each node operates in an independent manner.

A. State Sampling

Consider a certain node. Let i be the next number of transmission trials for a data packet in the node; $i = 0, 1, 2, \ldots, L$. Note that a node with $i = 0$ implies that it has no data packet. Let j denote the number of energy packets in the node; $j = 0, 1, 2 \ldots, E$. Then, from the single-capacity assumption for data packets, the state of each node in a slot can be described by a two-tuple (i, j).

In order to facilitate construction of a DTMC model with respect to the state of a node, we first define the order of the stochastic events in a time-slot. As addressed in the previous section, there may be four stochastic events in a slot; data packet generation with probability λ, harvest of an energy packet with probability ε, data packet transmission with probability p and outcome of data packet transmission; success or failure. At the end of a slot, a node which just transmitted a data packet receives a positive acknowledgment (ACK) or a negative acknowledgment (NAK).

B. Two-Dimensional DTMC Model

According to the sampling timing of the state, we can construct a node-centric two-dimensional DTMC model with respect to (i, j) for $i = 0, 1, \ldots, L$ and $j = 0, 1, \ldots, E$. States (i, j) for $i > 0$ imply that a node is backlogged; that is, a data buffer at a node is occupied. States (i, j) for $i > 0$ and $j > 0$ are those in which a node can transmit a data packet with probability p, since it has sufficient energy packets for data packet transmission.

C. State Transition Probabilities

State transition probability $p_{(i,j),(k,\ell)}$ from State (i, j) to State (k, ℓ) can be obtained by taking into consideration the four stochastic events between two consecutive sampling points in Fig. 2;

1) data packet transmission at a backlogged node with non-empty energy buffer,
2) outcome of data packet transmission; success or failure,
3) data packet generation at a node with empty data buffer,
4) arrival of an energy packet.

First, suppose that a node has no data packet; that is, a node is in State $(0, j)$ for $j = 0, 1, \ldots, E$. If no data packet arrives and if an energy packet arrives, then a node moves to State $(0, j + 1)$. We have state transition probabilities for thsese events;

$$p_{(0,j),(0,j+1)} = (1 - \lambda)\varepsilon \tag{1}$$

for $j = 0, 1, \ldots, E - 1$. If a data packet arrives without an energy packet, then state transition probabilities are

$$p_{(0,j),(1,j)} = \begin{cases} \lambda(1 - \varepsilon) & \text{for } j = 0, 1, \ldots, E - 1 \\ \lambda & \text{for } j = E \end{cases} \tag{2}$$

If both data and energy packets arrive, then we have

$$p_{(0,j),(1,j+1)} = \lambda\varepsilon \tag{3}$$

for $j = 0, 1, \ldots, E - 1$.

Next, suppose that a node is backlogged. A state transition from (i, j) to $(i, j + 1)$ occurs, if a data packet is not transmitted and if an energy packet arrives. Then, we have

$$p_{(i,j),(i,j+1)} = \begin{cases} \varepsilon & \text{for } j = 0 \\ (1 - p)\varepsilon & \text{for } j = 1, 2, \ldots, E - 1 \end{cases} \tag{4}$$

for $i = 1, 2, \ldots, L$. Note that no data packet can be transmitted if a node is in State $(i, 0)$, since it has no energy. A node moves from State (i, j) to State $(i+1, j-1)$, if a data packet transmission results in failure and if no energy packet arrives:

$$p_{(i,j),(i+1,j-1)} = pP_{\text{fail}}(1 - \varepsilon) \tag{5}$$

for $i = 1, 2, \ldots, L - 1$ and $j = 1, 2, \ldots, E$, where P_{fail} is the probability of transmission failure of a data packet, which is formulated later. If a node fails in data packet transmission and if an energy packet arrives, then state transition probabilities are

$$p_{(i,j),(i+1,j)} = pP_{\text{fail}}\varepsilon \tag{6}$$

for $i = 1, 2, \ldots, L-1$ and $j = 1, 2, \ldots, E$. If a data packet transmission succeeds and if no data packet and no energy packet arrive, then a node transits from State (i, j) to State $(0, j-1)$:

$$p_{(i,j),(0,j-1)} = \begin{cases} p(1-P_{\text{fail}})(1-\lambda)(1-\varepsilon) \\ \qquad \text{for } i = 1, 2, \ldots, L-1 \\ p(1-\lambda)(1-\varepsilon) \quad \text{for } i = L \end{cases} \tag{7}$$

for $j = 1, 2, \ldots, E$. Note that for $i = L$, a node moves from State (L, j) to State $(0, j-1)$ when it transmits a data packet irrespective of success or failure. If only an energy packet arrives without a new data packet after successful data packet transmission, then we have

$$p_{(i,j),(0,j)} = \begin{cases} p(1-P_{\text{fail}})(1-\lambda)\varepsilon & \text{for } i = 1, 2, \ldots, L-1 \\ p(1-\lambda)\varepsilon & \text{for } i = L \end{cases} \tag{8}$$

for $j = 1, 2, \ldots, E$. If a new data packet is generated without an energy packet generation after successful data packet transmission or after data packet discard, state transition probabilities are

$$p_{(i,j),(1,j-1)} = \begin{cases} p(1-P_{\text{fail}})\lambda(1-\varepsilon) \\ \qquad \text{for } i = 1, 2, \ldots, L-1 \\ p\lambda(1-\varepsilon) \quad \text{for } i = L \end{cases} \tag{9}$$

for $j = 1, 2, \ldots, E$. If an energy packet is generated for the above case, then we have

$$p_{(i,j),(1,j)} = \begin{cases} \lambda p(1-P_{\text{fail}})\varepsilon & \text{for } i = 1, 2, \ldots, L-1 \\ \lambda p\varepsilon & \text{for } i = L \end{cases} \tag{10}$$

for $j = 1, 2, \ldots, E$.

Finally, the state transition probabilities for trivial state transitions are obtained as

$$p_{(i,j),(i,j)} = 1 - \sum_{(k,\ell)\neq(i,j)} p_{(i,j,),(k,\ell)} \tag{11}$$

for $i = 0, 1, \ldots, L$ and $j = 0, 1, \ldots, E$.

D. Steady-State Probabilities

Let us denote the steady-state probability of State (i, j) by $\pi_{(i,j)}$ for $i = 0, 1, \ldots, L$ and $j = 0, 1, \ldots, E$. Then, according to the theory of Markov chains, the steady-state distribution $\{\pi_{(i,j)}\}$ can be obtained by solving a system of linear equations;

$$\begin{bmatrix} \vdots \\ \pi_{(m,n)} \\ \vdots \end{bmatrix} = \begin{bmatrix} & \vdots & \\ \cdots & p_{(i,j),(m,n)} & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} \vdots \\ \pi_{(i,j)} \\ \vdots \end{bmatrix} \tag{12}$$

and

$$\sum_{i=0}^{L} \sum_{j=0}^{E} \pi_{(i,j)} = 1. \tag{13}$$

IV. Performance Analysis

A. Fixed Point Equation

Let τ denote a ratio of transmitting nodes tentatively. Since a backlogged node with one or more energy packets transmits its data packet with probability p, we have

$$\tau = p \sum_{i=1}^{L} \sum_{j=1}^{E} \pi_{(i,j)} \tag{14}$$

As shown in the previous subsection, the steady-state distribution $\{\pi_{(i,j)}\}$ is a function of the probability of transmission failure P_{fail}, which can be formulated as

$$P_{\text{fail}} = 1 - (1-\tau)^{N-1} \tag{15}$$

from the assumption of the independent operation of nodes underlaid in EPA. Here, a combination of (14) and (15) together with (1)–(13) provides a fixed point equation with respect to τ for given N, L, E, λ, ε, and p, which can be numerically solved.

Once we obtain the value of τ, we can evaluate various performance measures as follows.

B. Throughput

The offered traffic is the average number of nodes which are transmitting their data packet in a slot. It follows from the independent operation assumption of nodes that

$$G = N\tau. \tag{16}$$

Then, we can evaluate the throughput as the average number of successful nodes per slot;

$$S = (1-P_{\text{fail}})G = N\tau(1-\tau)^{N-1}. \tag{17}$$

C. Average Transmission Delay

According to Little's result [7], the average transmission delay can be obtained as the ratio of the average number of backlogged nodes to the average number of nodes departing from the backlogged states. In the steady-state, the average number of backlogged nodes is given as

$$B = N \sum_{i=1}^{L} \sum_{j=0}^{E} \pi_{(i,j)}. \tag{18}$$

Nodes can depart from the backlogged states due to successful data packet transmission or discard of their data packet experiencing an excessive transmission failures. The average number of successful nodes per slot is given by (17). On the other hand, a data packet is discarded, if data transmission from a node in State (L, j) results in failure for $j = 1, 2, \ldots, E$. The average number of discarded packets per slot is evaluated as

$$N_d = NpP_{\text{fail}} \sum_{j=1}^{E} \pi_{(L,j)}. \tag{19}$$

Therefore, we can obtain the average transmission delay as

$$D = \frac{B}{S + N_d}. \tag{20}$$

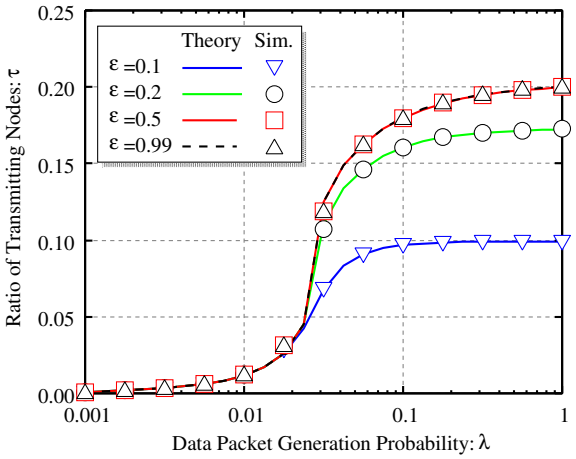


Fig. 3. Comparison of the ratio of transmitting nodes between the fixed point analysis and computer simulation for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$.

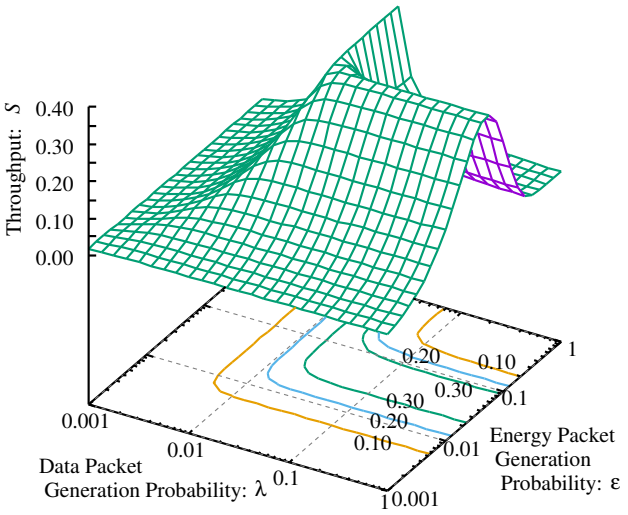


Fig. 4. Throughput for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$.

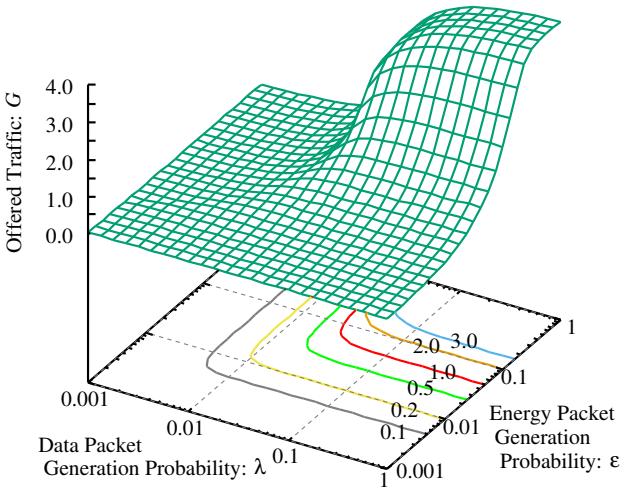


Fig. 5. Offered traffic for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$.

D. Discard Probability of Data Packet

When we impose the retry limit on a data packet, the probability that a generated data packet is discarded is important performance measure. A data packet is released from data buffer at a node because of successful transmission or compulsory discard. Thus, the ratio

$$P_d = \frac{N_d}{S + N_d} \tag{21}$$

provides the discard probability of a data packet.

V. Numerical Results

A. Accuracy Verification

Since EPA assumes the independent operation of nodes, it is required to verify the accuracy of the derived results. As shown in the previous section, various performance measure can be evaluated from the numerical result of τ obtained by solving the fixed point equation (14). Here, we examine the accuracy of our analysis via the ratio of transmitting nodes τ.

The analytical and computer simulation results are shown in Fig. 3 for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$. From Fig. 3 it is clear that the analysis using EPA offers sufficiently accurate numerical results. Also, it can be found that the ratio of transmitting nodes is independent of $\varepsilon \geq 0.5$

B. Performance Measure

Based on the fixed point equation (14), the numerical results for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$ in terms of throughput (17), the offered traffic (16), the average transmission delay (20) and the discard probability of data packet (21) are shown in Fig. 4, Fig. 5 Fig. 6 and Fig. 7, respectively, as a function of the data packet generation probability λ and the energy packet generation probability ε.

From Fig. 4, we can observe that the shape of throughput surface exhibits a weak symmetric relationship between data packet generation probability λ and energy packet generation probability ε. That is, we can recognize that throughput roughly depends on $\min[\lambda, \varepsilon]$. A backlogged node can transmit no data packet, unless it has one or more energy packets. Conversely, a node with empty data buffer can transmit no data packet, even if it has one or more energy packets. This relationship results in weak symmetry of throughput between λ and ε. It is widely confirmed that throughput of slotted ALOHA systems is maximized when the offered traffic is one data packet per slot. In fact, it follows from Fig. 5 that the offered traffic which achieves the maximum throughput in Fig. 4 is around one data packet per slot; $G \approx 1.0$. Comparing Fig. 5 to Fig. 4, we can find that the shape of the contours projected on the λ-ε plain is closely related. Also, we can perceive that the well-known relationship $S = Ge^{-G}$ approximately holds between Fig. 4 and Fig. 5.

Next, from Fig. 6, both the shapes of the surface of the average transmission delay and the contours on the λ-ε plain differ from those of those in Fig. 4 and Fig. 5. We can observe no symmetry between λ and ε. Transmission delay

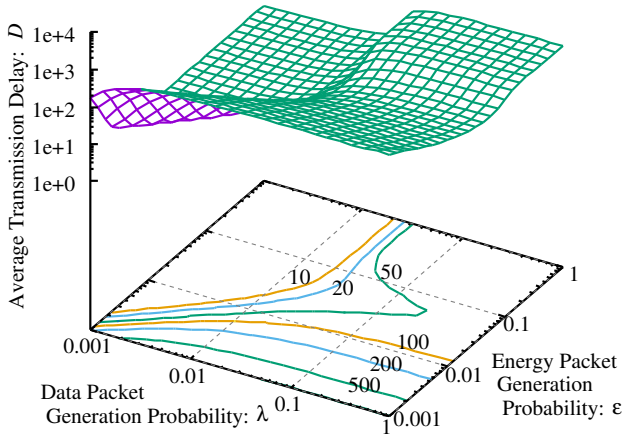


Fig. 6. Average transmission delay of data packet for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$.

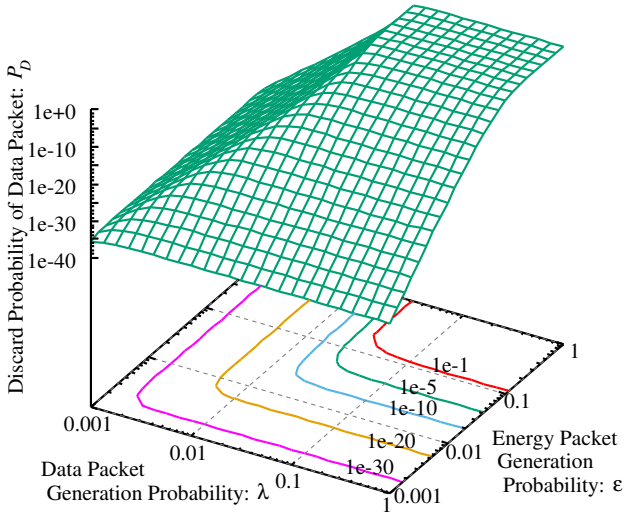


Fig. 7. Discard probabiltiy of data packet for $N = 20$, $L = 20$, $E = 5$ and $p = 0.2$.

is defined as the elapsed time-slots between the data packet generation and its departure from the system due to successful transmission or discard. Hence, transmission delay increases for small ε, since the generated data packet has little chances to be transmitted due to the lack of the energy despite the data packet generation probability λ. For small λ the average transmission delay tends to decrease according to an increase of ε. This is because the data packet has more chances to be transmitted for large ε and it has less possibility to collide with other simultaneously transmitted data packets because of small λ. For given $\lambda > 0.05 = 1/N$, the average transmission delay has its minimum at around $\varepsilon = 0.05 = 1/N$. For small $\varepsilon < 0.05$, the average transmission delay increases due to less chances for transmission. On the other hand, for $\varepsilon > 0.05$, increment of the offered traffic, as shown in Fig. 5, results in more packet collisions, so that the average transmission delay is enlarged.

Finally, the contour of the discard probability of data packet P_D in Fig. 7 exhibits the same tendency as throughput in Fig. 4 and the offered traffic in Fig. 5. However, in contrast to Fig. 4 and Fig. 5, the discard probability is rapidly degraded several orders of magnitude against small fluctuation of λ and ε even if λ and ε are sufficiently small.

VI. Conclusion

In this paper, we analyzed the performance of slotted ALOHA systems consisting of energy harvesting nodes with retry limit. We assumed that the capacities of data and energy buffer at a node are one packet and E packets, respectively, and that one data packet transmission consumes one energy packet. The data and the energy packet arrival processes are modeled as independent and identically distributed Bernoulli processes. Under these assumptions, we developed a node-centric two-dimensional discrete-time Markov chain model, whose states represent a node state described by a two-tuple of the number of data packets in the data buffer and the number of energy packets in the energy buffer. According to the concept of the equilibrium point analysis, the fixed point equation with respect to the ratio of nodes transmitting a data packet was derived.

Based on the numerical results obtained from the fixed point equation, we derived expressions of throughput, the offered traffic, the average transmission delay and the discard probability of data packet. We verified the theoretical results by means of computer simulation. The numerical results indicated that throughput, the offered traffic and the discard probability roughly depend on the minimum of the data packet generation probability and the energy packet generation probability.

Generalization and relaxation of the assumption such as an independent property of the energy packet arrival process are left for further investigation.

References

[1] H. H. R. Sherazi, L. A. Grieco and G. Boggia, "A comprehensive review on energy harvesting MAC protocols in WSNs: Challenges and tradeoffs," *Ad Hoc Networks*, vol. 71, pp. 117–134, Mar. 2018. DOI: 10.1016/j.adhoc.2018.01.004.

[2] M. Moradian and F. Ashtiani, "Throughput analysis of a slotted Aloha-based network with energy harvesting nodes," in *Proc. IEEE PIMRC 2012*, Sydney, Australia, pp. 351–356, Sep. 2012. DOI: 10.1109/PIMRC.2012.6362809.

[3] S. Foss, D. Kim and A. Turlikov, "Model with common energy harvesting for the random multiple access system," in *Proc. REDUNDANCY 2014*, St. Petersburg, Russia, pp. 39–42, June 2014. DOI: 10.1109/RED.2014.7016701.

[4] Y. H. Bae, "Modeling timely-delivery ration of slotted Aloha with energy harvesting," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1823–1826, Aug. 2017. DOI: 10.1109/LCOMM.2017.2693998.

[5] K. Sakakibara, H. Muta and Y. Yuba, "The effect of limiting the number of retransmission trials on the stability of slotted ALOHA systems," *IEEE Trans. Veh. Tech.*, vol. 49, no. 4, pp. 1449–1453, July 2000. DOI: 10.1109/25.875281.

[6] S. Tasaka, "Stability and performance of the R-ALOHA packet broadcast system," *IEEE Trans. Comput.*, vol. C-32, no. 8, pp. 717–726, 1983. DOI: 10.1109/TC.1983.1676309.

[7] M. E. Woodward, *Communication and Computer Networks*, IEEE Computer Society Press, Los Angeles, CA, 1994.

[8] G. Bianchi, ""Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000. DOI: 10.1109/49.840210.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 665–674
DOI: 10.15439/2018F203
ISSN 2300-5963 ACSIS, Vol. 15

Universal serial bus as a communication medium for prototype networked data acquisition and control systems – performance optimisation and evaluation

Andrzej Tutaj, Jacek Augustyn†
AGH University of Science and Technology
Faculty of Electrical Engineering, Automatics,
Computer Science and Biomedical Engineering
Department of Automatics and Robotics
al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: tutaj@agh.edu.pl

***Abstract*—Universal serial bus can be considered a cost-effective and high-throughput communication medium for sensor networks and multinode control or data acquisition systems, especially for prototyping purposes. In a prototype system, a PC or Mac computer with a general-purpose operating system is often selected as a host or root node for the USB bus and it acts as a central data collector, supervisory user interface, and network traffic scheduler. However, achieved communication performance is often unsatisfactory since USB stack drivers incorporated in Windows, Linux, or macOS operating systems are not optimised for such specific purposes. The paper shows how an appropriately selected and implemented user application communication schedule, making use of operating system drivers pipelining and multitasking capabilities, can substantially improve USB network throughput and reduce communication latency.**

Keywords*—*universal serial bus; sensor network; distributed and networked control and data acquisition systems; rapid prototyping; communication scheduling; USB stack pipelining and multitasking.

I. Introduction

A COMMUNICATION channel constitutes a crucial part of every sensor network, distributed data acquisition system, or a networked control system. Its performance affects the overall system quality of service. For measurement data acquisition solutions, which often process large streams of data, the most important network characteristic is its throughput. For closed loop control applications the most critical parameter is the round trip time as it directly influences the net loop time delay. There are various networks and protocols available with different properties and characteristics. They differ in popularity, openness, initial costs and implementation efforts.

For small-scale distributed control and data acquisition systems, a full-speed variant of the Universla Serial Bus (USB) 2.0 can be considered an attractive and convenient choice, especially well suited for prototyping purposes. It provides low-cost, high-throughput communication channel with favourable performance-to-price ratio. The network infrastructure can be built using inexpensive and easily available hardware components. Software USB stacks and drivers are readily available for most common general-purpose operating systems (OS), like Windows, Linux, or macOS, which can host popular rapid control prototyping (RCP) software engineering tools like MATLAB/Simulink or LabVIEW. Modern microcontrollers (MCU) and system on chips (SoC), on which network nodes are likely to be built, are routinely equipped with a USB device port peripheral with an integrated PHY module. High-performance USB device stacks are usually available free of charge from MCU manufacturers.

Unfortunately, a USB stack incorporated in a general-purpose OS is usually not well suited for measurement data acquisition or real-time closed-loop control systems, since it has been designed and optimised for different applications. Hence, the performance of such prototype configurations can be poor unless special measures are undertaken. The paper shows how an appropriately selected communication schedule can substantially improve throughput and timing characteristics of a multinode USB 2.0 network comprising PC computer running Windows 7 OS and MATLAB RCP tool as a host node and several MCU-based full-speed device nodes. The schedule is realized by a user application coded as a MATLAB M-file script and does not require any modifications to the standard OS USB stack. Hence, it could be easily implemented under any RCP tool and executed by any unprivileged OS user. The solution takes advantage of a pipelined and multitasked processing implemented by the OS USB driver stack.

The topic of USB bus applications for data acquisition or control purposes is relatively popular in the literature where numerous examples of various such systems can be found. However, less work is reported concerning communication performance optimization and characteristics adaptation. Some researchers restrict their investigations to one-to-one communication systems, comprising a single host and a single device node [1], [2], [3], [4], [5]. Their solutions employ either a USB-to-UART adapter [1], an integrated circuit standalone USB device controller connected to an MCU [2], [3], or

This work was supported by AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland, grant No. 11.11.120.396.
† Deceased.

an MCU or SoC device with an integrated USB peripheral module [4], [5]. Performances of such systems, expressed in terms of data throughput or timing properties, are considered in [6], [7], [8], [9]. Other authors present applications of multinode USB networks for various data collecting or control purposes, including industrial systems, home automation, or virtual instrumentation for power monitoring [10], [11], [12], [13], but do not investigate system performances extensively. Such a study can be found in [14], while the problem of nodes synchronisation is addressed in [15], [16]. Some authors propose hybrid solutions with the USB bus connecting a host computer and a single controller of another multidrop bus, like RS485, CAN, or I2C [17], [18].

The abundance of USB control and data acquisition solutions, on one hand, and rarity of extensive performance analysis and communication optimisation recommendations, on the other hand, encouraged the authors of this article to devote their research to the latter topic.

The paper is organised as follows. An introduction with motivations and a literature review has been given in section I. Section II presents hardware and software architecture of the test bench system that has been used to verify effectiveness of proposed communication schedules for network performance improvement. Four different schedules considered in the paper are elaborated in Section III. Results of experiments are given and discussed in section IV. Section V provides final remarks and further considerations and is followed by acknowledgements and a reference list.

II. Hardware and software architecture of a test system

A. Host and device nodes

Main hardware and software components of a test system, built in order to measure communication performances for various polling schedules, are shown in Fig. 1. Their technical characteristics are given in Tab. I. A portable PC computer running MS Windows OS and hosting MATLAB application is used as a host node of the USB network. A user application responsible for polling all device nodes is coded as an M-file script running in MATLAB environment. The standard USB driver stack of the OS is employed and standard OS Application Programming Interface (API) is used. Device nodes are implemented on an MCU with integrated USB device port using C++ language and bare metal programming approach. A software USB device stack provided by the MCU manufacturer is employed, however some modifications of the stack code for latency reduction are implemented. In a real application, the device node is expected to interface with a physical system being controlled or monitored. However, for communication performance evaluation, this system functionality is irrelevant and has been omitted.

B. USB transfer mode and speed selection

Out of three possible transmission speeds offered by the USB 2.0 specification: low (LS), full (FS), and high (HS), the full speed is a reasonable choice for moderately demanding applications. The data rate is relatively high compared to CAN, RS-485, or similar standards, and the hardware implementation on the USB device side is simplified, as most modern MCU-s and SOC-s incorporate complete full-speed USB peripheral modules. Hence, the FS variant has been selected for USB devices in the study presented in the paper.

TABLE I
Hardware and software components of the test bench

Host node	Hardware	DELL Latitude E6400, Core 2×2.54 GHz, 4 GB RAM notebook PC computer E-Port Plus PRO2X docking station
	Software	MS Windows 7 Professional SP1 operating system CDC USB class driver ver. 6.1.7601.17514 USB host controller driver ver. 6.1.7601.17586 MATLAB ver. 7.9.0.529 R2009b rapid development environment
Device node	Hardware	Olimex SAM7-EX256 Rev. A evaluation board Atmel SAM7X microcontroller based on ARM7TDMI core, 48 MHz
	Software	USB device stack framework streamlined by the authors system-less bare-metal application written by the authors
USB hub	Hardware Software	USB 2.0 high speed hub

There are four transfer modes available with the USB 2.0 protocol: control, interrupt, isochronous, and bulk [19], [20]. Of these, the bulk mode is a natural choice for a distributed system transferring potentially a large amount of data. Unlike the isochronous one, it provides error detection and correction features. Number of transactions allowed in a single USB frame is not limited as with interrupt mode. Although there is no bandwidth reservation for a bulk transfer, it can consume up to 100% of the available bandwidth, provided that there are no other modes transfers scheduled. Large allowable data packet size helps to reduce transmission overhead and thus allows high data throughput.

A standard and popular Communication Device Class (CDC) has been selected for the test application. Software drivers for CDC class are routinely incorporated in most OS-es, making it attractive for rapid prototyping purposes, as no extra programming is required from the user. Virtual Com Port (VCP) driver is used on the host site. It allows standard OS API as well as standard MATLAB functions set for serial port handling to be employed.

C. Data and control flow in the system

Fig. 2 shows relations between MATLAB API function calls, OS API function calls, USB bus transactions, and device node actions. A user M-file script implements polling policy with each individual device node contacted once in a single cycle. The cycle is repeated endlessly. MATLAB *serial* object as well as *fwrite* and *fread* functions are used. Their calls are translated into *write* and *read* OS API calls and interact with the OS USB stack via VCP and CDC drivers. Hardware host and device controllers on the PC and MCU side, respectively, as well as USB 2.0 hubs are engaged in data transfer over

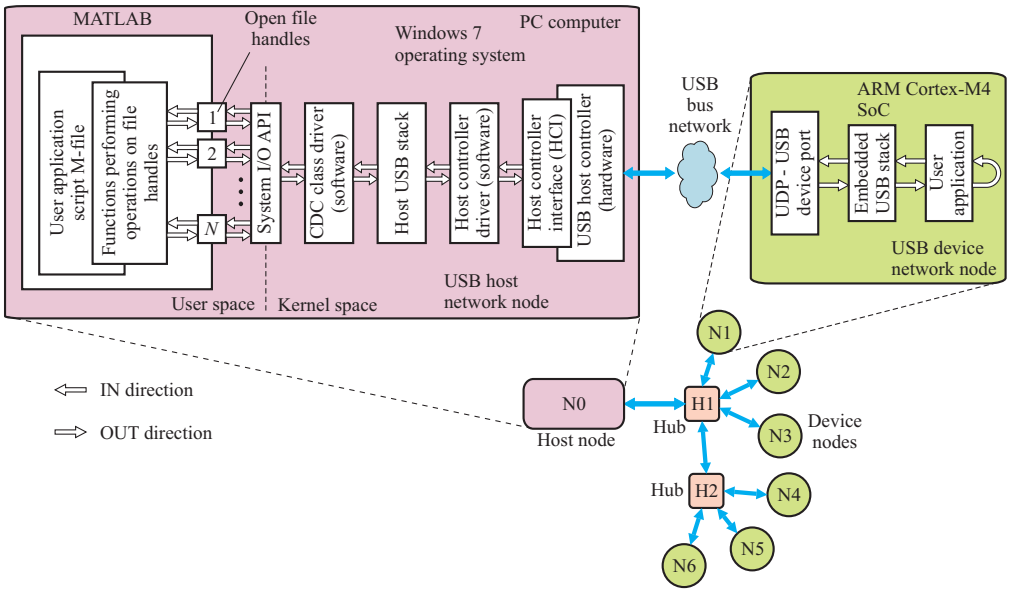


Fig. 1. Hardware and software architecture of the test bench system.

the network. The embedded MCU application on the device side responds to each host query with a predefined amount of data. Greyed and dash-dotted components to the right in the figure, usually present in a real systems, are omitted in the test configuration. The USB device stack provided by the MCU manufacturer has been streamlined by the authors to reduce latency it introduces. It helps to focus the performance study on the network rather than device properties.

In a real control or measurement acquisition system the host is supposed to send to the device control values to be fed to a control plant or parameters controlling the measurement process. The device, on the ohter hand, sends to the host measurement results. One can expect data size asymmetry between *write* and *read* operations with small units of data being sent to the device and large amounts of data being received due to a multichannel or high speed measurements. This expected asymmetry has been taken into account during tests presented further in the paper.

III. User Application Polling Schedules

Four different schedules of device nodes polling by the host side user application are investigated in the article. They are defined, explained and named in the following subsections.

A. Direct interleaved schedule.

Arguably the simplest and the most natural polling scheme is presented in Fig. 3. The user application running on the host node uses *write* call to send a query to a device node and then calls *read* function to wait for a reply. As soon as the data arrives, the host proceeds to the next device. Having finished a full cycle, the host begins a subsequent one. Let us call the duration of a single cycle the *network repetition time* (NRT). It will be used for communication performance evaluation. We will refer to the presented scheme as *direct interleaved schedule*, as *write* and *read* calls alternate and the scheme does not involve any distinct preparatory stage.

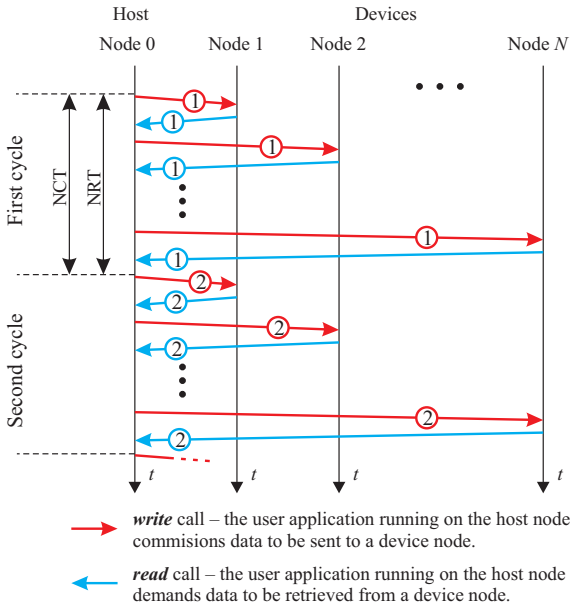


Fig. 3. Direct interleaved schedule – pattern of *write* and *read* I/O functions calls by the user application running on the host node. Encircled numbers on arrows help to match related writing and reading operations (the query and the response corresponding to it).

B. Advanced interleaved schedule

The *advanced interleaved schedule* shown in Fig. 4 differs from the one presented in the previous subsection in having a special initial stage. During this state the host *in advance* writes data to all devices in turn without waiting for any reply. Then it proceeds as with the *direct interleaved schedule*, applying *read* and *write* operations pair to each device in turn

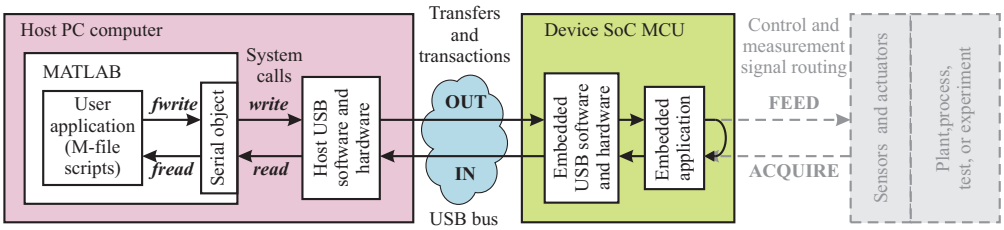


Fig. 2. Relations between user application and OS API function calls and USB bus transaction types.

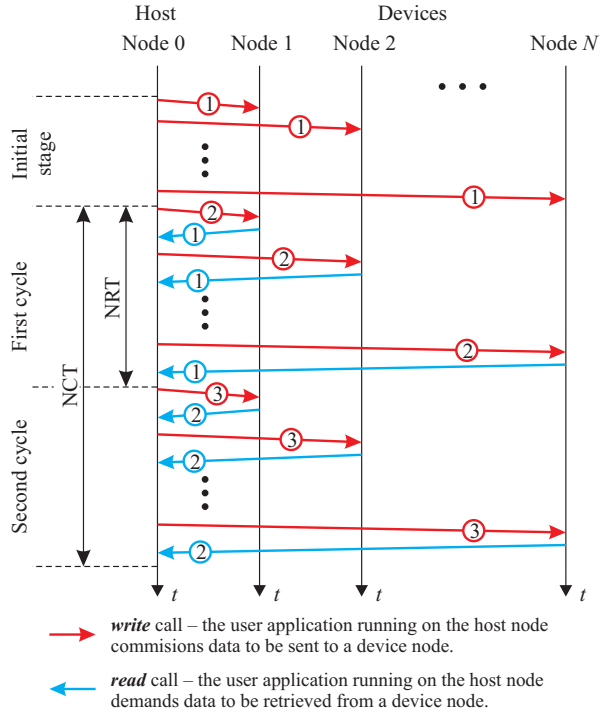


Fig. 4. Advanced interleaved schedule. The *network cycle time* (NCT) equals twice the *network repetition time* (NRT).

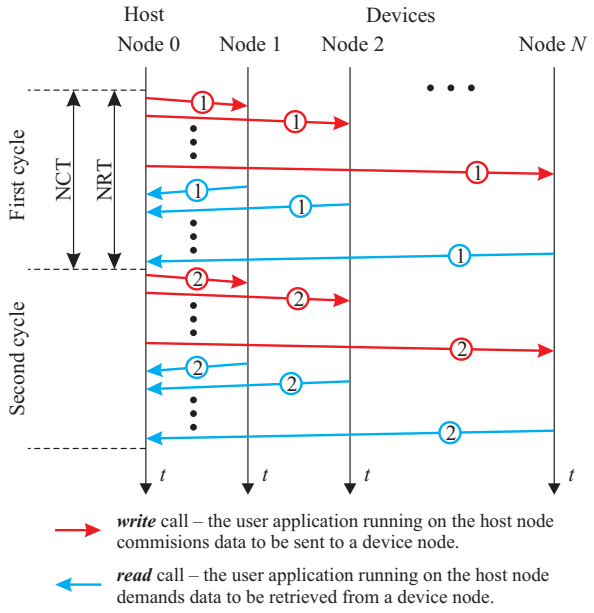


Fig. 5. Direct aggregated schedule. The *network cycle time* (NCT) is equal to the *network repetition time* (NRT).

and repeating the cycle. Thus, there is a sustained excess of writes over reads for each device. It can improve the communication performance by taking advantage of OS USB stack pipelining, buffering, and multithreading capabilities. A new measure called *network control time* (NCT) is introduced in Fig. 4. It is equal to the time elapsing between the beginning of a cycle where *write* operation are effected and the end of a cycle where corresponding *read* calls are completed (note numbers in circles on arrows in the figure). Because of the presence of the initial stage of the schedule, the NCT parameter equals twice the NRT in average. The *control* term in the NCT name alludes to the fact that in a closed-loop control application, NCT rather than NRT parameter influences the quality of control as it contributes to the net time delay in the loop.

C. Direct aggregated schedule

An important drawback of the schedule given in the previous section is that each response received from a device corresponds to last but one query instead of the last one. Hence, there is a one-step query–response shift or delay. Should it be unacceptable for a particular application, one may choose an alternative approach shown in Fig. 5. There is no special initial stage. In every regular cycle the host *aggregates* all write and all read operations in two separate groups, with all writes executed before all reads. That approach provides the stack with an additional time reserve for collection of device replies and does not introduce any shift in messages exchange order.

D. Advanced aggregated schedule.

A combination of *advancing* and *aggregation* techniques is presented in Fig. 6 where the last proposed polling scheme is explained. There is an initial stage comprising *write* calls only while all consecutive cycles start with aggregated writes succeeded by grouped reads. One may expect this schedule to provide further performance improvement by a synergy effect.

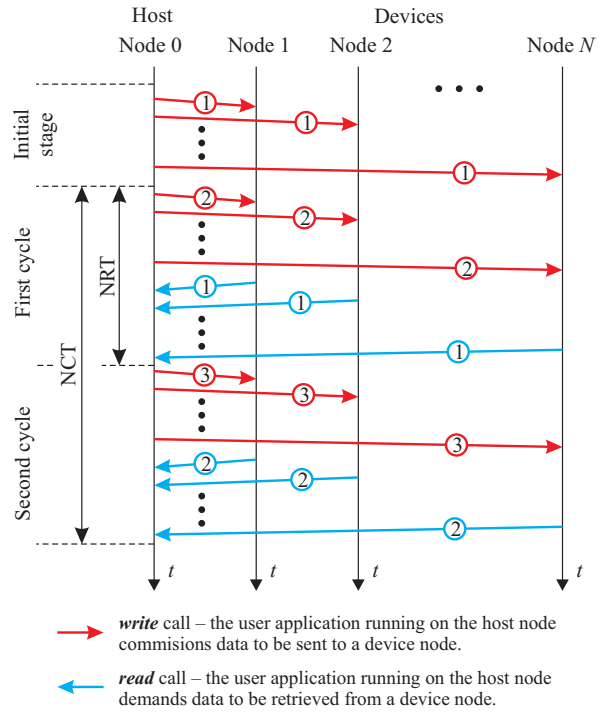


Fig. 6. Advanced aggregated schedule. The *network cycle time* (NCT) equals twice the *network repetition time* (NRT).

IV. Experimental results

A. Test conditions and performance measures

A lot of tests have been conducted for various experimental conditions gathered in Tab. II. Four different polling schedules, proposed in the previous sections, have been investigated in turn. Device nodes number N as well as the data length SIN for a single IN transfer (*read* call) have been varied, while the OUT transaction size SOUT (*write* operation) has been kept constant. A long data series of 10 000 samples have been collected for each experiment in order to compute several statistical performance measures. Four such quantities are used to compare performances of individual scheduling policies, based on network timing or throughput characteristics (see Tab. III). Timing is characterised by NRT and NCT parameters defined in the previous section. They are computed on the host side based on timestamps added to transferred data by device nodes employing hardware peripheral timers. Throughput corresponding to data transferred by IN transactions (*read* operations) is characterised by two related measures: *total network stream* (TNS) and *stream per node* (SPN), satisfying the equation $\mathrm{TNS} = N \times \mathrm{SPN}$ (as long as mean values are considered) where N is the number of active device nodes. For all four quantities their average values (avg) have been computed. For timing related NCT parameter its standard deviation (std) has been also determined. Time series and histograms of timing parameters obtained in selected experiments are presented in the next subsection in Fig. 7–10. Statistics computed from all tests results are gathered in Tab. IV. Discussion of results is also provided. The amount

TABLE II
Experiments conditions and parameters

Parameter or condition	Value or variant
polling schedule	direct interleaved, advanced interleaved, direct aggregated, advanced aggregated
number of active device nodes in the network	1, 2, 3, 4, 5, 6
data length for a single IN transfer (*read* call) SIN, B	48 B, 100 B, 200 B, 500 B, 750 B, 1000 B, 1250 B, 2000 B, 4000 B, 6000 B, 8000 B
data length for a single OUT transaction (*write* call) SOUT, B	16 B

TABLE III
Network performance measures

timing	NRT, ms	network repetition time
	NCT, ms	network control time
throughput	TNS, kB/s	total network stream
	SPN, kB/s	stream per node

of data presented in the table may seem to be intimidating for the reader. However, authors decided to include all results because they share a view expressed in [21]: *It is understood that real time systems are not tested with a single analysis that pronounces them correct. Testing of real time systems is a proof by exhaustion.*

B. Presentation and discussion of experimental results

Fig. 7 presents time series and histograms of the NCT parameter obtained for the *direct interleaved schedule* in an experiment with four active device nodes and IN transfer size of $\mathrm{SIN} = 100\,\mathrm{B}$. For a direct schedule, $\mathrm{NCT} = \mathrm{NRT}$ equality holds. The average (avg) NCT value is equal to 38 ms (see Tab. IV) and approximately matches performances observed by other authors for USB systems with a single device [7]. Minimum (min) and standard deviation (std) of NCT equal 11 ms and 21 ms, respectively. Large discrepancy between avg and min value as well as large std/avg ratio reveals a large room for improvement, since the min value estimates the best case scenario. From Tab. IV one can deduce that NCT is approximately proportional to the device nodes number N. That seems to be a natural behaviour for the schedule that executes two-way data exchange witch every devices in turn and does not employ any kind of parallelism. A relation between SIN and NCT is approximately affine with a considerable y-intercept and relatively small slope (NCT increases merely by 60% for SIN increasing over 160 times). It suggests that software components rather than a physical data exchange channel form a communication bottleneck and again suggests a potential for performance improvement by an appropriate polling schedule selection. The TNS is roughly proportional to SIN and almost independent on N. It shows the advantage of using large size transfers and reveals that the available throughput is equally shared by all device nodes.

Time series of NCT and NRT as well as NCT histograms for the *advanced interleaved schedule*, $N = 4$, and $\mathrm{SIN} = 100\,\mathrm{B}$

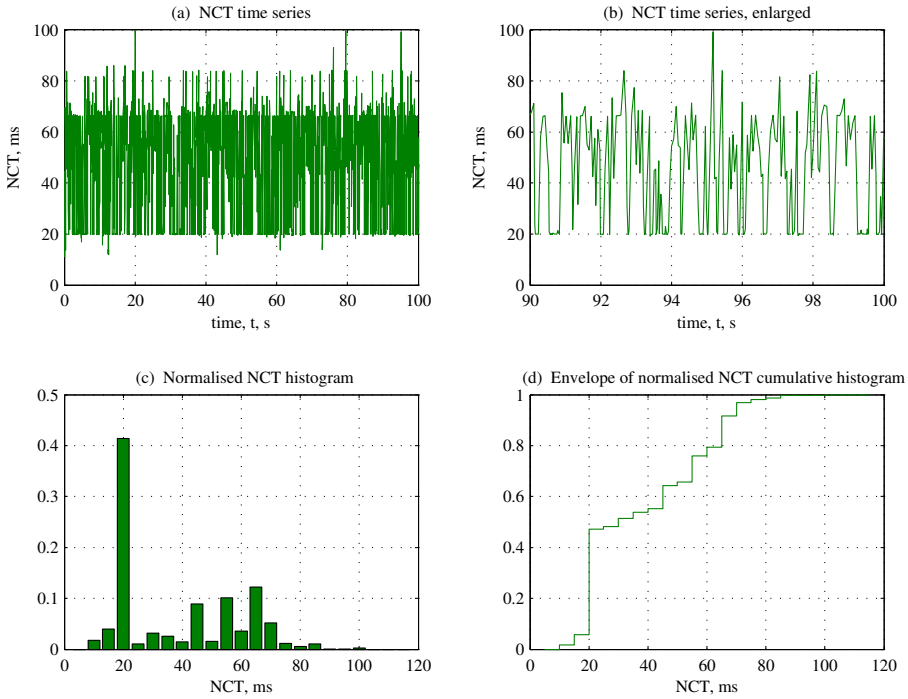


Fig. 7. Results of an experiment for the *direct interleaved schedule* with $N = 4$ active nodes and IN transfer size of 100 B. Time series and histograms of the *network control time* (NCT) parameter.

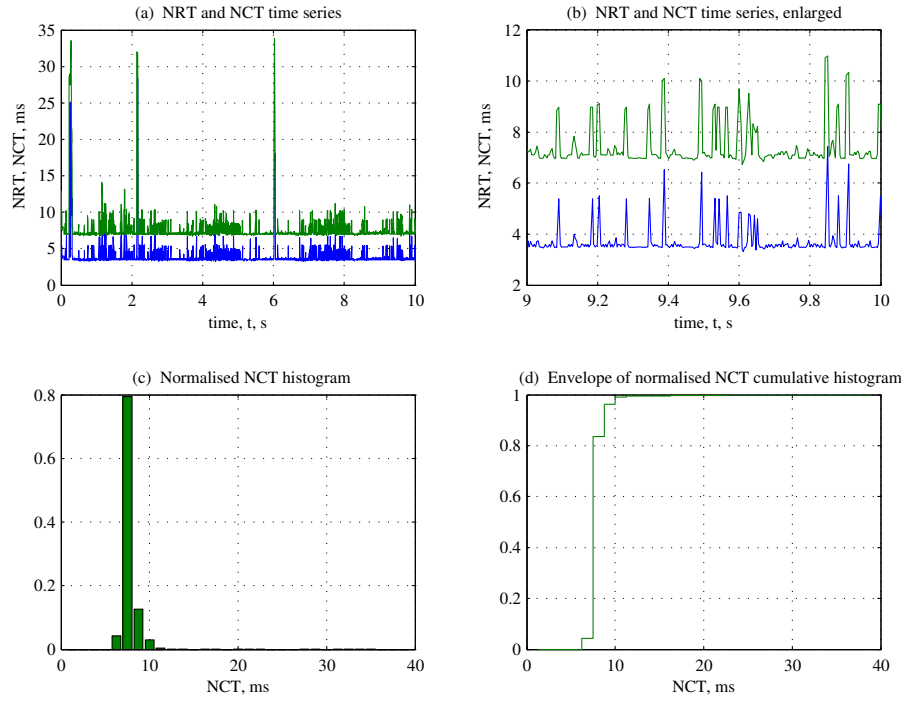


Fig. 8. Results of an experiment for the *advanced interleaved schedule* with $N = 4$ active nodes and IN transfer size of 100 B. Time series and histograms of *network repetition time* (NRT, blue) and *network control time* (NCT, green) parameters.

are presented in Fig. 8. Timing and throughput statistics for various conditions are gathered in Tab. IV. An enormous performance improvement can be observed compared to the *direct interleaved* scheme. The NCT is reduced five times from $38\mathrm{ms}$ to $7.5\mathrm{ms}$. The NRT reduction is even more impressive – about ten times from $38\mathrm{ms}$ to $3.7\mathrm{ms}$. Consequently, both TNS and SPN increase tenfold from $10\,\mathrm{kB/s}$ to $110\,\mathrm{kB/s}$ and from $2.6\,\mathrm{kB/s}$ to $27\,\mathrm{kB/s}$, respectively. The huge improvement is achieved exclusively by introduction of the initial, preparatory stage at the beginning of the cyclic polling schedule (compare Fig. 3 and 4). The one-step shift (and resulting time delay) introduced by the *advancing* technique is by far compensated by the timing performance improvement, revealed by the considerable reduction in both NCT and NRT measures. One can observe that the performance gain is more prominent for small SIN values. Apparently, when the data stream increase, the hardware limitations play more and more important role and diminish benefits brought by the polling scheme modification.

Results for the *direct aggregated schedule* are presented in Fig. 9 and in Tab. IV. For $N = 4$ and $\mathrm{SIN} = 100\,\mathrm{B}$, NCT and NRT become reduced over 6 times (from $38\,\mathrm{ms}$ to $6\,\mathrm{ms}$) compared to the *direct interleaved schedule* while TNS and SPN increase about 6.5 times (from $10\,\mathrm{kB/s}$ to $67\,\mathrm{kB/s}$ and from $2.6\,\mathrm{kB/s}$ to $17\,\mathrm{kB/s}$, respectively). Comparison of two improved polling schemes, the *advanced interleaved schedule* and the *direct aggregated schedule*, reveals that the former performs generally better as long as NRT, TNS, and SPN mean values are considered. However, for the average NCT, the latter scheme shows advantage for most N and SIN combinations. Consequently, for data acquisition systems, the *advanced interleaved schedule* is the preferred one while for the closed-loop control systems the choice should be made based of the number of nodes and IN transfers sizes.

Results of experiments for the *advanced aggregated schedule* are shown in Fig. 10 and in Tab. IV. They reveal a large improvement compared to the *direct interleaved* scheme. On the other hand, the table shows that the performance of this combined schedule is comparable to that obtained for the *advanced interleaved* one. Apparently, the *advancing* approach takes advantage of USB stack pipelinign, multithreading, and parallel computing capabilities to an extent that cannot be further intensified by incorporation of the aggregating technique.

C. Sporadic timing spikes

One can observe sporadic spikes on NCT and NRT time series presented in Fig. 7–10. They are several times higher than the average value of the considered timing parameter. They may result from an occasional lengthy or prolonged preemption of the USB stack or the user application by an unrelated time-consuming task, like hard disk servicing routine. One can expect such behavious since the Windows 7 is not a real-time OS. For a production system such a lack of determinism would be probably a prohibiting factor. For rapid prototyping purposes, however, it may be acceptable, since is far outweighed by development and testing benefits brought by RCP engineering tools like MATLAB or LabVIEW hosted by general-purpose OS-es.

V. Conclusions

The solution presented in the paper is intended for prototype rather than production systems and mainly for rapid prototyping approach. It allows to obtain high performance of a USB-based network despite the application of standard USB stack available in a general-purpose operating system. The dramatic communication improvement is achieved by employment of an appropriately modified read and write function call schedule on the user application side. It takes advantage of a multithreading, parallel computing, buffering and pipelining in the USB stack drivers to streamline the data exchange processes and improve data rate as well as timing characteristics. In a production real-time system designed for data acquisition or distributed control, one may expect protocol stacks to be adapted to the intended applications. That leaves less space for improvement with methods like those proposed in the paper.

All results included in the article have been obtained for a multinode system with several USB devices. However, some proposed methods and schedules can be used as well for a system comprising a single device communicating with a single host. Application of the advancing technique for such a system have been presented in authors' previous work [9].

The paper proves effectiveness of advancing and aggregating techniques in case of a network based on the USB bus technology. However, the authors expect that similar approaches may succeed also for other communication networks and protocols, provided that they make use of a similar software architecture.

The methods given in the article can be beneficial mainly for data acquisition systems, where data throughput maximization rather than closed loop latency (delay) minimization is the main objective. However, to a limited extent, they can also be employed in closed loop control applications as in some cases they also allow reduction of the round trip delay.

Acknowledgements

This work was funded by the AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Krakow, Poland, grant No. 11.11.120.396.

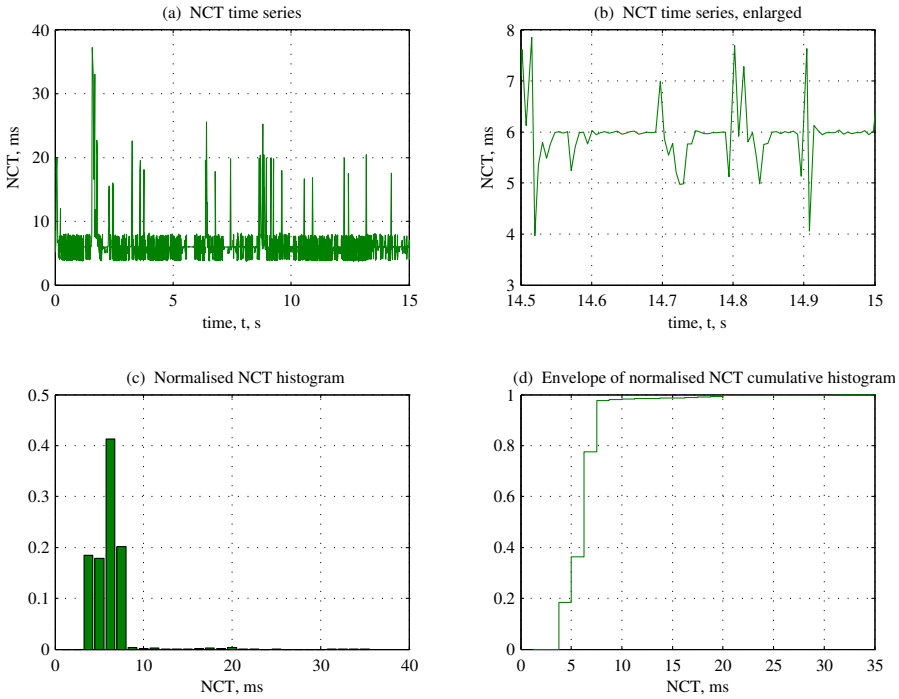


Fig. 9. Results of an experiment for the *direct aggregated schedule* with $N = 4$ active nodes and IN transfer size of 100 B. Time series and histograms of *network control time* (NCT) parameter.

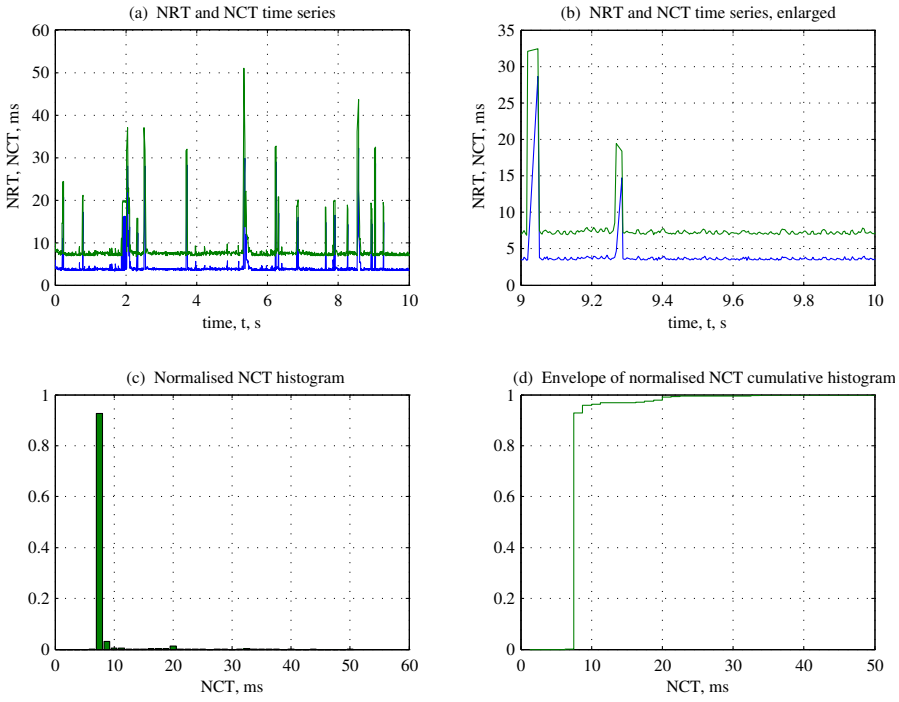


Fig. 10. Results of an experiment for the *advanced aggregated schedule* with $N = 4$ active nodes and IN transfer size of 100 B. Time series and histograms of *network repetition time* (NRT, blue) and *network control time* (NCT, green) parameters.

TABLE IV
RESULTS OF EXPERIMENTS – TIMING AND THROUGHPUT MEASURES STATISTICS FOR DIFFERENT POLLING SCHEDULES AND VARIOUS TEST CONDITIONS. ALL DATA ROUNDED TO TWO SIGNIFICANT DIGITS. RESULTS FOR $N = 4$ AND $\mathrm{SIN} = 100$ B MARKED IN BOLD.

		direct interleaved schedule					advanced interleaved schedule				
	SIN=	48	100	500	2000	8000	48	100	500	2000	8000
NCT avg ms	$N=1$	10	8.3	9.1	13	17	3.6	3.5	4.2	8.8	23
	$N=2$	18	20	25	27	34	4.9	4.9	6	8.8	28
	$N=3$	31	30	38	43	52	6.2	6.3	7.1	11	31
	$N=4$	37	**38**	49	57	69	7.4	**7.5**	7.8	10	28
	$N=5$	54	52	63	70	86	9	9.1	9.4	11	29
	$N=6$	62	57	77	85	100	11	11	11	12	32
NCT std ms	$N=1$	7.6	6.6	7.9	6.9	5.6	1.1	1	1.7	5.2	8.2
	$N=2$	12	12	13	11	8.3	1.1	1.5	0.78	4	9.9
	$N=3$	16	16	17	13	11	1.5	4	1.4	4.4	11
	$N=4$	21	**21**	22	16	13	1.1	**1.5**	5	2.7	9.5
	$N=5$	24	25	26	19	15	1.2	1.9	4.6	2.6	8.7
	$N=6$	29	29	28	22	17	1.2	1.6	1.9	3	6.4
NRT avg ms	$N=1$	10	8.3	9.1	13	17	1.8	1.7	2.1	4.4	12
	$N=2$	18	20	25	27	34	2.5	2.5	3	4.4	14
	$N=3$	31	30	38	43	52	3.1	3.1	3.6	5.3	15
	$N=4$	37	**38**	49	57	69	3.7	**3.7**	3.9	5.1	14
	$N=5$	54	52	63	70	86	4.5	4.6	4.7	5.4	14
	$N=6$	62	57	77	85	100	5.4	5.4	5.5	6.1	16
TNS avg kB/s	$N=1$	4.8	12	55	160	460	26	58	240	450	680
	$N=2$	5.2	10	40	150	460	39	81	330	900	1100
	$N=3$	4.7	10	39	140	460	46	95	420	1100	1500
	$N=4$	5.2	**10**	41	140	460	52	**110**	520	1600	2300
	$N=5$	4.5	9.6	40	140	460	53	110	530	1900	2800
	$N=6$	4.6	11	39	140	460	54	110	550	2000	3000
SPN avg kB/s	$N=1$	4.8	12	55	160	460	26	58	240	450	680
	$N=2$	2.6	5.1	20	74	230	19	41	170	450	560
	$N=3$	1.6	3.4	13	47	150	15	32	140	380	520
	$N=4$	1.3	**2.6**	10	35	120	13	**27**	130	400	570
	$N=5$	0.89	1.9	8	29	93	11	22	110	370	560
	$N=6$	0.77	1.8	6.5	23	77	9	19	91	330	500

		direct aggregated schedule					advanced aggregated schedule				
	SIN=	48	100	500	2000	8000	48	100	500	2000	8000
NCT avg ms	$N=1$	8.8	7.3	8.1	12	18	3.3	3.4	4	8.7	24
	$N=2$	3.3	4.4	9.3	12	20	4.4	4.6	5.1	7	25
	$N=3$	5.2	5.2	6.8	15	23	6.1	6.2	6.2	7.7	24
	$N=4$	6	**6**	6.7	13	25	7.6	**7.9**	8	8.5	20
	$N=5$	7	7.1	7.5	10	26	9.5	9.9	9.6	10	20
	$N=6$	7.1	7.3	8.3	11	30	11	12	12	12	27
NCT std ms	$N=1$	6.6	5.8	7.2	7.5	7.6	1.4	1.6	1.5	5.1	9.5
	$N=2$	2.5	4.2	7.4	7.8	7.3	1.7	2	3.4	2.5	9.8
	$N=3$	3.4	3.2	5	8.9	9.4	2.6	2.5	3.8	3.8	9.4
	$N=4$	2.2	**2.3**	3.3	8.5	8.9	2.2	**2.7**	4.2	2.6	4.6
	$N=5$	3.1	3	3.7	4.8	9.6	2.5	3.1	2.6	2.6	4.9
	$N=6$	3.4	2.9	3.4	3.9	8.3	2.6	3	2.6	3	3.3
NRT avg ms	$N=1$	8.8	7.3	8.1	12	18	1.6	1.7	2	4.3	12
	$N=2$	3.3	4.4	9.3	12	20	2.2	2.3	2.5	3.5	13
	$N=3$	5.2	5.2	6.8	15	23	3	3.1	3.1	3.9	12
	$N=4$	6	**6**	6.7	13	25	3.8	**3.9**	4	4.3	9.8
	$N=5$	7	7.1	7.5	10	26	4.8	4.9	4.8	5.1	9.9
	$N=6$	7.1	7.3	8.3	11	30	5.7	5.8	5.8	6.1	13
TNS avg kB/s	$N=1$	5.5	14	62	170	430	29	59	250	460	660
	$N=2$	29	46	110	340	780	43	87	390	1100	1300
	$N=3$	28	58	220	400	1000	48	97	480	1500	2000
	$N=4$	32	**67**	300	600	1300	51	**100**	500	1900	3300
	$N=5$	34	71	330	1000	1500	50	100	520	2000	4000
	$N=6$	41	83	360	1100	1600	51	100	520	2000	3500
SPN avg kB/s	$N=1$	5.5	14	62	170	430	29	59	250	460	660
	$N=2$	14	23	54	170	390	22	44	200	570	630
	$N=3$	9.2	19	74	130	340	16	32	160	520	660
	$N=4$	8	**17**	75	150	320	13	**25**	120	470	810
	$N=5$	6.8	14	67	200	310	10	20	100	400	800
	$N=6$	6.8	14	60	190	260	8.5	17	87	330	590

REFERENCES

[1] M. A. Ahmad, A. N. K. Nasir, N. S. Pakheri, N. M. Ghani, M. A. Zawawi, and N. H. Noordin, "Microcontroller-based input shaping for vibration control of flexible manipulator system," *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 597–610, 2011.

[2] C. Qiong, P. Zhuo, and C. Hui, "The communication design of simulation and measurement for excitation system based on USB2.0," in *2nd International Workshop on Intelligent Systems and Applications (ISA)*, Wuhan, China, 22-23 May 2010. doi: 10.1109/IWISA.2010.5473535 pp. 1–4. [Online]. Available: https://doi.org/10.1109/IWISA.2010.5473535

[3] T. Baohua and Q. Shuhai, "A high speed data acquisition card based on USB bus," in *International Conference on Machine Vision and Human Machine Interface (MVHI)*, Kaifeng, China, 24-25 April 2010. doi: 10.1109/MVHI.2010.179 pp. 357–360. [Online]. Available: https://doi.org/10.1109/MVHI.2010.179

[4] A. Kumar, I. P. Singh, and S. K. Sud, "Energy efficient and low cost indoor environment monitoring system based on the IEEE 1451 standard," *IEEE Sensors Journal*, vol. 11, no. 10, pp. 2598–2610, 2011. doi: 10.1109/JSEN.2011.2148171. [Online]. Available: https://doi.org/10.1109/JSEN.2011.2148171

[5] G. Wang, X. Cheng, and Z. Wang, "Terminal design of the intelligent data acquisition system based on USB interface," *Applied Mechanics and Materials*, vol. 380-384, pp. 3629–3632, 2013. doi: 10.4028/www.scientific.net/AMM.380-384.3629. [Online]. Available: https://doi.org/10.4028/www.scientific.net/AMM.380-384.3629

[6] L. Ramadoss and J. Y. Hung, "A study on universal serial bus latency in a real-time control system," in *34th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1-5, Orlando, Florida, USA, 10-13 November 2008. doi: 10.1109/IECON.2008.4757930 pp. 19–24. [Online]. Available: https://doi.org/10.1109/IECON.2008.4757930

[7] R. P. Gomez, J. J. E. Rodriguez, G. A. Hemandez, and A. M. Sibaja, "USB bulk transfers between a PC and a PIC microcontroller for embedded applications," in *5th Electronics, Robotics and Automotive Mechanics Conference Proceedings (CERMA)*, Cuernavaca, Mexico, 30 September - 3 October 2008. doi: 10.1109/CERMA.2008.21 pp. 559–564. [Online]. Available: https://doi.org/10.1109/CERMA.2008.21

[8] J. Augustyn and A. Bieñ, "Real time performance of USB interface in embedded control and measurement systems," *Przegląd Elektrotechniczny*, vol. 85, no. 7, pp. 1–7, 2009.

[9] J. Augustyn and A. Tutaj, "Evaluation and optimisation of communication performance in a hybrid measurement and control system," *Studies in Informatics and Control*, vol. 23, no. 4, pp. 341–351, 2014. doi: 10.24846/v23i4y201404. [Online]. Available: https://doi.org/10.24846/v23i4y201404

[10] A. Depari, A. Flammini, D. Marioli, and A. Taroni, "USB sensor network for industrial applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 7, pp. 1344–1349, July 2008. doi: 10.1109/TIM.2008.915487. [Online]. Available: https://doi.org/10.1109/TIM.2008.915487

[11] Y. S. Kim, H. S. Kim, , and C. G. Lee, "The development of USB home control network system," in *8th International Conference on Control, Automation, Robotics and Vision (ICARCV 2004)*, vol. 1-3, Kunming, Peoples Republic of China, 6-9 December 2004. doi: 10.1109/ICARCV.2004.1468839 pp. 289–293. [Online]. Available: https://doi.org/10.1109/ICARCV.2004.1468839

[12] C. P. Young, M. J. Devaney, and S. C. Wang, "Universal serial bus enhances virtual instrument-based distributed power monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 6, pp. 1692–1697, December 2001. doi: 10.1109/19.982969. [Online]. Available: https://doi.org/10.1109/19.982969

[13] P. P. Stang, S. M. Conolly, J. M. Santos, J. M. Pauly, and G. C. Scott, "Medusa: A scalable MR console using USB," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 370–379, 2012. doi: 10.1109/TMI.2011.2169681. [Online]. Available: http://dx.doi.org/10.1109/TMI.2011.2169681

[14] P. E. Guerrero, I. Gurov, A. Buchmann, and K. V. Laerhoven, "Diagnosing the weakest link in wsn testbeds: A reliability and cost analysis of the USB backchannel," in *Proceedings of the 37th Annual IEEE Conference on Local Computer Networks (LCN 2012)*, Clearwater, Florida, USA, 22-25 October 2012. doi: 10.1109/LCNW.2012.6424085 pp. 934–942. [Online]. Available: https://doi.org/10.1109/LCNW.2012.6424085

[15] J. Dvorak and J. Havlik, "Data synchronization for independent USB devices," in *2011 International Conference on Applied Electronics (AE)*, Pilsen, Czech Republic, 7-8 September 2011, pp. 1–3.

[16] P. Foster, A. Kouznetsov, N. Vlasenko, and C. Walker, "Sub-nanosecond distributed synchronisation via the universal serial bus," in *IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication (ISPCS 2007)*, 1-3 October 2007. doi: 10.1109/ISPCS.2007.4383772 pp. 44–49. [Online]. Available: https://doi.org/10.1109/ISPCS.2007.4383772

[17] E. J. Bueno, A. Hernandez, F. J. Rodriguez, C. Girón, R. Mateos, and S. Cobreces, "A dsp- and fpga-based industrial control with high-speed communication interfaces for grid converters applied to distributed power generation systems," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 3, pp. 654–669, 2009. doi: 10.1109/TIE.2008.2007043. [Online]. Available: https://doi.org/10.1109/TIE.2008.2007043

[18] U. Saranli, A. Avci, and M. C. Ozturk, "A modular real-time fieldbus architecture for mobile robotic platforms," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 3, pp. 916–927, 2011. doi: 10.1109/TIM.2010.2078351. [Online]. Available: https://doi.org/10.1109/TIM.2010.2078351

[19] *Universal Serial Bus Specification, Rev. 2.0*, 27 April 2000.

[20] J. Axelson, *USB Complete. Everything you need to develop Custom USB peripherals*, 3rd ed. Lakeview Research LLC, 2005.

[21] M. Hall, "Windows CE 5.0 for real time systems," *Embedded Computing Design*, vol. 3, no. 6, pp. 37–43, 13 November 2005.

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems. Events that constitute IT4BMS are:

- AITM'18—16th Conference on Advanced Information Technologies for Management
- ISM'18—13th Conference on Information Systems Management
- KAM'18—24th Conference on Knowledge Acquisition and Management

16th Conference on Advanced Information Technologies for Management

WE are pleased to invite you to participate in the 16th edition of Conference on "Advanced Information Technologies for Management AITM'18". The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

TOPICS

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

EVENT CHAIRS

- **Andres, Frederic,** National Institute of Informatics, Tokyo, Japan
- **Dudycz, Helena,** Wrocław University of Economics, Poland
- **Dyczkowski, Mirosław,** Wrocław University of Economics, Poland
- **Hunka, Frantisek,** University of Ostrava, Czech Republic
- **Korczak, Jerzy,** International University of Logistics and Transport, Wrocław, Poland

PROGRAM COMMITTEE

- **Abramowicz, Witold,** Poznan University of Economics, Poland
- **Ahlemann, Frederik,** University of Duisburg-Essen, Germany
- **Atemezing, Ghislain,** Mondeca, Paris, France
- **Cortesi, Agostino,** Università Ca' Foscari, Venezia, Italy
- **Czarnacka-Chrobot, Beata,** Warsaw School of Economics, Poland
- **De, Suparna,** University of Surrey, Guildford, United Kingdom
- **Dufourd, Jean-François,** University of Strasbourg, France
- **Franczyk, Bogdan,** University of Leipzig, Germany
- **Januszewski, Arkadiusz,** University of Science and Technology in Bydgoszcz, Poland
- **Kannan, Rajkumar,** Bishop Heber College (Autonomous), Tiruchirappalli, India
- **Kersten, Grzegorz,** Concordia University, Montreal, Canada
- **Kowalczyk, Ryszard,** Swinburne University of Technology, Melbourne, Australia
- **Kozak, Karol,** TUD, Germany
- **Krótkiewicz, Marek,** Wroclaw University of Science and Technology, Poland
- **Leyh, Christian,** University of Technology, Dresden, Germany
- **Ligęza, Antoni,** AGH University of Science and Technology, Poland
- **Ludwig, André,** Kühne Logistics University, Germany
- **Magoni, Damien,** University of Bordeaux – LaBRI, France
- **Michalak, Krzysztof,** Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pankowska, Malgorzata,** University of Economics in Katowice, Poland
- **Pinto dos Santos, Jose Miguel,** AESE Business School Lisboa, Portugal
- **Proietti, Maurizio,** IASI-CNR (the Institute for Systems Analysis and Computer Science), Italy
- **Rot, Artur,** Wroclaw University of Economics, Poland
- **Stanek, Stanislaw,** General Tadeusz Kosciuszko Military Academy of Land Forces in Wroclaw, Poland
- **Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Tazi, El Bachir,** Moulay Ismail University, Meknes, Morocco
- **Teufel, Stephanie,** University of Fribourg, Switzerland
- **Tsang, Edward,** University of Essex, United Kingdom
- **Wątróbski, Jarosław,** University of Szczecin, Poland
- **Wendler, Tilo,** Hochschule fur Technik und Wirtschaft Berlin
- **Wolski, Waldemar,** University of Szczecin, Poland
- **Zanni-Merk, Cecilia,** INSA de Rouen, France
- **Ziemba, Ewa,** University of Economics in Katowice, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 679–682
DOI: 10.15439/2018F126
ISSN 2300-5963 ACSIS, Vol. 15

Attribute Selection with Filter and Wrapper: An Application on Incident Management Process

Claudio A. L. do Amaral, Marcelo Fantinato, Sarajane Marques Peres
Graduate Program in Information Systems, Universidade de São Paulo, Brazil
{claudio.amaral, m.fantinato, sarajane}@usp.br

***Abstract*—Few approaches allow assertive estimates for ticket completion time in incident management. The accuracy level of prediction models depends on how useful the used attributes are. Moreover, to effectively use computational resources, a canonical attribute subset must be used. This paper proposes two automated attribute selection methods to build prediction model. A filter method and two wrapper search techniques were combined with annotated transition systems to automate attribute selectors applied to a real-life incident management process. The results show that the wrapper method surpassed human experts' decision making.**

***Index Terms*—process mining; attribute selection; incident management; ITIL; annotated transition system.**

I. Introduction

IN Information Technology (IT), optimization is sought by adopting frameworks such as the Information Technology Infrastructure Library (ITIL) [1]. ITIL covers several IT service management processes, including incident management [2], which is responsible for correcting failures and restoring the normal service operation, as soon as possible, minimizing the impact on business [1]. One of the most relevant monitoring indicators related to this process is the completion time for incident resolution (a.k.a. 'ticket completion time') [2].

Assertive and reliable estimates for completion time is still challenging [3]. A common reason for poor estimates is to conduct predictions based only on a naive and superficial abstraction of the actual process being performed. Fortunately, many companies are using process-aware information systems and recording events about the activities executed. The large amount of data recorded in event logs can be explored in detail through different process mining techniques, which allow to infer a more realistic process model [4]. For example, representing the process as an Annotated Transition System (ATS) allows to estimate the process completion time based on statistics aggregated into the process model [5]. To achieve a proper ATS model representing an incident management process, both the event log and a set of descriptive attributes need to be considered. However, depending on management context, the number of descriptive attributes that may be associated with process instances can be large and complex enough to render unfeasible *(i)* the use of all descriptive attributes, which could generate inefficient ATSs to predict completion time as well as *(ii)* a non-automated decision making about which attributes should be considered to build the ATS. Therefore, two concerns should be considered when building a proper process model: not all attributes are necessarily useful and much computational resource may be required. Thus, a canonical subset of descriptive attributes must be selected; i.e., an ideal minimum subset of descriptive attributes that minimizes the computational cost and contains the maximum information relevant to build the model.

This paper proposes to apply two classic methods of attribute selection to automatically determine the canonical subset of descriptive attributes. The filter [6] and wrapper [7] methods have been applied to an event log obtained from a real-world enterprise incident management system. For the experiments, ATSs were created using attribute subsets selected by human experts in addition to the two automated methods. The results show that the proposed automated methods surpass human experts in selecting attributes for prediction using the ATS built based on the selected attributes, having wrapper surpassed filter. The remainder of this paper presents: background, related work, proposed solution, experiments and results, and conclusion and future work.

II. Background

The transition systems used in process mining was proposed by Aalst *et al*. [8] and then extended with annotations to describe statistical data that allow predicting the completion time of a process instance [5]. To create the ATS, each state is annotated with data collected from all traces that have visited it [5]. For time analysis, for example, data about the completion time of the instances related to each earlier trace is used. The data are aggregated in each state producing statistics such as average time, standard deviation, median time etc. Two of the proposed strategies are applied here: *maximal horizon* and *representation*, including per *sequence*, *multiset* and *set* [5].

Attribute selection is essential to build a model capable of predicting ticket completion time, by deciding which features to describe the concept to be learned and how to combine them [9]. Methods for selecting attributes are typically classified as filters, wrappers and embedded [6]. In this paper, a filter method based on correlation analysis was applied. Each attribute is individually evaluated based on its correlation with the target attribute (i.e., the ticket completion time). Moreover, two well-known search techniques were applied: hill-climbing and best-first search [7]; having ATSs as the learning model and Mean Absolute Percentage Error (MAPE) [10], [11] as the metric to evaluate the learning model accuracy.

III. Related Work

The proposal presented in this paper is based on the extension of transition systems with annotations, which was originally proposed by Aalst, Schonenberg and Song [5], to predict completion time of running traces. According to them, ATSs include alternatives for state representation, allowing to address overfitting and underfitting in prediction tasks. They concluded that their prediction approach overcomes simple heuristic approaches. Other authors, as Polato *et al*. [12], extended ATS to solve the same task by combining the probability of occurrence of activities with a regression model.

In addition to transition systems, Petri nets have also been used as a technique for prediction work. Rogge-Solti, Vana and Mendling [13] introduced time series Petri net models, making it possible to handle the analysis of temporal aspects of processes. Hinka *et al*. [14], Evermann, Rehse and Fettke [15] and Tax *et al*. [16] presented approaches closer to the study presented in this paper. Tax *et al.* [16] presented a comparison of their approach with ATSs used as predictor and concluded that they obtain more accurate predictions except for instances with a reduced number of events. Approaches were assessed with cross-validation and prediction accuracy metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) [10].

Although with different strategies addressing completion time prediction, there is a lack of concern on choosing the input log configuration for the predictor induction. A preprocessing work for attribute selection, as proposed in this paper, has the potential to improve results of the related work.

IV. Proposed Solution

When an incident occurs, it is identified and reported by a caller. Afterwards, a primary expectation is to know the incident completion time. The usual estimates follow ITIL best practices, which are based on some specific incident attributes such as urgency, category etc. This approach is fairly general and inaccurate since it aggregates a large number of different situations and common target completion times. As the process evolves from early stage to initial support and investigation, some attributes are updated and new ones are added. Depending on the system used, it can usually lead to a number close to 100 attributes. There is an open issue related to providing assertive estimates on incident completion time that is not adequately solved by simple statistical methods. Incident management systems store descriptive information of process instances and audit information about the history of updates of the process in progress. The combination of both types of information allows executing a detailed step-by-step process evaluation and hence deriving estimates for each event.

The problem addressed here lies in such a scenario, where, one needs to discover an attribute subset that allows generating a model capable of minimizing the prediction error of the incident completion time during the process of its resolution. The process starts with a sequence of actions to build the enriched event log used to induce the prediction models. After that, it is possible to apply the three attribute selection methods explored in this paper: *(i)* expert-driven selection, *(ii)* filter with ranking and *(iii)* wrappers with the hill-climbing and best-first search techniques. The search is performed in the context of our attribute selection strategy. In these algorithms, there is a construction function to build an ATS and the evaluation function of the ATS. They use, respectively, a training log excerpt and a testing log excerpt, which represent disjoint subsets of the original event log generated in a cross-validation procedure. After that, the evaluation function is applied and returns the MAPE for the ATS under evaluation. The minimization function applied to the ATS evaluation returns the index of the model that produces the lowest MAPE when applied to the testing log. As final result, it is returned the ATS with lowest MAPE in the set of ATSs under evaluation.

For all the selection methods, ATS is applied as the prediction model responsible for generating the estimates of the incident completion times, including to act as a state evaluator in the wrapper search spaces. For practical purposes, the basic idea is that ATS can be generated from an attribute subset which adequately describes the currently completed incidents. From this point, ATS can be applied to predict the completion time of new incidents while they are running.

V. Experiments and Results

This section presents the used event log, experiments setup and execution details, and results obtained. A cross-validation method with 5 folds was applied to the enriched event log to build the prediction models. The ATS accuracy is given by the average MAPE of test folds in terms of mean and median of incident completion time. In addition, the ATS completeness (or non-fitting) was evaluated as the accounting of how many event records do not have a corresponding state in ATS.

A. Enriched Event Log

An enriched event log of the incident management process was extracted from an instance of the platform used by an IT company (Table I). Information was anonymised for privacy reasons. This enriched event log is composed by information gathered from the audit system and the relational model of the platform. A preprocessing step filtered out noise and sorted audit records in a sequence compatible with event log format.

Some statistical data on the enriched event log is shown in Table II. A well-defined behavior for the incident management process is observed, as most incidents (75%) go up to 7 updates, 50% up to 5 updates and on average 6 updates are needed to close incidents. However, there are outliers, with 58 as the maximum number of updates for one incident. Regarding time, the behavior resembles an exponential distribution.

B. Experiments Setup and Results

For the three experiments conducted, the ATSs parameters were: an enriched event log was randomly sampled by creating two subsets with 8,000 (A) and 24,000 (B) incidents, having $A \subset B$. The maximum horizon parameter values used were: 1 – case with the last event per incident trace; 3, 5, 6 and 7 – most frequent behaviors in this incident process (statistic ‘by incident’ in Table II); ‘infinite’ – uses all events in trace.

TABLE I
INCIDENT ENRICHED EVENT LOG EXCERPT

number	*incident_state*	*sys_updated_on*	*category*	*assignment_group*
INC001	New	3/2/2016 04:57	Internet	Field Service
	Active	3/2/2016 18:13	Internet	Field Service
	Awaiting UI	3/2/2016 19:15	Internet	Field Service
	Active	3/3/2016 12:43	Internet	Field Service
	Resolved	3/4/2016 11:02	Internet	Field Service
	Closed	3/9/2016 12:00	Internet	Field Service

TABLE II
ENRICHED EVENT LOG STATISTICS: DISTRIBUTIONS BY INCIDENT/DAY

	1st Quart.	2nd Quart.	3rd Quart.	Max.	Mean	Std. Dev.
By incident	3	5	7	58	6	3.67
By day	0.01	0.40	5.29	336.21	6.67	21.20

1) Experiment #1 – Expert-Driven Selection: first, attribute selection was driven by data about the domain held by human experts. According to ITIL best practices, in the first stage of incident management process, the caller should provide the initial information, which is complemented by the service desk agent, especially with information related to the incident category and priority. Additional information (i.e, textual descriptions) is also provided to help the support agents; its use is out of the scope of this work. Based on these practices, *incident_state*, *category* and *priority* were considered the most adequate attributes to define the process model in ATS: *incident_state* reports the stage at which the incident is; *category* indicates the type of service the incident belongs; and *priority* determines the focus requested by business. In this scenario, using event log sample with 24,000 incidents and varying the horizon and state representation parameters, 18 ATSs were generated and used as completion time predictor. The best results were obtained with horizon 3, state representation *sequence* and are shown in Table III.

TABLE III
EXPERIMENTS – AVERAGE PREDICTION RESULTS. BEST ATTRIBUTE SUBSETS SELECTED BY SPECIALIST, FILTER AND WRAPPER. LOG SAMPLE: 24,000 INCIDENTS. METRIC: MAPE. NF = % OF NON-FITTING INCIDENTS. BOLD: BEST RESULTS ON EACH EXPERIMENT.

Max	Set			Multiset			Sequence		
Horiz	*Mean*	*Median*	*NF*	*Mean*	*Median*	*NF*	*Mean*	*Median*	*NF*
Experiment #1 – Attribute subset: *{incident_state, category, priority}*									
3	106.93	77.46	0.98	91.35	75.87	1.23	**72.36**	**63.66**	**1.38**
Experiment #2 – Attribute subset: *{caller, assigned_to}*									
5	90.73	76.30	33.31	**69.69**	**57.85**	**35.67**	80.97	69.10	35.73
Experiment #3 –Attribute subset: *{incident_state, location}*									
5	**50.45**	**24.49**	**1.11**	41.90	29.35	2.30	35.09	27.28	2.74

2) Experiment #2 – Filter with Ranking: second, attribute selection was driven by filter using a ranking strategy. This approach follows consolidated concepts of specialized literature [6], [7], [9]. In this paper, ranking was applied as preprocessing, as suggested by Kohavi and John [7], to create a baseline for attribute selection, regardless of the prediction model in use. It was created through a variance analysis by correlating the independent variables (i.e., descriptive attributes) and the dependent variable (i.e., attribute 'closed', prediction target attribute). Since most of descriptive attributes are categorical, the statistic η^2 (Eta squared) was applied, as explained by Richardson [17]. As a design decision, the 15 attributes with the highest correlation were selected to compose the ranking. The variance analysis was carried out on the entire enriched event log. The attributes and correlation scores are listed in Table IV. Using ranking results filter was executed by combining the attributes as follows: *{Caller (1^{st})}*; *{Caller (1^{st}), Assigned to (2^{nd})}*; ...; *{Caller (1^{st}), Assigned to (2^{nd}), ..., Knowledge (15^{th})}*. For this scenario, 18 ATSs were generated for each attribute subset. In the second part, the best two result sets obtained in the first part where chosen to generate new sets of ATSs, however, using event log sample with 24,000 incidents (best results in Table III). The prediction results with the ranked attribute subsets were slightly worse than those obtained in the experiment #1. Analyzing the results, it is noticed that, resource-related attributes impair the generation of the prediction model, i.e., they do not reflect the process behavior with the same fidelity as the control attributes do. Particularly regarding to non-fitting, a possible explanation for these poor results could be the frequent changes in the values of the human resources assigned to solve different incidents.

TABLE IV
THE 15 DESCRIPTIVE ATTRIBUTES WITH THE HIGHEST CORRELATION WITH THE DEPENDENT VARIABLE AND THE RESPECTIVE η VALUES.

Ord	Attribute	η	Ord	Attribute	η	Ord	Attribute	η
1^{st}	Caller	0.54	6^{th}	Incident state	0.32	11^{th}	Created by	0.21
2^{nd}	Assigned to	0.37	7^{th}	Subcategory	0.32	12^{th}	Opened by	0.20
3^{rd}	Assig. group	0.35	8^{th}	Category	0.27	13^{th}	Location	0.14
4^{th}	Symptom	0.33	9^{th}	Active	0.25	14^{th}	Made SLA	0.14
5^{th}	Sys upd. by	0.33	10^{th}	Priority conf.	0.24	15^{th}	Knowledge	0.12

3) Experiment #3 – Wrappers with Hill-Climbing and Best-First: lastly, the attribute selection was driven by wrapper using a forward selection mode[1], with the hill-climbing and best-first search techniques [7]. The search space is composed by all possible combinations of the 15 attributes pre-selected by the filter with ranking strategy, i.e., attributes listed in Table IV. Since each combination represents a state in such a space, whose quality measure is calculated as the predictive power achieved by the ATS generated with the attribute subset associated with this model[2], an exhaustive search procedure is unfeasible and hence the use of a heuristic search procedures is justified. Wrapper was carried out on the enriched event log sample with 8,000 incidents. For the best-first search technique, the maximum number of expansion movements with no improvement was set to 15. Both search techniques selected this same best attribute subset: *{incident_state, location}*. Despite the high agreement between them, this can be highlighted: with hill-climbing, the stopping criterion was reached after the third expansion movement and 42 states of the search space were explored; using best-first, 17 expansion movements were done and 172 states of the search space were

[1]In the forward selection, the search start point is a singleton attribute subset to which a new attribute is incorporated at each new step in the search.

[2]The search space had $2^{15} = 32,768$ states, taking the 18 ATSs generated for each state, the range of the horizon and state representation parameters.

explored. The best results were obtained with horizon 7 and state representation *set*; however, the results obtained with the other state representations for the same horizon were very good as well. These results are significantly better than those obtained by the filter and expert-driven selections. Overall, the low non-fitting results are promising. As a second part of experiment #3, a new set of ATSs was generated using as parameters those of best results and enriched event log sample with 24,000 incidents. The best results (Table III) were obtained with maximum horizon set to 5 and overcome those obtained in the previous experiments. The MAPE results are less than half of those measures obtained by the expert-driven selection, keeping non-fitting values at the lowest level.

C. Analysis of Results

When analyzing the results, it is verified that the strategies expert-driven and filter with ranking allow us building models with a similar predictive power. However, when checking the model non-fitting capabilities, differences (1.38% and 35.67%, respectively) are observed between them for the best results. Such differences were caused due to the different process perspectives represented by the attribute subset used in each case. For the former, the ATS generation was driven by incident descriptive attributes recommended by the ITIL best practices suggested by human experts for incident clustering and routing; then, the resulting model was able to accurately represent the process. For the latter, the set of attributes automatically selected to build the ATS represents organizational and resource perspectives of the incident management process. In this case, the ATS captured the way that teams (i.e., people) act to support user requests and became highly specialized and incapable of generalizing the real process behavior. This phenomenon happens because the attributes selected represent information that presumably changes frequently (i.e, 'caller' and 'technical people' in charge of the incident). The MAPE results obtained for the experiment #1 were compared to those obtained for the experiment #2, using the paired Wilcoxon statistical test. This test showed that there is no statistical difference among the distributions of the MAPE values, seeing that, with $p_{value} = 0.3125$, it is not possible to reject the null hypothesis for equal distributions.

The wrapper-based experiment achieved an average MAPE measure (24.49) that is 38.47% of the average MAPE achieved in the expert-driven experiment. The model non-fitting continued in an even lowest level (1.11%) as that obtained in the first one. The paired Wilcoxon statistical test was applied to compare the MAPE results obtained for the experiment #1 with those obtained for the experiment #3. The null hypothesis for equal distributions can be rejected with $p_{value} = 0.0312$. This result allows to affirm that the attribute selection obtained with wrapper is better than the expert's choice.

The attribute subset selected by wrapper is the union of expert knowledge with an organizational perspective, which produced a completion time predictor with high accuracy and low non-fitting rates. Moreover, it was very similar with results obtained with the hill-climbing and best-first search techniques. This behavior has already been observed in experiments executed by Kohavi and John [7], in which, for different types of datasets, additional search effort did not produce better results.

VI. Conclusion and Future Work

Wrapper made it possible to select a set of attributes that supported a significant improvement in the accuracy of the ATS to be used as a prediction model when compared to both filter and expert knowledge. Furthermore, such search process pointed out that the maximum horizon and different types of state representations have a high influence on the prediction model results. This approach has the potential to be used as a useful preprocessing step prior to the application of other prediction methods, in addition to the ATS method used here.

As next steps, it is necessary to verify the influence of outliers throughout the process (search and prediction performance), since the results obtained in the experiments presented some variation degree. The use of other search methods such as genetic algorithms or other induction algorithms such as neural networks and the combination of the best models of ATSs with other regression models are points to be explored.

References

[1] itSMF, "Global survey on IT service management," The IT Service Management Forum, 2013, http://www.itil.co.il.
[2] M. Marrone, F. Gacenga, A. Cater-Steel, and L. Kolbe, "IT service management: A cross-national study of ITIL adoption," *Communic. of the Association for Inform. Sys.*, vol. 34, pp. 49.1–49.30, 2014.
[3] M. de Leoni, W. M. van der Aalst, and M. Dees, "A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs," *Inform. Syst.*, vol. 56, pp. 235–257, 2016.
[4] W. M. P. van der Aalst, *Process Mining – Discovery, Conformance and Enhancement of Business Processes*, 2nd ed. Springer, 2016.
[5] W. van der Aalst, M. Schonenberg, and M. Song, "Time prediction based on process mining," *Inform. Syst.*, vol. 36, no. 2, pp. 450–475, 2011.
[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. of Machine Learning Res.*, vol. 3, pp. 1157–1182, 2003.
[7] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intel.*, vol. 97, no. 1, pp. 273–324, 1997.
[8] W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, and C. W. Günther, "Process mining: A two-step approach to balance between underfitting and overfitting," *Software & Systems Modeling*, vol. 9, no. 1, 2008.
[9] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intel.*, vol. 97, no. 1-2, pp. 245–271, 1997.
[10] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *Int. J. of Forecasting*, vol. 8, no. 1, pp. 69 – 80, 1992.
[11] A. de Myttenaere, B. Golden, B. L. Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.
[12] M. Polato, A. Sperduti, A. Burattin, and M. de Leoni, "Data-aware remaining time prediction of business process instances," in *Proc. of the 2014 Int. Joint Conf. on Neural Netw.* IEEE, 2014, pp. 816–823.
[13] A. Rogge-Solti, L. Vana, and J. Mendling, "Time series petri net models – enrichment and prediction," in *Proc. of the 5th Int. Symp. on Data-driven Process Discovery and Analysis (SIMPDA)*, 2015, pp. 109–123.
[14] M. Hinkka, T. Lehto, K. Heljanko, and A. Jung, "Structural feature selection for event logs," pp. 20–35, 2017.
[15] J. Evermann, J.-R. Rehse, and P. Fettke, "Predicting process behaviour using deep learning," *Decision Supp. Sys.*, vol. 100, pp. 129 – 140, 2017.
[16] N. Tax, I. Verenich, M. La Rosa, and M. Dumas, "Predictive business process monitoring with lstm neural networks," pp. 477–492, 2017.
[17] J. T. E. Richardson, "Eta squared and partial eta squared as measures of effect size in educational research," *Educational Research Review*, vol. 6, no. 2, pp. 135–147, 2011.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 683–689
DOI: 10.15439/2018F115
ISSN 2300-5963 ACSIS, Vol. 15

Scoring method versus TOPSIS method in the evaluation of e-banking services

Witold Chmielarz
University of Warsaw, Faculty of Management, in Warsaw
ul. Szturmowa 1/3, 02-678 Warsaw, Poland
Email: witold@chmielarz.eu

Marek Zborowski
University of Warsaw, Faculty of Management, in Warsaw
ul. Szturmowa 1/3, 02-678 Warsaw, Poland
Email: mzborowski@wz.uw.edu.pl

***Abstract*—The main objective of this article is to compare of the utility of the multi-criteria TOPSIS method with scoring method and scoring method with preferences. The data was taken from Poland banking sphere in 2017. e-Banking websites were assessed from the point of view of an individual client. e-Banking helps to strengthen the bank's position in a competitive market environment; therefore, the high quality of a website frequently plays a decisive role in the perception of a bank as an organization. In the research the authors have used a scoring method, a scoring method with preferences and – in order to verify the correctness of the results and compare them – the TOPSIS method. An additional problem which appeared in the course of the study was the questions of whether the sophisticated multi-criteria methods produce better quality results than simple methods based on a scoring evaluation. Subsequently, on the basis of the obtained findings, the authors have carried out comprehensive and multi-dimensional analyses and presented the conclusions and recommendations drawn on the basis of the aforementioned analyses. The authors' contribution to the research was specifying the criteria for the evaluation of the websites as the main indicators of the perception of the quality of websites, identifying the best e-banking websites and formulating conclusions that may become a starting point for creating an effective quality management systems of e-banking services.**

I. Introduction

THE importance of electronic banking in Poland is evidenced by the constant pace of its development. In relation to the fourth quarter of 2016 the number of individual clients who have potential access to account increased in relation to the fourth quarter of 2017 by over 7%, reaching 35,512 million users (92% of the population), where the number of active individual clients rose by more than 3.5%, reaching the level of 15,889 million (45% of the population) [9]. It is the fastest growing banking sector – nothing seems to interfere with these positive trends. Over the last ten years, the number of customers has increased by more than 25 million. Thus, it is the sphere, the development of which should be carefully examined and analysed.

The problems related to the functioning of websites, in particular, access to e-banking services, are widely discussed in the literature on the subject, and there is no single formula which would allow their unambiguous assessment and the improvement of their quality. Numerous analyses also do not indicate what effect they have on the development of banking in the countries where they are being examined. There is an ongoing process of searching for the method which would best reflect the tendencies in this sphere and at the same time would be most convenient from the point of view of its users. Literature review shows that bank websites may be analysed from the point of view of their usability (site map, address catalogue) [2], functionality (search, navigation, content) [19], interactivity (accessibility and responsiveness) [14], [18] visualization (colour scheme, background, graphics, text) [7], reliability [1], cost-effectiveness (costs of purchase, transport, the difference in prices in traditional and online shops) [3].

Most of the evaluation methods of e-banking websites are traditional scoring methods based on specific criteria sets, evaluated according to a fixed scale. Among the criteria which are most frequently applied there are technical and functional criteria. Many of them contain factors which may be evaluated in a highly subjective way: text clarity, the attractiveness of the colour scheme, images and photos, the speed of finding specific functions and using them) etc. In addition, some users do not treat particular criteria sets in an equivalent way. On the other hand, there are also numerous problems with determining preferences and relations between them. These problems – according to relevant literature on the subject – are solved by multi-criteria methods. However, the question arises whether indeed their more complicated use may in some way be compensated for when compared to the ease and convenience in using simple methods. The authors will attempt to address this question in the article.

II. Description of the research method and the sample

The research in this paper has been conducted using the authors' own criteria sets used for electronic access to accounts of particular banks. The criteria sets were applied since 2006 and they were created on the basis of relevant literature and verified following consultations with the experts. The evaluation criteria were established during an internet discussion conducted with the participation of scientists and researchers representing leading universities dealing with electronic banking in Poland, based on the literature on the subject. At the moment of economic crisis of 2008, a set of anti-crisis criteria, i.e. the selected measures which – in the experts' opinion – were supposed to counteract the potential effects

of the banking crisis [15] - was added to the set of evaluation criteria used to assess the access to banking services. The second modification took place in 2017 where the authors verified the correctness, comprehensibility and importance of the se-lected criteria for the users with the participation of 244 respondents. Finally, after this verification and consideration of users' comments, the criteria adopted in the studies into the evaluation of banking websites were divided into three main groups: economic criteria, technical, visualisation and security criteria, anti-crisis measures. The respondents evaluated their preferences with regard to criteria groups as well as individual criteria. Specific criteria with preferences calculated as an arithmetic mean of the scores, motivation and justification of their choice are presented in details in [16].

Among the groups, the most important set were economic criteria which obtained on average a 62% score (including the most important criterion – account maintenance PLN/month (average: checking and savings account) – nearly 8%), subsequently, technical and security criteria – on average 32% (the most important security measures 6%) as well as anti-crisis measures, on average - 6%.

The presented study constitutes the next stage of the research carried out systematically from 2006 whose primary objective is to evaluate the factors which impact the quality of websites that provide access to individual accounts in banks. Frequently, it is the quality of the website which turns out to be decisive in retaining and acquiring new customers. It is important to notice that the present ranking evaluating the quality of e-banking websites includes also economic factors which are the specific reflection of the current bank policy. In order to evaluate particular criteria in the banks which were selected by the clients, the authors used a standardised, simplified Likert scale [10], in which lack of a particular quality is represented by the value equal to zero, its complete fulfilment is equal to one, average fulfilment of the feature – 0.5 and intermediary values such as good fulfilment is equal to 0.75; and sufficient fulfilment amounts to 0.25.

The study has been conducted with the initial application of a simple scoring method and a scoring method with preferences. In the simple scoring method, the authors measure the distance from the maximum value which can be obtained (according to the adopted scale). It concerns the value of the measure of the criterion and in the sense of a distance, it is the same when we measure the distance from the first and second criterion and vice versa. However, the relationship between individual criteria is not determined. Assigning the preference scale, which adds up to the value of 100%, to particular criteria (or criteria groups) can be regarded as such a measure. The normalised linear preference scale determines the participation of particular criteria in the final score. It is important to indicate that scoring methods are seen as subjective evaluation methods, even though their subjectivity appears to be limited together with the number of the interviewed respondents and the application of a preference scale. Despite their drawbacks, these methods are commonly applied and their scores are easy to interpret. The methods which are believed to be more objective, for example, AHP method [11], Promethee II, Electre I and III method, the TOPSIS method and other solutions are rather complex to use and sometimes it is difficult to interpret their findings. The authors' experience, mainly related to the application of AHP method used to evaluate websites, points to the fact that the completion of survey questionnaires is very complicated from the perspective of the respondents participating in the studies. As a result, this may lead to ill-considered and random assessments, and the final scores may frequently be determined by the order of particular criteria. In order to eliminate such prob-lems, the authors have devised their own evaluation method – a conversion method. The data which are used in calculations are collected in the form of the same input tables as in the case of a scoring method. This method combines the simplicity and unambiguity of a scoring method with the precision of relational methods. It consists in establishing the relations of each criterion in relation to other criteria, based on averaged distances from the potential maximum value previously established on the basis of a scoring method.

Based on the above assumptions, in December 2017, the authors have conducted the research into the quality of the e-banking websites of the banks which are most popular among individual clients in Poland. The sample of the study covered 721 respondents. Among them, there were 83 (nearly 12% of the population) people holding and evaluating two accounts in two different banks, 38 respondents (5%) having and assessing three accounts in three different banks. In total, the survey participants carried out 1002 evaluations of 28 banking websites. Among the 28 websites, seven responses concerned one bank, none of them was complete and correct, and thus the authors used 21 banks in further analyses. Correct responses were provided by 290 individuals (40% of the respondents), out of which 16 (almost 6%) people evaluated two websites, and four participants (over 1%) assessed three of them. In total, there were 334 fully and correctly completed evaluations of banking websites (33% of all completed survey questionnaires). The participants were 19-50 years old students from randomly selected students groups. More than 98% of respondents were 18-25 years old, which could have influenced the results of the survey (15.6% of the population in Poland are potential clients of e-banking, including over 50% of active clients in 2016). Among the survey participants, there were 72% of women and 28% of men. The majority (55%) described themselves as working students, 45% as students. Most people (26%) stated that their place of birth was a town below 50,000 inhabitants, almost the same number of respondents - cities with more than 500,000 residents, and 23% - villages.

The greatest number of electronic access accounts was indicated in the case of the clients of mBank (15%), then iPKO PKO BP S.A. (13%) and Millenium (12%). The smallest shares in the examined group were clients holding accounts in: BGŻ Optima and Orange Finance (each approximately 1%). The spread between the smallest and the largest share of electronic access to accounts in particular banks in the entire sample amounts to 14%. Only in six out of twenty-one banks,

the participation of clients was above the average amounting to 5%.

III. Comparative analysis of internet access to accounts in electronic banking with the application of a scoring method

In the analysis with the application of a scoring method, the authors used input tables, where each of the clients evaluated bank offers concerning the selected e-banking services and the fees related to using bank accounts which can be managed via the Internet. Next, on the basis of the completed surveys, the authors created one summary table of averaged criteria evaluations generated by the users. On this basis, it was possible to carry out analyses and discuss the obtained findings. There occurred a great discrepancy in the evaluations of the analysed banks. In 2017 it amounted to nearly 13 percentage points (as compared to 2.25 percentage points in 2008), which confirms the thesis that the period of crisis in 2008 increased the radicalism of the evaluations and increased the requirements with regard to tools providing access to account. The best in the ranking were: Orange Finanse (81.80%) and Bank Millenium S.A. (80.12%). The next positions were taken by ING Bank Śląski and Raiffeisen Bank. Interestingly, the first place was taken by a mobile bank which was created on the basis of the cooperation between the most innovative bank, i.e. mBank (taking the fifth position in the ranking) and one of the largest mobile operators, namely, Orange, on the basis of mBank experience. The worst in the ranking were: Bank Pocztowy S.A. and Credit Agricole Polska S.A.. The first thirteen banks in the ranking obtained the scores which were above the average amounting to 76.77%.

In the analysed banks, the transfer to a bank where we hold an account is evaluated as average (over 87%) and to a different bank (over 84%). The service of issuing a debit card is evaluated at a slightly lower level (over 83%), similarly to many different access channels (over 82%). The exceptionally low interest rates on deposits and relatively high-interest rates on credits in the analysed banks obtained the lowest scores (approximately 64% of the maxi-mum possible scores). It emerges that the spread between the highest and the lowest scores was relatively high and amounted to nearly 24 percentage points. The scores recorded in the case of the average interest rate of current and savings accounts were alarmingly low and amounted to 67-69%, which undoubtedly does not motivate the users to save money. The discrepancy be-tween the lowest and the highest scores was relatively high and amounted to nearly 24 percentage points. In total, thirteen evaluation criteria were above the average equal to 76.77%, and only ten were below it. It may appear that the respondents generally have a high opinion of the bank websites since all the criteria were rated above the 50% of the maximum score. However, since websites in Poland strongly compete with each other for many years, the scores should not be seen as satisfactory. The first of the banks in this year's ranking was a new player in the electronic banking market which gained its position owing to banking application for smartphones and tablets. In recent years, however, the highest scores were recorded in the case of the banks holding an established position in traditional internet banking such as ING Bank Śląski S.A., Bank BPH or BZ WBK. Among the first ten positions, there were banks such as Millenium and Getin Bank, as well as the banks which started to implement electronic banking and which have their loyal customers, especially those falling in the middle-age range. The high – fourth position was taken by Raiffeisen Bank, which probably resulted from the introduction of numerous modernizations and innovations carried out in recent years and an ongoing advertising campaign.

IV. Comparative analysis of electronic access to the account in e-banking with the application of the scoring method with preferences

One of the methods which allow for limiting the specific subjectivity of the experts' and users' evaluations in the scoring method is applying unitary preferences with regard to particular criteria or criteria groups. In the study, the authors divided the criteria into three groups: economic, technical and anti-crisis measures. The fourth group adopted in the study was created according to the preferences of the clients indicated in the research preceding the analyses. In this variant the criteria are as follows: economic criteria are preferred in 62%, technological in 32% and anti-crisis factors only in 6%. For each of the remaining groups, the authors adopted a variant with a group of dominating criteria: economic criteria (70% for economic criteria and 15% in the case of each of the remaining ones), technological, visualisation and security criteria (70% for technological, visualisation and security criteria, 15% for each the remaining ones), anti-crisis criteria (70% for anti-crisis criteria, and 15% for each of the remaining ones).

In the first case – of economic preferences – the three leading positions are taken by Orange Finance in the first place, Raiffeisen Bank (which moved from the fourth position), and Nest Bank (which moved from the tenth place). In the technical, visualisation and security variant, the first position is also taken by Orange Finance, mBank moved to the second place, and ING Bank Śląski is next. In the variant connected with the anti-crisis measures the second place is taken by Raiffeisen Bank, and Bank Millenium ranks next. In the last variant – the users' variant – the first place is still occupied by Orange Finance, and subsequent positions are taken by Bank Millenium and Raiffeisen Bank. The results of the rankings with particular types of preferences have significantly changed the order in the ranking and have shown the advantage of particular characteristics in the considered banks.

V. Verification of the ranking of internet access to an e-banking account with the application of the TOPSIS method

The theoretical assumptions of the TOPSIS method are presented below. In order to evaluate 21 most popular e-banking websites in Poland in 2017 (A1, . . . , A21), the authors used the set of 20 criteria which were adopted by the users

(C1, . . . , C20).
With the use of the MCDA (Multi-Criteria Decision Analysis) selection frameworks provided in [4], [5], [6] the authors chose the TOPSIS method to perform the empirical research. The obtained closeness coefficient CC_i is the score value produced by the TOPSIS method and is used to construct the ranking of alternatives (see table 1).

To carry out an analysis based on the TOPSIS method, the authors used the input tables where each of the clients evaluated the bank offers related to selected e-banking services and fees connected with using bank accounts which can be managed via the Internet. Subsequently, on the basis of the completed survey questionnaires, they created one summary table of the averaged criteria evaluations generated by the users. On this basis, the authors were able to carry out relevant analyses and discuss the obtained scores.
In the empirical study, 21 banks (A1 – A21) were evaluated with the use of 20 criteria (C1 – C20). The constructed decision matrix is presented in Table 1. The preference direction for all the criteria was set to maximum. In the first step of the research, a ranking was created based on the weights obtained with the use of normalized means of all users opinions. The obtained ranking is presented in table 2. The analysis of the ranking allows to observe, that the leading alternative A19 (Raiffeisen Bank Polska SA) obtained approximately four times more score than the worst alternative A4 - Bank Pocztowy SA - (0.6406 compared to 0.1578). On the other hand, the second and third alternative in the rank (A16 - Orange Finance - and A2 - Bank Millennium SA - respectively) obtained only slightly less score (0.6337 and 0.6096) than the leading alternative. The average score was 0.4703. Twelve alternatives obtained more than average score and nine alternatives scored worse than average.
Subsequently, the authors studied how the ranking would change if the weights of the criteria were not taken into account. Therefore, in the second step of the research, each criterion obtained an equal score. The produced ranking is presented in table 2. Again, the alternative A19 (Raiffeisen Bank Polska SA) took the first position in the ranking, and the alternative A4 (Bank Pocztowy SA) the last one. However, the alternatives A16 and A2 switched places in the new ranking. The analysis of the table 2 allows to note that as much as 7 alternatives remained unchanged: A19, A17, A7, A20, A6, A9, A4. On the other hand, the alternative A15 (Nest Bank) underwent the most significant change, from position 4 to 9.

The results of the calculations presented in table 2 show that the use of the TOPSIS method produces basically similar results with regard to the ranking as using the scoring method. This confirms the thesis that if the initial set is not greatly diversified, then the application of the simpler method is comparable to the use of more complex methods (here: the multi-criteria TOPSIS method), and it does not require any additional complicated calculations. The interpretation of the findings is also equally possible and convenient. At least – this represents a greater possibility to differentiate the input data in order to examine different hypotheses concerning the distribution of preferences between the criteria of particular groups.
If we compare the scores obtained in the scoring method and the TOPSIS method without the consideration of weight differentiation (carried out with the participation of the research sample), then – as indicated previously, the scores concerning the positions in the ranking are similar, despite the differences in the presentation standard (in the scoring method, the point of reference is the maximum possible level of the quality). Nevertheless, the differences in the rankings are really small. In 21 banks – for equal weights – they occur only in ten cases, and the greatest difference for Raiffeisen Bank amounts to the change of three positions. Fourteen differences (out of 21) occur when comparing the version with preferences. The biggest difference (four places) is recorded in the case of BGŻ Optima. In total, the differences are not as great as those in the case of comparison with the conversion method or the AHP method. In general, apart from the spectacular advancement of Raiffeisen Bank, or the lower position of BGŻ Optima, the differences are on average, the move of one place in the ranking. For equal weights, there is a considerable difference in the spread in the ranking for the TOPSIS method – more than 34% and simultaneously, twenty percentage points smaller (less than 10%) for the scoring method. Even greater differences between the high-est and the lowest values are indicated in the case of the version with preferences – for the TOPSIS method amounting to over 34%, for the scoring method to less than 4%.

VI. Conclusions

The presented analysis has shown the diversity of the opinions of individual clients on the usability of e-banking websites, in particular, their views concerning the selection and use of websites to meet the daily needs of users related to banking services. At this point, it is important to indicate that the demand is high and greatly diversified. At the end of 2017 in Poland there were 32.6 million of e-banking clients, including 14.7 million of active customers (at least one contact with checking and savings account per month) [12], including 8.9 million of clients using mobile devices to contact the bank (website or application) [13], including 2.2 million users of strictly mobile banking (only via a smartphone application) [8]. Even in 2017, we dealt mainly with the first trend, yet the bank analysts predict that this year the remaining trends will be taking a dominating position.
In this paper, the authors have not differentiated the clients with regards to the devices they use and the tools by means of which they contact their banks. Nevertheless, they evaluated them from the point of view of the device which made it possible to communicate with the bank. The evaluation of the devices allowed drawing the following main conclusions:

- it appears that mobile access to banking services is the most important phenomenon in the electronic banking market. This is evidenced by the position of Orange Finance in this ranking.

TABLE I
THE DECISION MATRIX – INPUT FOR THE TOPSIS METHOD; SOURCE: THE AUTHORS' OWN WORK

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**
A1	3.5232	4.1151	4.4438	4.2822	4.0845	4.1656	3.7889	3.3620	3.2592	3.2759	3.9667
A2	3.5702	4.1488	4.5620	4.3802	4.2314	4.3471	3.8430	3.4876	3.4215	3.2810	4.2231
A3	3.1429	4.1429	4.0000	3.9286	3.8571	4.1429	3.4286	3.2857	2.9286	2.8571	3.5714
A4	3.1739	3.5652	4.1304	3.8261	3.5652	3.7826	3.4348	3.1304	3.0870	3.0870	3.4783
A5	2.8421	3.9474	4.4737	4.2632	4.3158	4.2632	3.4737	2.8947	3.0000	2.9474	3.5263
A6	3.3000	4.0429	4.3286	4.0286	4.0000	3.9143	3.6571	2.9714	3.0000	2.9714	3.9143
A7	3.3684	3.9211	4.5000	4.3684	3.9474	4.2105	3.9737	3.1316	3.1053	3.1842	3.9737
A8	4.0000	4.0000	4.4444	4.2222	3.5556	4.1111	3.3333	3.8889	4.0000	3.8889	3.6667
A9	3.2667	4.0000	4.1667	3.8000	3.8667	4.1000	3.7000	2.9333	3.1667	3.0333	3.7667
A10	3.4516	3.6129	3.9677	3.7742	3.9355	4.0968	3.6774	3.1935	3.1290	3.1613	3.7742
A11	3.1818	3.6364	4.7273	4.8182	3.9091	4.1818	3.6364	3.5455	3.3636	3.3636	4.0000
A12	3.7857	4.0714	4.3571	4.5000	4.0000	4.3571	3.9286	3.6429	3.1429	3.0000	3.7857
A13	3.5217	4.1304	4.4239	4.2935	4.1739	4.2500	3.7174	3.5435	3.2174	3.2826	3.8696
A14	3.4805	4.1364	4.4610	4.3312	4.0974	4.1948	3.7792	3.3052	3.2208	3.2987	3.7597
A15	3.6129	4.4516	4.3871	4.1613	4.1613	4.3226	4.4194	3.6774	3.5161	3.2258	4.1935
A16	3.8000	4.7000	4.8000	4.6000	4.2000	4.7000	4.3000	3.2000	2.8000	2.9000	4.5000
A17	3.5188	4.0602	4.3083	4.1353	3.9248	3.9549	3.7444	3.2932	3.1353	3.2481	3.9173
A18	3.7143	4.0000	4.1429	4.2143	3.9286	3.9286	3.7857	3.5000	3.1429	3.0000	3.5000
A19	3.8158	4.3684	4.5000	4.2105	4.1579	4.0526	3.8421	3.6316	3.5000	3.6842	4.0000
A20	3.3043	4.0000	4.4348	4.2609	4.0000	4.0870	3.7826	3.4348	2.9130	2.9565	3.4348
A21	3.1818	4.1818	4.4545	4.3636	4.0000	4.6364	4.0909	3.5455	3.3636	3.5455	4.0000
	C10	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**
A1	3.2759	3.9667	3.9993	4.3634	4.1983	4.1147	3.9995	4.1010	4.1614	4.3418	3.7495
A2	3.2810	4.2231	3.9008	4.3223	4.1736	4.0661	3.9339	4.0909	4.1570	4.3058	3.6777
A3	2.8571	3.5714	3.7857	4.2857	4.2143	3.7857	3.5000	3.5714	3.7143	3.6429	3.5000
A4	3.0870	3.4783	3.2174	3.4348	3.5652	3.3478	3.3913	3.4783	3.3478	3.4348	3.3913
A5	2.9474	3.5263	3.5263	4.5263	4.3158	3.8421	3.7368	4.0526	4.0000	4.0526	3.7368
A6	2.9714	3.9143	3.6714	4.1286	3.9143	3.6714	3.6714	3.9000	3.8857	3.9286	3.4571
A7	3.1842	3.9737	3.6579	4.2105	4.0000	3.8158	3.9474	4.0526	3.8158	3.9737	3.6579
A8	3.8889	3.6667	3.6667	4.0000	4.2222	3.6667	3.4444	3.4444	3.6667	3.7778	4.0000
A9	3.0333	3.7667	3.5333	4.1000	3.6000	3.3000	3.5000	3.3667	3.8333	3.7667	3.2000
A10	3.1613	3.7742	3.7742	4.0645	3.8065	3.5806	3.7419	3.6452	3.8065	3.6452	3.4839
A11	3.3636	4.0000	3.9091	4.0000	3.9091	4.0909	4.1818	4.0909	4.2727	4.0909	3.9091
A12	3.0000	3.7857	3.7857	4.1429	4.1429	4.0000	3.7143	3.7143	3.7857	4.0000	3.6429
A13	3.2826	3.8696	4.0217	4.3696	4.0870	4.3152	4.1630	4.2174	4.2391	4.3478	3.7283
A14	3.2987	3.7597	4.0519	4.4221	4.1883	4.2403	4.1623	4.2273	4.2143	4.4416	3.7662
A15	3.2258	4.1935	3.3871	3.7419	3.6452	3.6129	3.8065	4.0323	3.9032	3.9677	3.5806
A16	2.9000	4.5000	3.7000	4.2000	4.1000	4.4000	4.6000	4.3000	4.2000	4.3000	3.5000
A17	3.2481	3.9173	4.0451	4.3459	4.2331	4.0376	3.9023	3.9850	4.1128	4.2782	3.8045
A18	3.0000	3.5000	3.5714	3.7857	3.9286	3.4286	3.5714	3.6429	3.7857	3.7857	3.7857
A19	3.6842	4.0000	3.9737	4.1316	3.8947	4.0526	4.0789	4.0263	4.0000	4.0263	3.8684
A20	2.9565	3.4348	3.6087	3.9130	3.9565	3.7826	3.6957	3.7826	3.9565	4.0000	3.3043
A21	3.5455	4.0000	4.0000	3.9091	4.1818	4.1818	4.0000	4.2727	4.0000	3.4545	3.8182

TABLE II
COMPARISON OF RANKINGS FOR THE 21 BANKS, OBTAINED WITH THE USE OF COMPUTED WEIGHTS AND EQUAL WEIGHTS FOR TOPSIS AND SCORING METHOD AND SCORING METHOD WITH PREFERENCES; SOURCE: THE AUTHORS' OWN WORK

Method		**TOPSIS Computed Weights**		**TOPSIS Equal Weights**		**Scoring method with preferences User Weights**		**Scoring method Equal Weights**	
Alternative		**TOPSIS computed weights**	**Rank**	**TOPSIS equal weights**	**Rank**	**Scoring method with preferences user weights**	**Rank**	**Scoring method equal weights equal weights**	**Rank**
A1	Alior Bank SA	55.54%	8	0.5795	6	0.3192	8	79.30%	6
A2	Bank Millennium SA	60.96%	3	0.622	2	0.3246	2	80.12%	2
A3	Bank Ochrony Środowiska SA	33.34%	18	0.3279	19	0.2955	19	73.29%	19
A4	Bank Pocztowy SA	15.78%	21	0.1456	21	0.2812	21	68.87%	21
A5	Bank Polska Kasa Opieki SA	37.37%	15	0.4106	14	0.3036	14	75.74%	14
A6	Bank Zachodni WBK SA	35.81%	17	0.3597	17	0.3006	17	74.36%	16
A7	BGŻ BNP Paribas SA (BNP)	46.06%	13	0.4689	13	0.3111	13	76.82%	13
A8	BGŻ Optima	55.71%	7	0.5279	10	0.3152	11	77.00%	12
A9	Credit Agricole Bank Polska SA	29.05%	20	0.281	20	0.2937	20	72.00%	20
A10	Deutsche Bank Polska SA	32.42%	19	0.3367	18	0.2969	18	73.32%	18
A11	Euro Bank SA	50.29%	12	0.5554	8	0.3181	9	78.82%	8
A12	Getin Noble Bank SA	53.29%	10	0.5068	12	0.3153	10	77.50%	11
A13	ING Bank Śląski SA	56.24%	5	0.6023	4	0.3212	5	79.91%	3
A14	mBank	55.02%	9	0.5905	5	0.3199	7	79.78%	5
A15	Nest Bank	58.38%	4	0.5419	9	0.3203	6	77.81%	10
A16	Orange Finance	63.37%	2	0.62	3	0.3318	1	81.80%	1
A17	PKO Bank Polski SA (iPKO)	50.67%	11	0.5236	11	0.3129	12	77.98%	9
A18	PKO Bank Polski SA (INTELIGO)	41.78%	14	0.3768	15	0.3018	16	74.14%	17
A19	Raiffeisen Bank Polska SA	64.06%	1	0,6375	1	0.3244	3	79.82%	4
A20	T-Mobile Usługi Bankowe	36.74%	16	0.3721	16	0.3027	15	74.61%	15
A21	Volkswagen Bank Polska SA	55.76%	6	0.5769	7	0.3217	4	79.18%	7

- the position of Credit Agricole Bank Polska has significantly decreased in relation to other rankings, and the bank, when compared to previous rankings, lost its position among the top ten banks which obtained the best scores, similarly to T-Mobile Usługi Bankowe,
- the vast majority of active bank clients (62%) believe that economic criteria, i.e. the first three positions among all the most frequently used services are the most significant criteria in the evaluation of internet access to banking,
- however, more and more people admit that they are inclined to consider the ease of access to mobile banking (nearly 80%) and the number of access channels (82%) when selecting a given website,
- the issues related to anti-crisis measures also fell below the average (73%), and it emerges that users slowly start to forget about the crisis of 2008,
- the scale of inactive clients (approximately 55% appears to be alarmingly large in relation to those customers who can potentially use electronic banking. It is true that a few years ago the estimates did not exceed 20%, but the pace of increase in the customer activity in this field is still very slow.

The increasing diversification of banking services necessitates new approaches to the use of tools to assess their suitability and usability for the clients. The calculations obtained in the study carried out with the use of the scoring method and the TOPSIS method create the initial basis for such comparisons. Taking into consideration the basic features of these methods one can conclude that generally the obtained findings are largely similar. Therefore, the general conclusion is that in the case of large sets of homogeneous, uniform data both of these methods seem to be equivalent, and in the analyses it is recommended to use the simplest possible methods because they offer greater possibilities of "manual" analyses. The basic features which are characteristic of both methods are presented in table 3.

The fact that starting from last year, the long-awaited discrimination between banking services via mobile devices and mobile banking induces the authors to conduct thorough analyses of the "strictly" mobile banking carried out by means of the applications running on smartphones and tablets. The diversity in the sphere of banks operating independently, or in alliances with mobile operators also necessitates the consideration of the justification of making a separate evaluation of e-banking, e-banking used via mobile devices and mobile banking. The problem consists in the fact that clients who use mobile devices are not always fully aware that connecting to a website using a mobile device is not mobile banking. The second problem is that in the course of the previously conducted studies [17], clients claimed that they only engage

TABLE III
COMPARATIVE CHARACTERISTICS OF THE POSSIBILITIES TO USE THE SCORING METHOD AND THE TOPSIS METHOD; SOURCE: THE AUTHORS' OWN WORK

Criterion	Scoring method	TOPSIS method
Method	Simple method with the possibility to apply user preferences	Multi-criteria method with equal weights and calculated preferences
Obtaining input data	Easy	Easy
Initial data processing	Easy	Easy
Computing method	Easy	Relatively more difficult
Interpretation of the findings	Easy	Easy
Criterion	Scoring method TOPSIS	method
Extended analyses	Easy	Relatively more difficult

in low-value transaction when using a smartphone (by means of a website or application), and the remaining operations are carried out by means of personal and desktop computers, frequently not noticing or recording which of these transactions are conducted by means of applications. Thus, this area requires continuous and ongoing research in the field.

REFERENCES

[1] F. J. Miranda, R. Cortes, C. Barriuso, "Quantitative Evaluation of e-Banking Web Sites: an Empirical Study of Spanish Banks", in: *The Electronic Journal Information Systems Evaluation*, No. 2(9), 2004, pp. 73–82, http://www.ejise.com/issue/download.html?idArticle=766.

[2] H. Bauer, M. Hammerschmidt , T. Falk, "Measuring the Quality of E-Banking Portals - an Empirical Investigation", *International Journal of Bank*, Vol. 23, No. 2, 2005, pp. 153–175; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=962227.

[3] H. W. Webb, L. A. Webb, "SiteQual: an integrated measure of Web site quality", in: *Journal of Enterprise Information Management*, no. 6, vol. 17., 2004.

[4] J. Wątróbski, "Outline of multicriteria decision-making in green logistics", in: *Transportation Research Procedia*, 16, 2016, pp. 537–552.

[5] J. Wątróbski, J. Jankowski, "Guideline for MCDA method selection in production man-agement area", in: *New frontiers in information and production systems modelling and analysis*, Springer, Cham, 2016, pp. 119-138.

[6] J. Wątróbski, J. Jankowski, Z. Piotrowski, "The selection of multicriteria method based on unstructured decision problem description", in: *International Conference on Computational Collective Intelligence*, Springer, Cham, 2014 pp. 454–465.

[7] M. B. Mateos, A. C. Mera, F. J. Gonzales, O. R. Lopez, "A New Web Assessment Index: Spanish Universities Analysis", Internet Research: *Electronic Application and Policy*, no. 3, vol. 11., 2001.

[8] mBank, "Grupa mBanku - wprowadzenie, Najbardziej udany wzrost organiczny w Polsce", https://www.mbank.pl/pobierz/msp-korporacje/wyniki-finansowe/introduction_to_mbank_pol_2015_q3.pdf?noredir.

[9] NetBank, "Bankowość Internetowa i Płatności Bezgotówkowe IV kwartał 2017 r.", https://zbp.pl/public/repozytorium/wydarzenia/images/marzec_2018/konf/Netbank_Q4_20180329.pdf.

[10] R. Likert, "A Technique for the Measurement of Attitudes:, in: *Archives of Psychology*, No. 140, 1932.

[11] T. Saaty, "How to Make a decision. The Analytic Hierarchy Process", in: *European Journal of Operational Research*, Volume 48, Issue 1, 5, 1990, pp. 9–26.

[12] W. Boczoń, "Raport PRNews.pl: Liczba klientów mobile only – IV kw. 2017", https://prnews.pl/raport-prnews-pl-liczba-klientow-mobile-only-iv-kw-2017-433554.

[13] W. Boczoń,"Raport PRNews.pl: Rynek bankowości mobilnej – IV kw. 2017", https://prnews.pl/raport-prnews-pl-rynek-bankowosci-mobilnej-iv-kw-2017-433527.

[14] W. C. Chiou, C. C. Lin, C. Perng, "A strategic framework for website evaluation based on a review of the literature from 1995–2006", *Information & Management*, no. 5-6, vol. 47, 2010,

[15] W. Chmielarz, "Methodological Aspects of the Evaluation of Individual E-Banking Services for Selected Banks In Poland", chapt. 11 in: *Infonomics for Distributed Business and Decision-Making Environments. Creating Information System Ecology*, ed. M. Pańkowska, 2010.

[16] W. Chmielarz, M. Zborowski, "Analysis of e-Banking Websites' Quality with the Appli-cation of the TOPSIS Method – A Practical Study", *International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2018, in press.

[17] W. Chmielarz, M. Zborowski, "Comparative Analysis of Electronic Banking Websites in Poland in 2016", in: *Information Systems: Research, Development, Applications, Education*, eds. S. Wrycza, J. Maślankowski, Lecture Notes in Business Information Processing, 10th SIS-SAND/PLAIS EuroSymposium 2017 Gdańsk, Poland, Springer International Publishing, 2017, pp. 43–56.

[18] Y. K. Migdadi, "Quantitative Evaluation of the Internet Banking Service Encounter's Quality: Comparative Study between Jordan and the UK Retail Banks", in: *Journal of Internet Banking and Commerce*, no. 2, vol. 13., 2008.

[19] Z. Yang, S. Cai, Z. Zhou, N. Zhou, "Development and validation of an instrument to measure user perceived service quality of information presenting Web Portals", in: *Information & Management*, no. 4, vol. 42, 2005, pp. 234–254.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 691–695
DOI: 10.15439/2018F67
ISSN 2300-5963 ACSIS, Vol. 15

Analysis of Selected Internet Platforms of Distributors of Computer Games in the Assessment of Users

Witold Chmielarz
University of Warsaw Faculty of Management ul.
Szturmowa 3, 02-678 Warszawa, Poland
Email: witold@chmielarz.eu

Oskar Szumski
University of Warsaw Faculty of Management ul.
Szturmowa 3, 02-678 Warszawa, Poland
Email: oskar.szumski@uw.edu.pl

***Abstract*—The aim of this article is to execute a comparative analysis of services and distribution platforms used to purchase computer games. The article is a continuation of research focused on the popularity, use and impact of games on the behavior of a prominent population and analysis of the typical e-shops with games. For the analysis it was chosen the four most common amongst students computer game services and platforms simultaneously found in the first hundreds of searches in Google. CAWI analyzes were used for nine groups of features: transparency, site navigation, quality of information, product search quality, advertising, the quality of the order processing, way of visualization and product promotion from the customer's perspective. Platform analysis was carried out according to: installation package support, application design, search engine quality, transparency, product presentation and security. A qualitative study was conducted to select a sample of selected students using the point method and the point method with preferences to evaluate the distinctive features of the services. A group of over seven hundred randomly selected people from the university was examined. This approach was based on the structure of the article consisting of the presentation of the research hypothesis, the description of the methodology and the research sample, and the analysis of the results and their discussion together with the resulting conclusions. The results of the work may be used by the owners of computer distribution services or platforms and website designers.**

I. Introduction

THE MAIN objective of this study is to analyse the quality of selected Internet platforms of distributors of computer games. This is the fifth of the series of studies carried out among the representatives of academic youth concerning the possibilities of using computer games in education, entertainment, sport, etc. This time the study focuses on the potential distribution platforms providing access to various types of games. The digital distribution of computer games is considered by the authors to be a method of disseminating products consisting in making installation files or a code required to run the application via the Internet available to the player by the publisher. Generally, it is carried out with the application of a platform which is designed especially for this purpose and which provides support in the process of sale and after-sales customer service.

The study considers the data provided by GamesIndustry.biz and a report on the global gaming market in 2017 [1]. According to the published data, the value of the entire gaming market in 2017 amounted to 116 billion dollars, which represents an increase of 10.7% in relation to 2016. Mobile devices whose share in the total value amounted to 42% (an increase by 23% in relation to 2016) were the dominant platform. The income from games running on personal computers and consoles presented a very similar tendency, and it amounted to 27% and 31%, respectively [1-2]. Thus, it is a market which is developing, and thus it appears to be the research field which deserves further analyses. It is worth noting that the share of the digital platforms used for distribution of computer games in the entire market amounts to around 92% [3]. In the literature, one may point to the sources which evaluate the influence of the digital gaming platforms on the players [4] or on the natural environment [5], however, there is no comparative assessment of particular platforms from the point of view of users.

The conducted analysis is designed to determine the most important features of websites and platforms distributing computer games as well as motivations of individuals visiting the websites and making purchases.

II. The assumptions of the research methodology

Four platforms were selected for the above-described comparative analysis: GOG (CD Projekt – Poland, (https://www.gog.com)), Origin - (Electronic Arts Inc.- USA, (https://www.origin.com)), Steam (Valve Corporation – USA, (https://store.steampowered.com)) and Uplay (Ubisoft Entertainment – France, (https://store.ubi.com)). The rationale for choosing these platforms was their greatest recognition in Poland and popularity of use (segment leaders) and the diversity with regard to the location of owners. This has been supplemented and verified by an additional analysis of the frequency of the emergence of individual platforms in the first hundred Google search results and the results of the initial part of the survey distributed among students.

The analysis has been divided into four parts: the frequency of the use of games, devices and platforms, payments; comparative analysis of websites of distributors; comparative analysis of installed applications (platforms) and comments.

The analysis of the selected websites was conducted according to a set of thirty-seven indicators, divided into nine

groups which were verified in previous research [6-7]: website clarity; website support; information quality; the quality of the product search engine; advertising on the site; quality of the ordering process; the method of product distribution; presentation of the product and promotions. The third part of the research was the analysis of final digital distribution applications for PCs with Windows. The applications of the same distributors which were analysed in the first part of the research, namely Valve Corporation (Steam), Electronic Arts Inc. (Origin), Ubisoft Entertainment (Uplay) and CD Projekt (GOG Galaxy), were selected for analysis. The installation of Windows 10 on a virtual machine using VirtualBox 5.0 was used for the analysis. Six groups of parameters were analysed containing in total such criteria as: installation package, the appearance of the application, search engine, transparency, product presentation and security.

Under the circumstances of a dynamic and complex market environment as well as high innovation and competitiveness of solutions, the following questions arise: what characteristics and form should the software of websites and internet platforms adopt to increase their attractiveness to the client? which of the sets of attributes and functionality of these sites seem the most important?

Answers to these questions can be provided by means of a thorough analysis of the requirements of users of online platforms. The quantitative and qualitative research into users' opinions is needed for this purpose, and the study should examine, on the one hand, the use of software and, on the other, the websites which are the sources of its acquisition. Taking into account the fact that there are few and random studies regarding this sphere, both in national literature [8] as well as foreign sources [9-11], the research was based on the authors' own approach consisting of the following steps: selection of the test group, constructing an online survey characterizing websites and distribution platforms from the client's point of view, verification of the survey based on the test group, selection of the most important assessment criteria from the client's point of view, identification of the most important factors influencing the behaviour of the users of websites and distribution platforms and making the revised survey available again on the Internet, along with the dissemination of information about the possibility of its completion, analysis and discussion of the findings, drawing conclusions from research and design recommendations for growing platforms.

The simplified, standardized scoring method of R. Likert [12] was used for the assessment of each specified criterion. On this scale, each criterion was evaluated as follows: 0.00 - the criterion is not implemented, 0.25 - the criterion is implemented on a minimal, sufficient level, 0.50 - the criterion is implemented on the medium level, 0.75 - the criterion implementation level is good, 1.00 - full implementation of the criterion.

The selection of online gaming stores was made based on the analysis of the frequency of the first hundred of Google search results, after entering the keywords: *witryny dystrybutorów gier* (computer games websites) and *platformy dystrybutorów gier* (platforms of computer games distributors). This list has been verified by taking into account the opinions of students who frequently and intensely play computer games via digital platforms. The research was carried out in November 2017, on a sample of 713 randomly selected members of the academic community. 549 people completed the survey (over 77% of respondents), where - after a thorough analysis - 368 survey participants provided complete answers regarding the selected four platforms, which constitutes 67% of the respondents. 33% of the sample uses other platforms than the four websites selected for the present analysis.

The online survey was made available on the servers of the Faculty of Management at the University of Warsaw. The sample included the representatives of the academic community, students of all types of studies at two universities: the University of Warsaw and the Vistula Academy of Finance and Business, who were interested in completing the online survey. The study was carried out in two stages. The first stage concerned the evaluation regarding which of the distribution platforms are the most popular among students, which assessment criteria are the most suitable for the evaluation of distributors' websites and which should be applied to evaluate the applications of the selected Internet platforms. In the second stage, the services and selected applications were analysed according to user-defined criteria for the four previously mentioned online platforms. The survey was completed correctly by 368 respondents, evaluating only those websites they were familiar with out of the four websites considered in the study. Some of them rated two (55 people) or three websites (26 people). This resulted in a total of 446 observations. GOG was rated by 116 people, Origin by 121 respondents, Uplay by 108 and Steam by 103 individuals. Among the survey participants, there were 72.5% of women and 27.45% of men. The average age of respondents was 20.6 years, which was typical for BA and BSc students, mainly with secondary education (over 94%). Over 42% of respondents were non-working students, and almost 58% were working students. Almost 30% of respondents declared coming from the city of over 100,000 residents, over 26% from cities with 11-100 thousand inhabitants, and over 43% from villages or towns up to 10,000 residents.

III. Analysis and discussion of the findings concerning computer game distributors

The questionnaire was divided into three parts: an introductory part comprising eight questions, an analytical part (37 questions about websites and 6 questions regarding the application), a field of opinions, comments and recommendations of users, as well as a data sheet describing the test sample.

A. Introductory information

In order to analyse the obtained data, the scoring method was applied. Each of the respondents assessed individual

criteria in a subjective way. The assessments were then added together, structured and averaged, followed by a reference to the maximum possible assessment of each indicator, both in the cross-section of websites/applications and the criteria for the respondents evaluating them. From the authors' experience, the findings obtained with the application of comparative analyses of websites by scoring method are just as valuable as those received in the case of more sophisticated methods (AHP/ANP, Electre, Promethee and others) [13].

The first question concerned the moment when people using distribution platforms started playing computer games. Nearly 53% of respondents said that they started playing computer games in primary school, more than 12% in preschool age, almost 11% have been playing games from middle school, high school and college, and 24% did not play computer games at all. Among the gamers, almost 47% play games occasionally (once or twice a month), nearly 21% several times a week, 15% of the share - several times a month, more than 10% very rarely - several times a year, and more than 7% play computer games daily. After specifying this question, it turned out that 59% of the sample spend less than an hour a week playing computer games, and 16% play only 1-2 hours a week. Thirty percent of gamers use only a smartphone for their games, over 23% a PC or a desktop computer, 19% use a PC and a notebook, over 20% use a console or a portable console, and 8% - a tablet.

The next question concerned PC and console platforms where students have their accounts. Over 25% have an account on the Origin platform, over 21% on Steam, almost 10% on GOG and Uplay, and the remaining 33% of the share on the Windows Store, Xbox Games Store, Battle.net , PlayStation Store and Nintendo eShop.

Over 53% describe themselves as experienced players (many years of experience and a wide range of games they played), 9% believe that they are advanced players (they play almost every day, different games on different equipment at least from elementary school, 15% claim that they are casual players (novices or people playing only occasionally, e.g. once a month or every three months), and 23% do not play at all.

The students play mainly games that are free of charge - 74%, they do not spend any money on it. Of the remaining 26%, 13% of this share spend up to PLN 20, and only a little over 1% of the respondents spend over PLN 100.

B. Analysis of the findings concerning websites

Analyses of the findings concerning the evaluation of websites of the selected computer game distributors were made in two cross-sections: according to the websites and according to the evaluation criteria. The first cross-section was created by averaging the scores obtained for all specified criteria. The assessment for the detailed criterion was calculated as its percentage share in the potential maximum score that could be obtained during the implementation of a given criterion. All websites included in the ranking have achieved ratings exceeding 50% of the maximum possible score, so it emerges that the clients are generally satisfied with the services they offer. The average assessment in the respondents' opinions is close to 67%. Among the analysed websites, Steam received the highest rating with over 68%, the lowest rating was indicated in the case of GOG Galaxy - almost 66%. Thus the spread of results is in the range of 3%, which is very low value with regard to websites. The leading position in the case of Steam was caused by the highest score obtained for such groups of criteria as: product distribution, website clarity, the quality of product search and ordering process quality - where the average rating of these four criteria was 72%. The lowest rating of the GOG Galaxy website resulted from the highest rating in three categories: on-page ads, information quality and website navigation, with an average of these three criteria equal to 63%. Origin, where product presentation received the highest scores and Uplay (the best promotions) have taken the middle positions. The Steam service was the only website which ranked above the average; however, the difference amounted to less than 2 percentage points. The rating of the remaining websites was slightly below average.

Out of all nine groups of criteria, the highest scores were assigned to the product presentation (76.63%), the quality of the ordering process (72.05%) and the quality of the product search engine (69.67%). This demonstrates the pragmatism of website designers, who first of all pay attention to the most important factors from the point of view of sales that may encourage the clients to visit the website again. The lowest scores were recorded in the case of quality of advertising presented on the website (55.02%), information quality (61.51%) and promotion (64.08%). This is an interesting phenomenon, which shows that there is a growing dissatisfaction (compared to previous surveys of websites) with the excessive and redundant advertising on the website. The discrepancy between ratings is very large in this case, reaching 23 percentage points. The scores which are above average (which amounts to 66.74%) were recorded in the case of four groups of criteria – i.e. the above-mentioned factors listed in the three positions, as well as the criterion described as website navigation (67.15%). More detailed analysis will be present in extended version of the article.

C. Comparative analysis of platforms

Analyses of the evaluation results of the installed applications of the selected computer game distributors were also made in two cross-sections: according to websites and according to the evaluation criteria. The first of them was based on the average calculated for each distribution platform based on detailed assessments. The calculation indicates the absolute domination of the Steam application, expressed by the five highest average ratings in five categories of criteria, out of all six possible. In the opinion of respondents, only in the group of the product presentation (i.e. information about products and tools), the GOG Galaxy application gained the advantage of 3 percentage points. At the same time, it is a group of criteria with the highest average

assessment (74.02%). The lowest score (67.98%) on all platforms was assigned to the installation package (size and ease of access and ways of its distribution). The difference between the maximum and minimum scores amounts to 6 percentage points. Only the Steam platform with its results is above the average of all results. The lowest score was obtained in the case of the GOG Galaxy platform with the result of 69.93% by 4 percentage points less than Steam. The scores of Origin and Uplay platforms were only about one percent higher.

The order of the groups of detailed criteria determined the ranking order. The average for Steam was 73.72%. This was mainly due to the high search engine rating (in terms of usability and mode of operation), the appearance of the application, the layout of the elements, scalability to the screen size, clarity and readability (presentation of notifications and messages and the ease of finding information on the menu). The evaluations of Origin (70.95%) and Uplay (70.55%) are only slightly lower. The presentation method and the quality of the search engine received high scores, the worst scores were assigned to the installation package in each of the selected applications. Apart from this feature, the scores below the average (amounting to 71.23%) were indicated in the case of security – i.e. securing payments and access to purchased products.

For this study, the installation files of applications were downloaded directly from the distributors' websites in the November 2017, bypassing intermediaries. The application designed for the Windows platform was evaluated because all platforms operate on this operating system. Uplay has no alternative to other operating systems, while other platforms also have Mac installations, and Steam also distributes installations on Linux. All installation files were easily accessible from the manufacturer's website.

All platform applications require an active account. Each of them offers the opportunity to work offline, and the ability to add games from outside the platform. The installation process on all platforms proceeded in a similar manner and was not complicated. The most consistent and simplified installation process occurs on the Origin platform. It should also be noted that the installation is usually performed once.

IV. Analysis of the findings with the application of a scoring method with preferences and the discussion of the findings

One of the methods limiting the subjectivity of the experts' evaluations of users in the case of a scoring method (apart from the prior averaging of scores) is the application of unitary preferences, to particular criteria or selected criteria sets. In this study, the authors divided the criteria into three groups which are important or particular categories of users: novice – this category of a user is characterised by the interest in what he or she may evaluate at first sight, as well as the ease of obtaining a product and making a payment; gamer – a person who perceives a game as entertainment, frequently first plays games which are available free of charge on their smartphone, the switches to PC or console games; professional – a person who plays very frequently (every day), is passionate about playing games, plays all the latest games and is ready to pay for using them, plays professionally and can even earn money on the activity, etc. (he or she is mainly interested in the functional aspects: the presentation of the product, the quality of information (the amount of information on a website, the possibility and manner in which one may ask questions, ease of access to information), the quality of the product search engine (the number of modifications, the number of filters, accuracy of answers, clarity of the scores). For each group, the authors adopted one variant with a group of dominating criteria: novice (70% for technological aspects, 15% for the remaining ones); gamer (70% for service-related criteria, 15% for the remaining ones); professional (70% for functional criteria, 15% for the remaining ones).

Assigning preferences to particular groups of criteria resulted in slight changes in the rankings. The greatest changes could be observed in the case of the novice category in the case of games where Uplay platform moved from the last place to the second position in the ranking. In the remaining cases, the authors only recorded the reduction of the distance in relation to the previous experiment. Small differences confirm that despite the significant differences with regard to the strategy of the development of the examined platforms, there emerges a specific standardization of the product/services ranges offered to clients. The summary of the positive features of the ranking points to the dominating position of the Steam platform. Only in the variant of an e-gamer, the scores obtained for the Steam and Origin platforms were above the average, and in the case of a gamer only Steam platform reaches the scores beyond average, in the case of the novice, such scores were obtained for Steam and Uplay platforms.

V. Conclusions and recommendations

The conducted survey studies, supplemented with the opinions and comments of clients of computer games shops (the original wording of the respondents' opinions was retained), lead to the following conclusions: the majority of users are satisfied with the appearance and functioning of the websites distributing computer games, which is evidenced (mostly) by high scores (above 50%) of the specified criteria and their average values, both in the case of the selection of the examined companies as well as the evaluation criteria sets; the respondents emphasise the fact that in the case of the analysed websites they pay particular attention to the website features which allow them to easily obtain information on the content of the game (product information – 83% on average, and the information clarity – over 78%), as well as to find out whether they will be able to use all the functionalities of the game (hardware requirements - over 78%), ease of registration and payment (registration method and payment options – over 72% of approval), the lowest scores were obtained in the case of the excessive and intrusive advertising on the website (over 43%), despite the fact

that these are mainly advertisements aimed at self-promotion; Due to the fact that most of the individuals using the platforms consider themselves to be gamers, or even advanced gamers (65% in total), they are not interested in the manner and possibility to ask questions (51% and 56% respectively), as well as restrictions for unregistered users. Majority of the platforms is not designed for novices; clients value the simplicity and clarity of the analysed websites and the scores as well as the ease and intuitiveness of navigation, including the product search, the website selling computer games needs to be easy and clear to use, and a client cannot have any problem finding what he or she needs; the respondents pay attention to an important role of visualisation in attracting customers to computer games shops, simultaneously being aware of the fact that an excessive number of graphic elements may disturb visitors in making a selection and purchase,

At present, Steam is a world leader in the category of computer games distribution, and its high score in this study confirms its position and the awareness of the market where it operates. The lower scores indicated in the case of other platforms may be explained with the fact that it is targeting a rather narrow group of recipients, who do not mind certain shortcomings with regard to the website since they are satisfied with the high quality of the final product they purchase. Origin and GOG have a modern look and very good tools to communicate with the user. The design of the Ubisoft website appears to be obsolete, and thus it may be seen as unattractive. Moreover, another disadvantage of Ubisoft is the lack of consistency with regard to the naming of the website and the application which is installed on a PC to use the purchased games.

One of the limitations of the study was undoubtedly focusing on only four distribution platforms. These are the leading platforms on the market, and in total they constitute – according to statistics - approximately 70-80% of the market share.

The diversity of the opinions concerning the computer games websites causes some difficulties with regard to generalizing the evaluations. In the case of platforms offering games from independent designers there emerge various phenomena (such as “fake games”) [14], which may negatively affect the rating of a particular platform. Despite the high popularity of this type of websites, the similarity of their evaluations to the evaluations of internet shops in other areas of business [15]. Thus, it may be stated that there occurs a specific standardisation of views on how the website should look like. On the other hand, it gives also the idea on the discrepancy between the users’ expectations and the projects of their creators. Moreover, one may conclude that the traditional principles of designing websites are still up-to-date and applicable.

References

[1] GamesIndustry.biz. (2017) "GamesIndustry.biz presents... The Year In Numbers 2017", https://www.gamesindustry.biz/articles/2017-12-20-gamesindustry-biz-presents-the-year-in-numbers-2017,

[2] SuperData Research. (2018) "Market Brief — 2017 Digital Games & Interactive Media Year in Review", https://www.superdataresearch.com/market-data/market-brief-year-in-review/

[3] Lifewire. (2017) "Top PC Game Digital Download Services", https://www.lifewire.com/top-pc-game-digital-download-services-813065

[4] Polygon. (2014) "In the long run, do Steam sales harm gamers?", https://www.polygon.com/2014/1/15/5313142/in-the-long-run-do-steam-sales-harm-gamers

[5] Buonocore, Cathryn E. (2016) "Comparative Life Cycle Impact Assessment of Digital and Physical Distribution of Video Games in the United States", http://nrs.harvard.edu/urn-3:HUL.InstRepos:33797406

[6] Chmielarz W., Szumski O.: Analysis of users of computer games, Volume 8, Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, September 11-14, 2016, Gdańsk, Poland eds. M. Ganzha, L. Maciaszek, M. Paprzycki, PTI Warsaw, IEEE New York City, str. 1139-1146, ISSN 2300-5963,

[7] Chmielarz W., Szumski O.: Analysis of Predispositions of E-gamers and Its Relevance in the Use of Computer Games Didactic Process, in: Information Technology in Management, New Ideas and Real Solutions, 11-th Conference of ISM and 14-th Conference of AITM part of FedCSIS, 2016, Editor: E. Ziemba, Book, Lecture Notes in Business Information Processing, Gdańsk, Poland, ISBN 978-3-319-53075-8,, vol. 277, 2017, pp. 77-102,

[8] Ziemba E. (red.), Towards a Sustainable Information Society. People, Business and Public Administration Perspectives, Cambridge Scholars Publishing, Newcastle upon Tyne, 2016.

[9] Nielsen J., Projektowanie funkcjonalnych serwisów internetowych, Helion, Gliwice, 2003.

[10] Nielsen J., Mobile Website and Application Usability, Nielsen Norman Group Press 2013.

[11] Buonocore, C. E.: Comparative Life Cycle Impact Assessment of Digital and Physical Distribution of Video Games in the United States. Master's thesis, Harvard Extension School, 2016, https://dash.harvard.edu/handle/1/33797406?show=full

[12] Likert R., A Technique for the Measurement of Attitudes, [in:] Archives of Psychology, Nr 140, 1932, str. 1–55.

[13] Chmielarz W., Szumski O., Zborowski M., Kompleksowe metody ewaluacji witryn internetowych, Wydawnictwo Wydziału Zarządzania UW, Warszawa, 2011.

[14] USgamer. (2017) "Valve Removes Nearly 200 "Fake" Games from Steam", https://www.usgamer.net/articles/valve-removes-nearly-200-fake-games-from-steam

[15] Chmielarz W., Determinanty rozwoju serwisów dystrybucji treści komercyjnych w Polsce, [in:] Problemy Zarządzania, Wydawnictwo Naukowe Wydziału Zarządzania UW, Problemy wykorzystania systemów informatycznych zarządzania w gospodarce, vol. 13, nr 2 (52), t.1., 2015, str. 51-65, ISSN 1644-9584.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 697–700
DOI: 10.15439/2018F35
ISSN 2300-5963 ACSIS, Vol. 15

Cloud Platform Real-time Measurement and Verification Procedure for Energy Efficiency of Washing Machines

Pedram Memari
School of Industrial Engineering
University of Tehran, Tehran, Iran
memari.pedram@ut.ac.ir

Seyedeh Samira Mohammadi
School of Computer Engineering
Islamic Azad University South Tehran Branch, Tehran, Iran
st_sa.mohammadi@azad.ac.ir

Seyed Farid Ghaderi
School of Industrial Engineering
University of Tehran, Tehran, Iran
ghaderi@ut.ac.ir

***Abstract*— Industrial administrators are promoting approaches to improve energy efficiency and developing smart homes and appliances. Development of green technology requires accurate models. Real-time Measurement and Verification (M&V) procedure is used to quantify energy performance. It is conducted through short-term on-site measurements and engineering calculation. The period of this procedure lasts for several months or up to a year so the failure to immediately detect abnormal energy efficiency decreases energy performance so timely correction of appliances will be unable and the opportunity to adjust or repair them will be missed. In this study, a cloud computing platform is established to measure the washing machine energy consumption parameters and calculate energy savings which consist of load sensors and fuzzy control. Time-series data are transmitted to the cloud environment through the network and saved in databases. On this platform, for constructing accurate models, integration of the particle swarm optimization (PSO), M&V methodologies and multivariate regression analysis are used. After uploading energy consumption data directly, pre-installation energy baseline model is created and post-installation real-time energy performance calculation is obtained. Observing fluctuations of washing machine energy consumptions provides real-time monitoring or correction of the operating performance of the appliance or system and then good energy performance can be obtained. The aim of this study is to gain real-time and long-term energy performance information and automatic calculations of energy savings on washing machines. Using this cloud platform for home appliances could help the manufacturers to promote energy efficiency programs on smart appliances.**

I. Introduction

Energy consumed in buildings consists of residential and commercial end-users and accounts for 20.1% of the total delivered energy consumed worldwide. Energy performance is evaluated through measurement and verification (M&V) procedures under the Tradable White Certificate, Clean Development Mechanism, Demand-Side Management, Energy Service Companies (ESCO) and Energy-Saving Performance Contracting [1]. M&V procedures have become an important key in the energy efficiency policies. Related works were conducted. Bertoldi, et al. [2] standardized contracts and M&V procedures based on the results of reviews and analysis of ESCO industries in the Europe and help end-users understand the M&V procedures. International performance measurement and verification protocol has been used by many countries. This protocol describes M&V concepts and methodologies to determine energy savings for energy conservation measures (ECM) s in residential buildings and industrial processes. For M&V of energy savings in building energy management projects, the American Society of heating, refrigerating an air-conditioning Engineers Guideline 14 is used [3]. This guide contains M&V concepts and methodologies to calculate energy and water savings in residential, commercial and industrial buildings. Different methods are applied to various energy conservation measures to calculate the baseline model and energy saving. Many M&V studies have been conducted. Lee [4] proposed an accurate model for energy savings calculation by long-term monitoring and assumed this model for individual cases because actual lighting conditions may differ from data, which provided by clients, and this difference makes errors in energy saving calculations. Dong, et al. [5] presented a baseline model for energy consumption in buildings by using regression analysis and considered different parameters for the baseline model which included outdoor dry-bulb temperature, relative humidity, and global solar radiation. They used statistical indicators such as coefficient of determination (R2) and the coefficient of variation of the real-time-square error (CVRMSE) to verify the accuracy of the baseline model. Related works show that regression analysis methods are used for constructing M&V baseline models to evaluate the performance of energy conservation measures accurately. PSO algorithm creates more accurate models than the least error squares techniques.

In this study, a cloud computing platform for real-time measurement and verification of energy saving performance is created based on the M&V methodology and integrating of PSO algorithm, java programming language, and cloud computing techniques. While baseline models are applied automatically for pre-installation and real-time energy saving performance is calculated automatically after post-installation, the cloud computing platform reduces the time

and cost of M&V and increases the accuracy of energy saving calculations.

A. Measurement and Verification Procedure

Non-profit Efficiency Valuation Organization (EVO) proposed the International Performance Measurement and Verification Protocol (IPMVP) which explains how to determine baseline data and calculate energy savings. Baseline data should be established because energy consumption in the pre-installation period (i.e., baseline period) cannot be measured by instruments after energy conservation measures (ECM) have been implemented [6]. Therefore, adjustments (A) to the baseline data are required to determine the energy consumption of the equipment and systems during the baseline period (BP) and the energy consumption during the post-installation period (reporting period (RP)) under the same operating conditions [7] The amount of savings (S) can be calculated as shown in (1).

$$S = (BP - RP) \pm A \tag{1}$$

There are four options, A through D, for M&V procedure. In option A, only key parameter will be measured. But in option B all parameters will be measured. A and B options can be used when evaluating the performance of a single conservation measure. Energy consumption parameters will be measured for pre-installation and post-installation. Statistical methods or engineering calculations are used for evaluating the energy performance [8]. In option C, energy conservation measure is calculated by analyzing energy bills and statistical methods. In option D, first energy savings are calculated by specialized software and energy consumption for pre-installation and post-installation in simulated. Then the energy bill data are used to correct the simulation models. C and D options are used for evaluating energy conservation measures in whole buildings or in situations where measurements are difficult.

B. Particle swarm Optimization

PSO is a probability-based optimization technique that was developed by Eberhart and Kennedy [9] and inspired by the social behavior of birds or fish in finding food. It is assumed that a group of birds are randomly searched for food in a region, while food is only available in one part of the search area [10]. In PSO, each answer is a bird in the search area called the particle. Each particle has a fitness value obtained by the objective function. The bird that is closer to food is more desirable [11]. Steps of PSO algorithm and Equation of it are as follow:

$$V_i^{k+1} = W \times V_i^k + L_1 \times r_1\left(Pbest_i^k - X_i^k\right) + L_2 \times r_2\left(Gbest_i^k - X_i^k\right) \tag{2}$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \tag{3}$$

(1) First, a position and velocity is randomly considered for each particle in N-dimensional space.

(2) Next, an objective function is used to determine the fitness value for each particle.

(3) The fitness value is compared to the current best value for each particle (Pbest) and then the particle velocity is adjusted to improve the fitness value in the next iteration.

(4) A comparison between Pbest and the current best value for group (Gbest) is conducted, if Pbest is greater than the Gbest, then Gbest will be adjusted. The velocity and position of each particle are then adjusted based on Gbest for the next iteration.

(5) Each particle's velocity and position are updated using (2) and (3) as shown in Table II.

II. Cloud Computing Platform for M&V of Real-Time Energy Performance

The architecture of cloud computing platform contains three layers: first is the on-site measurement layer, second is the cloud computing layer and the last one is the user layer.

(1) *On-site measurement layer:* Energy consumption and other energy parameters are stored in storage devices and then all collected data are transmitted to the cloud computing platform via the internet.

(2) *Cloud computing layer:* All calculations are conducted in this layer, including creating files, receiving data, transmitting data, filtering data, sorting data, executing PSO algorithm operations, constructing baseline equations and energy savings calculations. In this cloud computing platform, the Google App Engine (GAE) is used as the cloud-based server. For displaying the results on graphs and curves and calculating energy savings automatically, web pages are created dynamically in hyper markup language (HTML) or JavaScript.

(3) *User layer:* In this layer, users can communicate with the cloud platform via their computers and follow the energy consumption fluctuations of any appliance at any time interval and on the curves, and also finds abnormal patterns, so they become aware of the critical issues of the appliance.

III. Software Architecture

Software architecture of this cloud platform contains PSORun (server), PSORun (client) and the Data Upload Program which are described as follow:

- *PSORun (server):* this program includes four subprograms and established on the GAE platform.
  1) *ProjectServiceImpl.java:* the functions of this program include project creation, project storage, project deletion and PSO operations. Which conducted by this primary program.
  2) *File Upload.java:* when pre-installation baseline data are uploaded and cloud computing results are returned, this program is called and executed.
  3) *DoGet and DoPost are hypertext transfer protocol requests:* DoGet protocol retrieves and

transmit data from the server to users and DoPost transmit data from the user to the server.

- *PSORun (client):* web page creation

Projectlist.java: this program is an interface which can be divided into two subprograms:

1) projectList-Savingview.java: is the interface for energy savings calculations
2) ProjectList-Manage.java: is the interface for managing projects.

- *Data Upload Program*: this program uploads post-installation data in a fixed format to the cloud platform for calculating energy savings.

IV. Constructing the Baseline Model Using PSO

Users can determine the number of independent variables in baseline regression model using cloud computing platform. In this study four energy consumption parameters (i.e. four independent variables) are used in the baseline regression model. The pre-installation energy consumption is assumed as P and the measurement of the four independent variables are considered as X, Y, V and Z. the PSO algorithm is applied to determine the coefficient of each independent variable (C_0- C_{14}) [1].

$$P = f(X, Y, V, Z) \tag{4}$$

$$P = C_0 + C_1 \times X + C_2 \times X^2 + C_3 \times Y + C_4 \times Y^2 + C_5 \times V + C_6 \times V^2 + C_7 \times Z + C_8 \times Z^2 + C_9 \times X \times Y + C_{10} \times Y \times V + C_{11} \times V \times Z + C_{12} \times X \times Z + C_{13} \times X \times V + C_{14} \times Y \times Z \tag{5}$$

In this process, the cloud computing platform first filters the pre-installation data using a regression analysis while retains data with less than 10% errors for dependent variables. Then, PSO regression analysis is used to determine the coefficients of independent variables to complete the baseline model. Further, the filtered independent variable data can be used as the upper and lower limits for post-installation data filtering, therefore users can be assured that the ranges of the baseline model and uploaded data are matched and the energy savings are calculated accurately. PSO algorithm includes the number of particles, number of iterations, inertia weight (w), cognitive learning factor (L_1), social learning factor (L_2), maximum particle velocity (Vmax), minimum particle velocity (Vmin), maximum particle position (Xmax), minimum particle position (Xmin), random numbers (r_1 and r_2) and convergence condition (Z_minAvg) which can be defined by the user. When the PSO algorithm determines that the convergence condition has been satisfied, the calculations will be determined. The convergence condition in this study is the relative error of dependent variable in (6).

$$Z = \sum_1^n \frac{depend\ variable - depend\ variable\ pso}{depend\ variable} \times 100 \tag{6}$$

V. Energy Saving Calculations

Energy saving calculations require uploading post-installation data to the cloud platform and a corresponding pre-installation project ID and URL should be entered like a data file name and file path. Inputting post-installation energy consumption values to the pre-installation baseline model yields the pre-installation energy consumption with the same conditions. Energy performance is obtained by subtracting the post-installation value from the pre-installation value (1).

While the post-installation data are uploaded to the cloud platform, a data filtering function is used to allow the appropriate data range to be set, so energy savings will be calculated accurately. In this cloud computing platform, energy performance information including average energy savings (KW), total kilowatt-hours savings (KWh), average percentage of energy savings, and energy cost saving are calculated.

In addition, constructing a time interval calculating function allow users to set a time interval so they can monitor fluctuations in the pre-installation energy consumption (Padjusted) and post-installation energy consumption (Pmeasured) on a graph and users can select a time point over the curve to see the associated data record and gain the results online.

VI. Case Study

The M&V cloud computing platform is used to assess the energy conservation for the washing machine. This platform automatically calculates the energy consumption, energy savings, and the energy cost savings. In addition, this system displays the results on the web pages and allows users to monitor the energy savings and energy performance and whenever the efficiency of the washing machine comes down, it warns the users, for example, washing their laundries in proper time or change the filter of machine. According to (4) and (5) five parameters, wash load (L_w), the amount of water (W_q), the temperature to be reached (T_f), energy consumption (P_e) and spin speed (V_s), are selected and shown in (7). First, the project ID, project name, and PSO parameters Table II were entered into the real-time energy performance M&V cloud computing platform. Next, pre-installation energy consumption data are uploaded for calculations. Cognitive Learning Factor is a broad theory that explains thinking and differing mental processes and how they are influenced by internal and external factors in order to produce learning in individuals. Social Learning Factor (Bandura) Bandura's Social Learning Theory posits that people learn from one another, via observation, imitation, and modeling. The theory has often been called a bridge between behaviorist and cognitive learning theories because it encompasses attention, memory, and motivation. Then the PSO algorithm is applied to determine the coefficients of these independent variables (C_0-C_{14}) and the construction of the baseline model is

complete and the message ‘‘true” was displayed for the field ‘‘computations completed”.

$$P_e = C_0 + C_1 \times L_w + C_2 \times L_w^2 + C_3 \times W_q + C_4 \times W_q^2 + C_5 \times T_f + C_6 \times T_f^2 + C_7 \times V_s + C_8 \times V_s^2 + C_9 \times L_w \times W_q + C_{10} \times W_q \times T_f + C_{11} \times T_f \times V_s + C_{12} \times L_w \times V_s + C_{13} \times L_w \times T_f + C_{14} \times W_q \times V_s \quad (7)$$

After the post-installation data for energy consumption and other energy parameters are uploaded to the cloud computing platform for the real-time M&V of the energy performance, the calculations indicates the energy savings accumulated over 11.5 h. Specifically, the average energy savings is 20.3 KW, the total kilowatt-hours saved is 2.8 KWh, the average percentage of energy savings is 25%, and energy cost savings was $523 as shown in Table I.

TABLE I.
THE POST-INSTALLATION ENERGY SAVINGS FOR THE WASHING MACHINE

Average energy savings(KW)	20.3KW
Total kilowatt-hours savings	2.8 KWh
Average percentage of energy savings	25%
Energy cost saving	523$

TABLE II.
PSO PARAMETERS VALUE

Interia weight (W)	0.4
Social learning factor (L2)	2
Minimum particle velocity (Vmin)	-5
Minimum particle position (Xmin)	-20
Maximum number of iterations	10,000
Cognitive learning factor(L1)	2
Maximum particle velocity (Vmax)	5
Maximum particle position (Xmax)	20
Number of particles	300
Convergence condition (Z_minAVG)	0.1

VII. CONCLUSION

Over the past few years, the energy efficiency and the market penetration of efficient washing machines has increased. Key parameters of washing machines are their energy and water consumption (full and partial loads), spin-drying efficiency and supply for hot fill. There are many washing machine features which make sense concerning energy efficiency containing load-auto sensor, automatic temperature control, automatic dispensers, spin speeds, fuzzy control and all water washing machines. The key parameters of washing machines can be achieved by these features. Observing fluctuations of the washing machine energy consumptions on a curve can help the users monitor the appliance operations and gain information about the energy consumption, a view of calculating energy savings and detect abnormal functioning. In this study a washing machine is considered with four independent variables including wash load, the amount of water, temperature to be reached, speed of spin and energy consumption for heating, mechanical actions, and pumping. A cloud computing platform for the M&V of the real-time energy performance that integrates the M&V methodology, PSO, and multivariate regression analysis modeling is conducted on the GAE cloud platform. The M&V cloud platform use uploaded real-time energy consumption data and create pre-installation baseline model and post-installation real-time energy performance calculation. When independent variables are uploaded onto the platform, a baseline regression model can be established and the energy performance can be calculated then the results will be displayed on the web page. Therefore, whenever energy consumption rises or energy savings decrease, the system alerts and suggests that the relevant problem be corrected for example the system suggests that the filter should be changed or declares the time inappropriate. This procedure can be used for evaluating energy performance accurately.

REFERENCES

[1] M.-T. Ke, C.-H. Yeh, and C.-J. Su, "Cloud computing platform for real-time measurement and verification of energy performance," *Applied Energy,* vol. 188, pp. 497-507, 2017.

[2] P. Bertoldi, S. Rezessy, and E. Vine, "Energy service companies in European countries: Current status and a strategy to foster their development," *Energy Policy,* vol. 34, no. 14, pp. 1818-1832, 2006.

[3] J. S. Haberl, D. Claridge, and C. Culp, "ASHRAE's guideline 14-2002 for measurement of energy and demand savings: How to determine what was really saved by the retrofit," 2005.

[4] A. H. Lee, "Verification of electrical energy savings for lighting retrofits using short-and long-term monitoring," *Energy conversion and management,* vol. 41, no. 18, pp. 1999-2008, 2000.

[5] B. Dong, S. E. Lee, and M. H. Sapar, "A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore," *Energy and buildings,* vol. 37, no. 2, pp. 167-174, 2005.

[6] S. Park, V. Norrefeldt, S. Stratbuecker, G. Grün, and Y. S. Jang, "Methodological approach for calibration of building energy performance simulation models applied to a common “measurement and verification” process," *Bauphysik,* vol. 35, no. 4, pp. 235-241, 2013.

[7] D. Jump, M. Denny, and R. Abesamis, "Tracking the benefits of retro-commissioning: M&V results from two buildings," in *Proceedings of the 2007 National Conference on Building Commissioning*, 2007, pp. 2-4.

[8] T. Giglio, R. Lamberts, M. Barbosa, and M. Urbano, "A procedure for analysing energy savings in multiple small solar water heaters installed in low-income housing in Brazil," *Energy Policy,* vol. 72, pp. 43-55, 2014.

[9] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, 1995, pp. 39-43: IEEE.

[10] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information processing letters,* vol. 85, no. 6, pp. 317-325, 2003.

[11] A. Ghodousian and M. R. Parvari, "A modified PSO algorithm for linear optimization problem subject to the generalized fuzzy relational inequalities with fuzzy constraints (FRI-FC)," *Information Sciences,* vol. 418, pp. 317-345, 2017.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 701–705
DOI: 10.15439/2018F125
ISSN 2300-5963 ACSIS, Vol. 15

Comparative Analysis of Big Data Analytics and BI Projects

Gloria J. Miller
SKEMA Business School
Lille, France
Email: gloriajean.gloria@skema.edu

Abstract—Decision support systems such as big data, business intelligence (BI), and analytics offer firms capabilities to generate new revenue sources, increase productivity and outputs, and improve competitiveness. However, the field is crowded with terminology that makes it difficult to establish reasonable project scopes and to staff and manage projects. This study clarifies the terminology around the data science, computation social science, big data, business intelligence, and analytics and describes their meaning relative to decision support projects. For BI and big data projects, it identifies the critical success factors, empirically classifies the project scopes, and investigates the similarities and differences between the project types. This comparative analysis provides unique insights into the factors and criteria that influence BI and big data project success. These results should inform project sponsors and project managers of the contingency factors to consider when preparing project charters and plans.

Index terms: Big data analytics, data science, business intelligence (BI), project success factors

I. Introduction

DATA Science and computational social science are emerging interdisciplinary fields that overlap in content with big data, business intelligence (BI), and analytics. As more data have become available on the internet, social media, and other source organizations have begun to collect it in growing volumes, new business models and algorithms are emerging, and data sales have become potential revenue sources [3]. Despite the increased attention to big data, the critical success factors for decision support projects have received little attention in the project management literature. Decision support projects are implementation projects that deliver data, analytical models, analytical competence, or all three, for unstructured decision-making and problem-solving. They include subspecialties such as big data, advanced analytics, business intelligence, or artificial intelligence. Without insights into the project's critical success factors, it can be challenging for project sponsors and managers to establish a reasonable project strategy and to achieve the desired benefits efficiently.

The emerging status of these fields means that terminology is not standardized. Scant research exists about the application and scientific and commercial implications of these domains to project management. There is little understanding of the impacts of those differences in transitioning organizations in order to benefit from those scientific areas and data platforms. This paper empirically investigates the critical success factors for decision support projects and provides a comparative analysis of big data and BI projects. This comparative analysis provides unique and previously unpublished results on the structural factors that contribute to decision support project success. The findings of this study add to technology and project management practices. In particular, it provides in-depth insights into what factors influence big data and BI project success. It provides information on the similarities and differences with regard to the criteria for measuring success. These results should inform project sponsors and project managers of the contingency factors to consider when preparing project charters and plans.

II. Literature Review

A. Business Intelligence

Business intelligence is used to refer to technology, processes, and software used to transform raw data into intelligence for computer-aided decision-making. The BI process includes the collection, evaluation, analysis, and storage of data and the production and dissemination of intelligence [6]. Davenport and Harris [7, p. 7] define analytics as a sub-category of BI that includes "*the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.*" Otherwise, the term analytics is not standardized in literature.

B. Big Data

As data volumes have grown and computing requirements have increased, technologies and tools used to manage, manipulate, and understand data have evolved [8, 9]. On the one hand, there is a consensus that "Big Data" refers to the attributes of data——velocity, variety, validity, and volume—— and on the other hand, it refers to innovative technologies and processes that allow for the use of data in novel ways [3].

C. Data Science

Data science is an interdisciplinary field that includes data analysis, statistics, data mining, and models; it has the goal to transform data into knowledge by finding patterns and trends in the data [8, 10]. The terminology originated as a role description for a single person who could act as a Business

This work forms part of a research project on "Decision Support Project: Project Success and Organizational Performance" conducted by the author.

Analyst, Statistician, Engineer, or Research Scientist [8]. Davenport and Patil [10] popularized the data scientist's role in their "Harvard Business Review" article; they described it as "a high-ranking professional with the training and curiosity to make discoveries in the world of big data" [10, p. 72].

D. Computational Social Science

Computational social science refers to revealing patterns related to group and individual behavior. It is based on the emergence of scientific research methods that can leverage big datasets [11]. Computational intelligence refers to approaches where algorithms intimate human information processes and reasoning [12].

E. Project Success Factors

Projects are a form of temporary organization for introducing changes and transitions into organizations. The transition is the change or transformation expected as a consequence of the temporary organization's tasks. Project success refers to the project delivering its expected output and achieving its intended objective [13, 14]. Success criteria and success factors are the main elements of success. Success criteria are used to judge the outcome of the project, and success factors influence the likelihood of achieving a successful outcome. That is, the success criteria determine measures or indicators for success, while the success factors refer to the circumstances, conditions, and events that support the project in achieving its objectives [13, 14].

F. Summary

There is significant overlap in the implementation processes and technologies used between big data, analytics, BI, decision science, and computational social science. We can summarize the terminology as follows for further use in this paper. BI, big data, and computational social science have similar technical processes and techniques but differ in use cases. BI focuses on the platforms, architectures, and tools for the provision of data and intelligence and has an enterprise orientation. Big data seeks use cases to monetize data directly through its sale or indirectly through data-driven business models or algorithms. Computational social science has a scientific research focus. Its use case focuses on leveraging big datasets for learning. Data science refers to the responsibilities of the people who use BI, big data, or analytical techniques; it describes a role or job.

III. Research Methodology

The research used a web-based survey with quantitative methods to collect data on decision support projects and to explore the difference between BI and big data analytics' project characteristics. The measurement items and composite variables are based on a review of the literature and quantitative analysis conducted by [15]. The survey collected the data over a ten-week period (September 2017 to December 2017) from a single informant. Project managers, team members, and sponsors from completed decision support projects were asked to take the survey. The responses were checked for scope, completeness, consistency, ambiguity, missing data, extreme responses, outliers, and leverage. Validity checks for common method bias, response bias, and reliability were conducted. No bias was found, and the data were considered to be reliable and valid.

The survey sample was comprised of 82 usable responses as follows: 76% of the respondents have a master's degree or higher; 38% perform IT roles; 18% are from a project management office; 48% are project or program managers; 5% agile coaches; 22% project team members; and 5% are project sponsors. The organizations sponsoring the projects were mostly publicly traded (51%) and were large, with more than 249 employees (83%) and US$50 million in revenue (78%). They are spread throughout 22 different industries and 24 countries. The majority of the participants were from Europe (74%).

Latent Class Analysis (LCA) was used to classify the types of projects. Descriptive statistics, mean ranking, Wilcoxon score, and correlation analysis were used to explore the characteristics, establish the validity and reliability, and explain the relationship between the variables.

IV. Construct Operationalize

Project critical success areas based on [1] and decision support success systems factors were used to formulate the measurement instrument. Table I includes an abbreviated description of the measurement items. Most of the items are based on a five-point Likert scale (1 = Not at all; 5 = To a great extent). To ensure completeness of data, "Don't know" or "Not Applicable" (N/A) was added to the Likert scale. Items that used different scales are described in the relevant sections.

A. Project Mission

The project mission represents the clarity of goals and directions [1]. The following deliverables proposed by [7] as valuable components of analytics projects were used in the analysis to define the project strategy and classify the project types: *Proprietary algorithms* or business models, *new data* that was not previously available in the company, deliverables embedded into *distinctive business processes*, and data science or *analytic competence*. The measurement instrument did not directly ask a question on the project's objectives. Thus, the organizational impacts of the project were used as proxies for defining the business strategy and vision. The impacts include the *cost* and effort-saving or increases in productivity, increases in *revenue,* and *strategic benefits* such as providing new and reusable learning or improving forecast and prediction accuracy.

B. Top Management Support

Top management support is needed to authorize the project and to provide the resources and authority to execute the project [1]. User contribution is divided into user participation and user involvement [17]. *Participation* represents an active role in the development process and *involvement* represents the importance and personal relevance an individual places on the system or project. Top management support was measured based on top management and senior management involvement.

TABLE I.
MEAN RANKING AND KRUSKAL-WALLIS TEST OF CRITICAL SUCCESS FACTORS (N = 82)

Critical Success Factor			**Big Data Analytics**		**Bus Intelligence**		**Kruskal Wallis**	
Project [1]	**BI & big data**	**Measurement Item**	**Mean**	**Rank**	**Mean**	**Rank**	**H**	**p-**
Project Mission	BI/BD Strategy [2], [4], [5]	New Data	3.38	29	3.30	21	0.19	0.67
		Distinctive Business Processes	4.1	8.5	3.25	22	15.2	0
		Proprietary Algorithms	3.52	26.5	2.90	25	5.1	0.02
		Analytic Competence	3.67	22.5	2.67	29	11.45	0
	Business Strategy / Vision [2], [4], [5], [9]	Cost Performance	3.46	28	3.56	11	0.02	0.90
		Revenue Performance	3.67	22.5	3.62	9.5	0.07	0.80
		Strategic Benefits	3.89	17	3.44	14	2.48	0.12
Top Mgt Support	Top Mgt Support [5], [9]	Sr Mgr Involvement	4.05	11	3.35	16.5	2.86	0.09
		Top Mgt Involvement	3.99	14	3.31	19.5	3.32	0.07
Project Schedule/Plans	Approach [4], [9], [16],[5],	Technological Uncertainty	2.14	39	2.28	32	0.36	0.55
		Pace	2.48	34	2.10	35	1.75	0.19
		Complexity	2.33	35	1.92	38	4.45	0.03
		Product Novelty	2.52	32	1.69	39	12.15	0
Client Consultation	Mgt Participation [7]	Sr Mgt Prj Directing	3.60	24	2.78	28	5.27	0.02
		Top Mgt Prj Steering	2.57	31	2.06	36	2.57	0.11
	User Participation [2]	Business User Participation	3.94	16	3.08	24	7	0.01
		Business User Prj Acceptance	2.50	33	2.23	33	0.67	0.41
Personnel	Specialized Skills [7], [10], [2], [16],[5], [19]	Analytical Competence	3.71	20.5	2.84	26	4.77	0.03
		Business Competence	4.05	11	3.62	9.5	1.69	0.19
		Data Competence	3.95	15	3.31	19.5	4.85	0.03
		Data Scientist in Team	4.10	8.5	2.79	27	10.65	0
		Technical Competence	4.19	3.5	3.77	2	6.13	0.01
Client Acceptance	Mgt Participation [7], [5]	Sr Mgt Fact-Based Decision	2.24	36.5	2.17	34	0.01	0.93
		Top Mgt Data Driven Action	2.14	38	1.99	37	0.06	0.81
	User [17], [21]	System Use	3.81	18	3.86	1	0.18	0.67
Technical Tasks	Analytic Tools [7], [20], [5]	Analytical Sophistication	3.33	30	3.33	18	0.07	0.79
	Data Architecture [2], [24], [10]	Data Availability	4.14	5.5	3.63	8	4.83	0.03
		Data Privacy	3.81	19	3.20	23	4.7	0.03
		Data Quality	4.21	2	3.66	5	5.72	0.02
		System Security	4.00	13	3.64	6.5	2	0.16
	Data [2], [24]	Data Velocity	2.24	36.5	2.44	31	0.12	0.72
		Data Volume	4.38	1	3.64	6.5	7	0.01
		Data Variety	3.52	26.5	2.64	30	6.12	0.01
	IT Infrastructure [7], [24]	Ease of Operations	3.54	25	3.37	15	0.45	0.50
		Performance Quality	4.19	3.5	3.73	3	4.92	0.03
	Service Quality [5], [22]	Business Support Quality	4.05	11	3.50	13	4.67	0.03
		Personal Qualities	3.71	20.5	3.35	16.5	2.17	0.14
		Technical Service Quality	4.13	7	3.72	4	3.91	0.05
Monitoring & Feedback	Prj Mgt [2], [4], [19]	Prj Mgt Competence	4.14	5.5	3.54	12	2.65	0.10

Abbreviations: BI-Business Intelligence, Bus-Business, BD-Big data analytics, Prj-Project, Mgr-Manager, Mgt-Management, Sr-Senior

C. Project Schedule/Plan

The project plan represents the steps needed to reach the project goal [1]. The Shenhar and Dvir [18] project classification model was used to define the attributes of the project. The three levels of *Technological uncertainty* refer to the degree to which the company is using a technology it has never used before: no, some, almost all, or all new technology. *Complexity* is described on three levels: assembly, system, and array. *Pace* describes the sense of urgency and the four levels are: regular, facts/competitive, time-critical, and blitz. Product *novelty* represents the uncertainty of the project goals and in the market: derivative, platform, and breakthrough.

D. Client Consultation

Client consultation involves engaging the internal and external stakeholders to give them the opportunity to air their views, influence the project plans, and know what has been decided [13]. Top management, senior manager, and business user involvement were evaluated using their participation in *steering* (establishing criteria), *directing* (steering and solving conflicts), *acceptance* (evaluating outcomes), and *participation* in projects tasks, specifically setting requirements and building models.

E. Personnel

Competency attributes were used to evaluate the specialized skills required for the staff [2, 8, 10]. The items measure the effort required to deliver quality services during the project for competencies highlighted by Debortoli, Müller and Vom Brocke [19] as being relevant to BI and big data, including *technical*, *business*, *data*, and *analytical* competence. The *data scientist* item considers the involvement of an analytically competent person or persons on the team.

F. Client Acceptance

Lucas Jr [20] suggests that the decision-making style of the users is a factor in their ability to comprehend and accept the results of decision support systems. Thus, *system use* is a measure of client acceptance. System use covers the mandatory and voluntary use as a measure of system success [21]. Top management and senior management participation include their usages of system results in *decision-making* or decision-making and *acting* on the results.

G. Technical Tasks

Success factors for technical tasks are addressed from multiple perspectives such as data sources and IT infrastructure, which includes the availability and quality of data, as well as the heterogeneity and sophistication of IT infrastructure [2]. New and existing items previously used in the DeLone and McLean (2003) [22] information systems success model were used to create factors by [15] and used to evaluate the technical tasks.

H. Monitoring and Feedback, Communication, Troubleshooting

The project manager and team members, the organization, and the external environment are four interrelated groups of project success factors [1, 23]. *Project management competency* is a quality measure for change management, planning, or agile competency [19]. Other environmental factors evaluated include: project demographics, budget, duration, number of departments involved, number of organizations involved, and team size. Organizational demographics include revenue, number of employees, market share position, and revenue position.

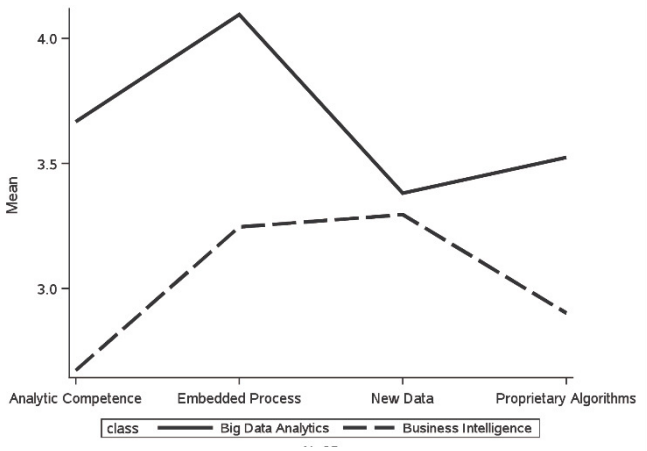


Fig. 1. Two-Cluster Model Structure

V. ANALYSIS

The latent class clustering technique was used to identify the homogeneous clusters. The dimensions were modeled in RStudio 1.0.153 using poLCA for the analysis. The four items for the BI/BD strategy in Table I were specified as indicators in the models. The two-cluster model was selected, given that it had the lowest Bayesian Information Criterion (BIC) measures and highest maximum likelihood. Fig. 1 includes the mean distribution for the two-cluster solution. The delivery of algorithms and analytic competency provide the most differentiation between clusters. Thus, each cluster has a strong differentiating feature. Cluster 1 was named Big Data Analytics and Cluster 2 was named Business Intelligence.

SAS Studio Release: 3.6 (Basic Edition) was used to perform the mean score ranking. The means were computed and ranked. The Wilcoxon test was used to compare the means of variables between the two classes of projects and to provide the significance of the comparison. The mean ranking and Wilcoxon score and significance for the variables for the two project classes are shown in Table I. The Wilcoxon scores with a p-value of less than 0.05 indicate that there are significant differences in the project types. The rank indicates the relative position—equated to importance—of the item for that project type.

VI. RESULTS

The 13 items for project attributes and demographics were not significant and are excluded from Table I for space reasons. Next, many of the results were expected. Specifically, *analytic competence* is ranked higher for big data analytics than for BI. Innovative technologies such as MapReduce- and Apache Hadoop-based systems exist specifically to process significant amounts of data and store

structured and unstructured data such as text, sound, images, video, etc. [9]. Thus, *data variety* was more highly associated with big data analytics, as expected. However, an unexpected result was that *data quality* was higher for big data analytics than BI. Given the *novelty* of big data analytics and its innovative uses, it is unusual that data quality would rank higher. Perhaps it can be explained by the big data solutions being used for the monetization of data. Monetizing big data means being able to process or derive intelligence from data that results in additional revenue from the sale, use, or reuse of the data or intelligence [3]. However, the organizational performance measures for revenue, costs, and strategy were not significantly different. The empirical results suggest BI and big data projects are differentiated based on analytics competence, providing models or algorithms. However, both project types have a similar level of technology uncertainty and pace. Next, while top management contribution is a critical success factor for all type of projects, senior managers and business users are significantly more involved in big data projects than they are in BI projects. The reason is elusive as the organizational performance measures do not differ significantly. However, the newness of the expected outcome, the technical, data, and analytical competence of the team, the data quality and variety, and the performance of the technical infrastructure are significant success factors for big data analytics.

VII. Conclusions

The practical implication is that sponsors and project managers should use this information to plan projects and to establish success criteria. First, based upon the literature review, it establishes the contingency factors for BI and big data analytic projects. That is, the measurement items offer a guide for defining the infrastructure, personnel, technical tasks, and governance for a project. For example, data privacy would be a critical factor in projects that produce proprietary algorithms and embed them in the business process. Thus, the project should include a specialist and activities for following data protection and privacy regulations. Consequently, the items can be used to facilitate discussions to assign accountable persons and human and financial resources to the project goals. The second area for using the results is to formulate success criteria that can be measured and monitored during the project. For example, measures could be defined around important aspects of service quality. This study contributes by adding clarity to BI and big data analytic project differences. The results of this study are not generalizable beyond decision support projects, and the findings are limited due to the small sample size.

References

[1] J. K. Pinto and D. P. Slevin, "Critical Success Factors Across the Project Life Cycle," *Project Management Journal,* vol. 19, no. 3, p. 67, Jun 1988.

[2] S. Olbrich, J. Pöppelbuß, and B. Niehaves, "Critical Contextual Success Factors for Business Intelligence: A Delphi Study on their relevance, variability, and controllability," in *45th Hawaii International Conf. on System Sciences*, Hawaii, 2012, pp. 4148-4157.

[3] P. Géczy, "Big Data Management: Relational Framework," *Review of Business & Finance Studies,* vol. 6, no. 3, pp. 21-30, Aug 2015.

[4] W. Yeoh and A. Koronios, "Critical Success Factors For Business Intelligence Systems," *The Journal of Computer Information Systems,* vol. 50, no. 3, pp. 23-32, Jul 2010.

[5] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electronic Markets,* vol. 26, no. 2, pp. 173-194, May 2016.

[6] T. Gilad and B. Gilad, "SMR Forum: Business Intelligence - The Quiet Revolution," *Sloan Management Review (1986-1998),* vol. 27, no. 4, pp. 53-61, Jun 1986.

[7] T. H. Davenport and J. Harris, *Competing on Analytics: The New Science of Winning*. Boston, MA, USA: Harvard Business School Press, 2007.

[8] J. Hammerbacher, "Information Platforms and the Rise of the Data Scientist," in *Beautiful data: the stories behind elegant data solutions*, T. Segaran and J. Hammerbacher, Eds., ed Sebastopol, CA, USA: O'Reilly Media, Inc., 2009, pp. 73-84.

[9] S. Sun, C. G. Cegielski, and Z. Li, "Amassing and Analyzing Customer Data in the Age of Big Data: A Case Study of Haier's Online-to-Offline (O2O) Business Model," *Journal of Information Technology Case and Application Research,* vol. 17, no. 3/4, pp. 156-165, Dec 2015.

[10] T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review,* vol. 90, no. 5, pp. 70-76, Oct 2012.

[11] R. M. Chang, R. J. Kauffman, and Y. Kwon, "Understanding the paradigm shift to computational social science in the presence of big data," *Decision Support Systems,* vol. 63, no. p. 67, Jul 2014.

[12] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big Data analytics and Computational Intelligence for Cyber–Physical Systems: Recent trends and state of the art applications," *Future Generation Computer Systems,* Nov 2017.

[13] R. J. Turner and R. Zolin, "Forecasting Success on Large Projects: Developing Reliable Scales to Predict Multiple Perspectives by Multiple Stakeholders Over Multiple Time Frames," *Project Management Journal,* vol. 43, no. 5, pp. 87-99, Oct 2012.

[14] L. A. Ika, "Project success as a topic in project management journals," *Project Management Journal,* vol. 40, no. 4, pp. 6--19, Dec 2009.

[15] G. J. Miller, "Decision Support Project: Project Success and Organizational Performance," DBA Thesis, Project and Program Management, SKEMA Business School, Lille, France, 2018.

[16] J. Thomas and J. Kielman, "Challenges for visual analytics," *Information Visualization,* vol. 8, no. 4, pp. 309-314, Jan 2009.

[17] H. Barki and J. Hartwick, "Measuring user participation, user involvement, and user attitude," *MIS Quarterly,* vol. 18, no. 1, pp. 59-82, March 1994.

[18] A. Shenhar and D. Dvir, *Reinventing project management: The diamond approach to successful growth and innovation*. Boston, MA, USA: Harvard Business School Press, 2007.

[19] S. Debortoli, O. Müller, and J. P. D. Vom Brocke, "Comparing Business Intelligence and Big Data Skills," *Business & Information Systems Engineering,* vol. 6, no. 5, pp. 289-300, Oct 2014.

[20] H. C. Lucas Jr, "Empirical evidence for a descriptive model of implementation," *MIS Quarterly,* pp. 27-42, Jun 1978.

[21] H. Barki and S. L. Huff, "Change, attitude to change, and decision support system success," *Information and Management,* vol. 9, no. 5, pp. 261-268, 1985.

[22] W. H. DeLone and E. R. McLean, "Information Systems Success: The Quest for the Dependent Variable," *Information Systems Research,* vol. 3, no. 1, pp. 60-95, Dec 1992.

[23] W. Belassi and O. I. Tukel, "A new framework for determining critical success/failure factors in projects," *International Journal of Project Management,* vol. 14, no. 3, pp. 141-151, Jun 1996.

[24] M. Halaweh and A. El Massry, "Conceptual Model for Successful Implementation of Big Data in Organizations," *Journal of International Technology and Information Management,* vol. 24, no. 2, pp. 21-34, Dec 2015.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 707–714
DOI: 10.15439/2018F262
ISSN 2300-5963 ACSIS, Vol. 15

Applying Formal Methods to Specify Security Requirements in Multi–Agent Systems

Vinitha Hannah Subburaj
School of Engineering, Computer Science and Mathematics
WTAMU Box 60767
West Texas A&M University Canyon, TX 79016 USA
Email: vsubburaj@wtamu.edu

Joseph E. Urban
Arizona State University
Tempe, AZ 85281 USA
Email: urban@asu.edu

***Abstract*—Security has become an important concern with the development of large scale distributed and heterogeneous multi-agent systems (MAS). One of the main problems in addressing security during the development of MAS is that security is often an afterthought. The cost involved to patch existing systems against vulnerabilities and attacks after deployment is high. If developers and designers can spend some quality time investigating security aspects before beginning to code then this cost can be reduced significantly. Also, using formal methods to specify the complex behavior of large scale software systems has resulted in reliable software systems. This research effort was focused on using formal methods early in the development lifecycle to specify security requirements for MAS. New solutions are emerging to fix security related issues, but how much thought gets in during the early phases of development in terms of security needs to be answered. In this paper, analysis of security requirements for MAS, existing solutions to secure MAS, and the use of formal methods to specify security requirements has been studied. Descartes – Agent, a formal specification language for specifying agent systems has been taken into study to model the security requirements of MAS early on in the development process. Functional specifications of MAS are modelled along with the non-functional security requirements using the Descartes – Agent specification language. A case study example is used to illustrate the specification of security requirements in MAS using the Descartes – Agent.**

***Index Terms*—multi-agent systems, security requirements, formal methods, Descartes - Agent**

I. Introduction

MAS are a set of software agents that work together to solve problems that are beyond the individual capacity of a single software agent. MAS are a comparably new software paradigm, which has been accepted widely in several application sectors that involve large and complex tasks. The autonomous, pro-active and dynamic problem solving characteristics of MAS have recently caught the attention of several application areas, such as: banking, transportation, e-business, and healthcare. In all these mentioned services, it is imperative that security must be assured. These services will face serious deployment issues if the security requirements are not being enforced. This approach is possible by considering the agent properties and the security aspects that relate with those specific properties.

The use of MAS in open, distributed, and heterogeneous applications, however may cause problems with security issues which in turn may affect the success of the various applications. Security in MAS is an upcoming field in a well-established field of study, such as security in networks, P2P, and web services communication. Hence, this paper analyzes the basic security concepts required to be applied to security of MAS.

This paper includes a review of the past and present work related to the security issues of MAS. Also, the research effort has studied the existing security technologies used as solutions to address the security issues of MAS. Mobile agents, host security, agent communication, and delegation are some of the current security technologies that are used to address security issues [1].

The need for systematic and secure system development has increased the use of formal methods. The following are some of the specific characteristics of using formal methods to specify secure software systems [5]:

- enable reasoning from logical/mathematical specifications of the behaviors of computing devices
- offer accurate proofs, so that all system behaviors meet desirable properties
- crucial for security goals
- rule out a range of attacks
- provide guidance for gapless construction and
- always use models.

Implementing formal methods in various areas such as verification of hardware system, embedded systems, analysis and testing of software has improved the quality of computer systems. There is a forecast that formal methods can bring similar improvement in the security of software systems. Formal methods have been associated with security

applications for a while [15], thereby offering new techniques for security goals across a wider range of components. Without the implementation of formal methods, security will always remain weak. In this paper, one such formal method has been used to specify the security requirements of MAS.

Wing, in her paper, has stated that security always had played a vital role in the development of formal methods in the 70s and early 80s [7]. There are a few questions that might arise regarding the formal methods. Has the scenario changed? Are the formal methods now ready to have a significant role in the production of more secure systems? The answer is yes, formal methods now play an important role in security systems. In this paper, limitations of formal methods, summary of the results on how model checking and theorem proving tools were discussed. Also, the challenges and opportunities for formal methods in analyzing the security of systems, beyond the protocol level are also elaborated. Formal methods need integration with 1) other methods that address issues on formalization (analysis must include several factors such as risk, hazard, fault, and intrusion detection) and 2) into the entire software development lifecycle (such as during requirements analysis, testing, and simulation). Finally, there is a necessity to introduce the human factor (cannot be ignored), which in principle is part of the system's environment. Research conducted on modeling of human behavior, human-computer interaction, and management of processes and organizations can all aggregate the formal nature of research on formal methods.

The remainder of this paper is structured as follows:
Section 2 discusses the existing work related to security issues in MAS, security solutions to MAS, and the use of formal specification to specify security requirements in MAS. Section 3 discusses the earlier extensions done to the Descartes specification language to specify agent systems. Section 4 discusses the security framework developed in this research effort to specify the security requirement of MAS using Descartes – Agent. Section 5 provides a case study example that illustrates the application of the developed security framework with an e-commerce application. Section 6 discusses the lessons learnt and Section 7 summarizes the paper with a brief discussion of future work.

II. Related Work

Jung, et. al [2] surveyed existing research efforts that exist related to security in MAS, with a special focus on access control and trust/reputation. The paper concluded that security of agent based environments is critical. In spite of several efforts, many problems still remain and appear to be challenging with the continuous development of new technologies that are developed.

The research described in the research paper [3] identified the various security issues encountered by MAS. In order to assure MAS security, the paper examined the following: 1) basic concepts of security in computing, 2) characteristics of agents and MAS that introduce new threats, and 3) different strategies to prevent attacks. However, despite the similarities, security in MAS has specific requirements which need the autonomy, mobility, and other agent features that are not usually found in most conventional systems.

A model (based on the concepts and models regarding agent's role and communications) is presented [4] for securing MAS. The model provides an adequate way to ensure the security requirements and design are combined with system functionalities during the development process. The proposed model also incorporates the general security requirements at the agent and system levels. The paper has considered and addressed several system level threats, such as 1) corrupted mobile agents attack the main system host, 2) fake agent, 3) insecure communication among the platforms, and 4) agent level threats. The research work has attempted to extend the Gaia methodology with the security model. Further research work is needed in order to provide developers with security solutions for MAS based on the Gaia methodology.

A secure-critical system is difficult to develop and there are several known research issues regarding the security weaknesses in many sectors. Hence, a good methodology to support secure systems development is immediately needed. The research paper [6] presents the aim to assist the difficult task of developing security-critical systems using an approach of the Unified Modeling Language. The extension UMLsec of UML [6] (that allows expressing security relevant information within the diagrams) in a system specification is described in this paper. The UMLsec is defined in the form of a UML profile using the standard UML extension mechanisms. In particular, the related constraints provide criteria to classify the security aspects of a system design, by attributing to the formal semantics of a simplified fragment of UML. Formal evaluation is possible since the behavioral parts of UMLsec are considered with formal semantics. Hence, even the security experts who undertake a formal evaluation for certification purposes also may benefit from the possibility of using a specification language that may be more adaptable than some conventional formal methods.

Even though security has a major role in the development of MAS, security requirements are usually considered after the design of a system. The main reason is because of the fact that agent oriented software engineering methodologies have not unified security concerns throughout their developing stages. Mouratidis and Giorgini [12, 20] in their paper, introduce extensions to the Tropos methodology to enable them to model security concerns throughout the entire development process. This paper also describes the new concepts and modeling activities getting integrated to the current stages of Tropos. Tropos is characterized by the following three key aspects.

- deals with all the phases of a system development, adopting a uniform and homogeneous way,
- attends to the early requirements (emphasizing the need to understand organizational goals), and
- builds a model of the system that is refined and extended from a conceptual level to executable level, by a sequence of transformational steps.

The Tropos methodology includes five main software development stages, such as early and late requirements analysis, architectural design, detailed design, and implementation. In order to extend Tropos with security related concepts, factors such as security concepts and security modeling activities are detailed in the paper. A real life case study from the health and social care sector is used to illustrate the approach using Single Assessment Process (eSAP) system.

MAS have become a promising architectural approach for constructing Internet-based applications. Recent research work in software architecture have resulted in the necessity to truly define languages for designing and formalizing agent architectures and more specifically secure ones. This paper describes the basic fundamentals for an architectural description language (ADL) to specify secure MAS. Mouratidis, et al. [13] in their paper introduce a set of system design primitives that is conceptualized with the Z specification language to build secure MAS architectures. The main concepts of SKwyRL-ADL, including the security aspects, are described in this paper. The Z specification language is used to describe SKwyRLADL concepts. Z is widely used as a formal specification language as it is clear, concise and easy to learn. The three sub-models of SKwyRLADL: agent model, security model, and architectural model are detailed in this paper. The concept is applied on an e-commerce example to illustrate the research effort. The illustration involves the description of formally specified architectural aspects, such as interfaces, knowledge bases, security objectives, security mechanisms, and plans of the e-Media system.

III. Background

The Descartes specification language, developed by Urban [10] was designed to be used throughout the software life cycle. The relationship between the input and the output of a system is functionally specified when using this specification language. Descartes defines the input data and output data and then relates them in such a way that output data becomes a function of input data. The data structuring methods used with this language are known as Hoare trees. These Hoare trees use three structuring methods namely direct product, discriminated union, and sequence.

Direct product provides for the concatenation of sets of elements. Discriminated union provides for the selection of one element out of a set of elements. A plus sign (+) is used to denote discriminated union. Sequence represents zero or more repetitions of a set of elements. Sequence is indicated by an asterisk (*) suffixed to the node name.

By definition of Hoare trees, a sequence node is followed by a sub node. A single node can accommodate a sequence of direct product or a sequence of discriminated union. In the Descartes specification language, a literal is any string that is enclosed within single quotes. Consider the following example,

agent

‘autonomous_agent’ wherein autonomous_agent is a literal.

The Descartes specification language was extended in 2013 by Subburaj [11] for specifying complex agent systems. The extensions made to the Descartes specification language follows a top-down modular development allowing for the decomposition and incremental development of large agent systems. Six new concepts were added to Descartes for specifying and validating agent software systems. The added concepts were: (1) agent construct; (2) agent goal; (3) agent attributes; (4) agent roles; (5) agent plans; and (6) communication protocol.

Agent systems consist of multiple autonomous agents. Each of the agents has a specific goal to achieve and a set of actions to perform in order to achieve a goal. The agent construct in an agent system is used to define the behavior of an agent, including the goal, different roles, type of events, the plans, and the knowledge base. Each agent in an agent system has a structure. The notion of declaring an agent can be compared to the identification of objects in an object oriented methodology. The declaration of an agent module is pre-pended with a unary “agent” reserved word. Consider the following example,

agent AGENT_MODLUE_NAME_(INPUT)

Every agent has a goal of achieving a certain state or task. For example, imagine an agent that would start running with a goal of cleaning a house. The initial goal of such an agent is to clean the house and perform actions accordingly to achieve the goal statement. In Descartes - Agent, the agent goal is specified by using a new primitive, “goal”, added to the Descartes syntax. An agent goal is an important attribute to be specified in an agent system. The plans that are executed by an agent solely depend upon the goal defined for a specific agent.

The agent roles are used to identify the key roles in an agent system. The notion and description of role models has been adopted from the Gaia methodology [8].

One of the most important aspects of agents is that they act autonomously to achieve their goals. This characteristic of agents to act autonomously in an environment is realized through the plans part in an agent system. The plans consist of a sequence of actions that an agent will take when a corresponding event occurs. The first part of the plan specified the list of events that trigger the execution of a specific plan by the agent. The second part context describes the contexts when the plan is applicable. The context part is

used to specify the current beliefs of the agent system. This part consists of a set of rules that can be specified with respect to specific agents. The context part also communicates with the knowledge/belief component in the agent framework to update and reads agent specific rules. The next extension is the reserved keyword "plans" used to specify the agent plans. The keyword triggered_events is used to list the triggered events. The keyword context is used to specify the agent specific rules and belief. The keyword method is used to specify the list of actions to be taken. In order to specify the context of the plan, new logical primitives were added to Descartes - Agent.

The knowledge/belief base in an agent system contains the knowledge that the agent has about itself and its environment. An agent's plan reads and modifies the knowledge/beliefs base. The knowledge/belief base consists of logical rules that are known initially before the agent starts to execute the plans. Also, based upon the execution of plans by the agents in the agent systems, the knowledge/belief base gets updated according to a current belief. In the Descartes - Agent processor, the knowledge/belief base was implemented as a separate component. The processor before executing the agent plans and also after executing the agent plans will access the knowledge/belief component to take appropriate decisions.

The last extension to Descartes - Agent for specifying agent systems is the communication protocol. Agents interact with other agents in the agent system and also with the environment to realize agent goals. The communication protocol in the extended Descartes is set up by the name tag (in upper case letters) of the calling agent module within parentheses followed by a period and the name of the relevant message within parentheses followed by the "^" symbol and then the name tag (in upper case letters) of the called agent module within parentheses.

IV. Specifying Security Requirements Using The Descartes – Agent

A. MAS properties

Wooldridge and Jennings [15] software agents come with the following properties:

Autonomy: An agent has its own goal and the ability to operate without any human intervention; more importantly, agent has control over its own state and can regulate its own functioning without outside assistance.

Sociability: An agent is capable of interacting with other agents and humans using an agent communication language. This approach allows an agent to seek and provide services.

Reactivity: An agent is capable of perceiving and acting on its close environment. The agent can respond to changes that occur in its surroundings.

Pro-activeness: Agents are not only capable of responding to the stimulus from their surroundings, but are also capable of exhibiting a goal-oriented behavior by taking initiatives.

In addition, there are some other characteristics, such as situadeness, mobility, rationality, veracity, and benevolence. Situadeness means agents are capable of sensing a special condition based on the inputs received from the environment.

Table I.
Agent properties and associated security concerns

Agent property	Description	Security concerns
Situatedness	If the agent gets to sense the input from its local host, then problems are less. But, instead if the information is coming from the Internet then there comes the problem of trust.	Trust, authentication, and integrity
Autonomy	Malicious agents can intrude without any request from humans or other agents.	Authorization
Social Ability	Enabling secure communications among agents and between humans and agents.	Confidentiality, integrity, availability, accountability, and non-repudiation
Mobility	By being able to self-migrate from one platform to other platforms, agents are prone to a number of security attacks.	Authentication, confidentiality, integrity, privacy, and faulty tolerance Damage, DoS, breach of privacy or theft, harassment, social engineering, event-triggered attack, compound attacks, masquerading, unauthorized access, copy-and-reply, and repudiation
Cooperation	Many agents cooperatively working together to access resources and internal status of other agents. This leads to security concerns.	Authentication and authorization

The term software agents covers a wide range of more specific agent types. Etzioni and Weld [16] and Franklin and

Graesser [17] provide a list of attributes that each agent must possess to a lesser or greater degree. The software agent attributes are as follows:

"Reactivity: the ability to selectively sense and act

Autonomy: goal-directedness, proactive and self-starting behavior

Collaborative behavior: can work in concert with other agents to achieve a common goal

Communication ability: the ability to communicate with persons and other agents with language more resembling humanlike "speech acts" than typical symbol-level program-to-program protocols

Inferential capability: can act on abstract task specification using prior knowledge of general goals and preferred methods to achieve flexibility;

Temporal continuity: persistence of identity and state over long periods of time

Personality: the capability of manifesting the attributes of a "believable" character such as emotion

Adaptivity: being able to learn and improve with experience

Mobility: being able to migrate in a self-directed way from one host platform to another."

B. Security requirements in MAS

The autonomous, pro-active, and dynamic nature of software agents thought proven to solve challenging problems, also comes with security concerns. Often, these security aspects get unnoticed until the deployment of the end-deliverables. Patching the security flaws after deployment has always resulted in high costs.

From the above properties, it is evident that software agents operate in an open environment and are free to interact with their surroundings to achieve their goal. This openness gives rise to a number of security and trust issues. Some of the commonly occurring security problems with agent based systems [2] are: confidentiality, integrity, availability, accountability, and non-repudiation.

Based on agent characteristics, there [2, 12] have been presented a list of security requirements of the MAS. Table I associates agent characteristics with their associated security problems.

C. Descartes – Agent Security Specifications

Among the list of security concerns listed above, in this paper we focus on two concerns namely: access control and confidentiality.

To provide access control there are two steps involved: first is to provide authentication to a group of agents enabling them to establish their true identity and then authorization that allows us to define the type of access privileges each agent obtains.

Every agent has a goal of achieving a certain state or task. In the secure agent framework specified by using the primitive, "goal" being prepended by a "!" symbol. With the specification of the secure agents, the goal is enclosed within a * symbol denoting the goal of secure agents.

Figures 1 and 2 illustrate the Descartes – Agent framework for specifying MAS and the Descartes – Agent secure framework for specifying MAS.

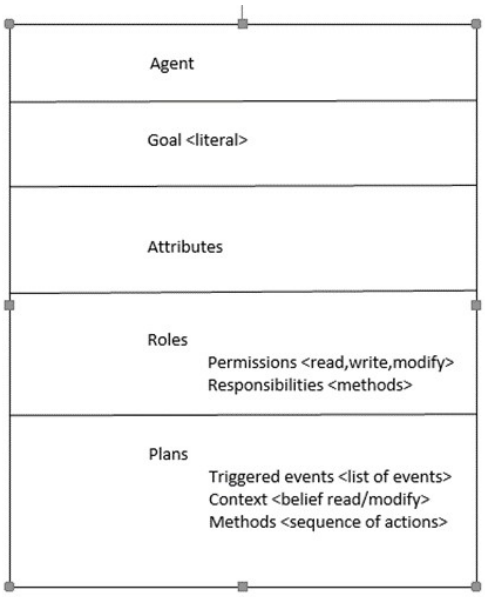


Fig 1. The Descartes – Agent framework for MAS

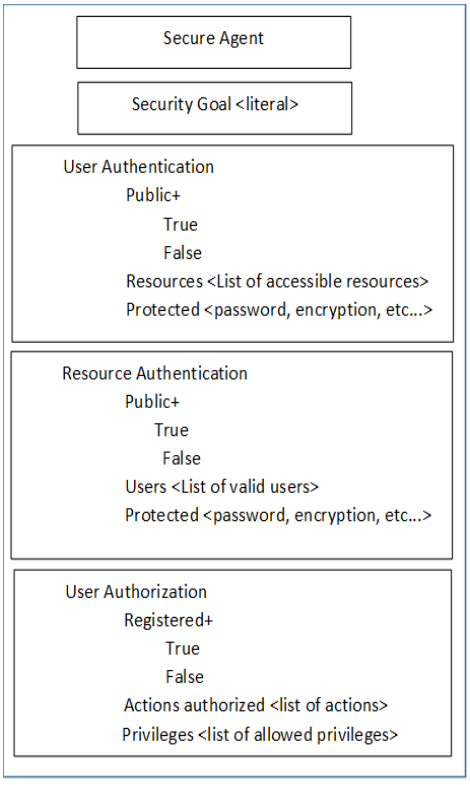


Fig 2. The Descartes – Agent security framework for MAS

1) Formally specifying authentication requirements for software agents

The authentication block in the security framework specified in the above Figure, is decomposed as user authentication and resource authentication. Within the user authentication, the node named public is a discriminated union meaning the public attribute can either be true or false. Resources in a secure agent system define the different types of resources that the secure agent can access. The protected attribute defines the method of authentication used by that user. There can be unique credentials such as passwords, encrypted passwords, and public-key-infrastructure schemes. The resource block is the same as the user except that there is a list of users that are given permission to access particular resources. Figure 3 illustrates the specification of the authentication requirement using Descartes – Agent.

```
agent TRANSACTION_AGENT
goal
   *! To successfully complete a secure transaction *
user1
   public+
        true
        false
      users
        files
            'list of files that can accessed'
          servers
            'list of servers that user1 can talk to'
      protected
          password*
            alphanumericstring

resource1
   public+
        true
        false
      resources
        users
            'list of users that can this resource'
      protected
         password*
            alphanumericstring
```

Fig 3. Authentication example using secure Descartes - Agent

2) Formally specifying authorization requirements for software agents

The user authorization block specified in the secure agent framework consists of three parts: registered, actions authorized, and privileges. The first part of the user authorization block specifies whether the user is a registered user or not. The second part of the authorization block allows one to specify all the actions that an authorized user can perform. The third part allows for the specification of all the privileges or access rights to specific resources. In order to specify the access privileges of users, new logical primitives were added to Descartes - Agent. Figure 5 illustrates the specification of an authorization requirement using Descartes – Agent. Figure 4, lists the newly added authorization primitives.

(<userx>)_HAS_READ_PERMISSIONS_TO (<resourcex>) – user X has read permission to access a specific resource

(<userx>)_HAS_WRITE_PERMISSIONS_TO (<resourcex>) – user X has write permission with a specific resource

(<userx>)_HAS_APPEND_PERMISSIONS_TO (<resourcex>) – user X has append permission with a specific resource

(<userx>)_HAS_EXECUTE_PERMISSIONS_TO (<resourcex>) – user X has execute permission with a specific

Fig 4. Authorization primitives

```
user2
      registered+
          true
          false
       actions_authorized
          update_database
            'update_the_database'
          receive_payment_info_from_agenty
            'receive_payment_info'
          process_payment
            'process_payment_info'
          confirm_transaction
            'confirm_the_completion_of_a_transaction'
        privileges
              (USER2)_HAS_WRITE_PERMISSIONS_TO
              (TRASACTION_RECORD)
              (USER2)_HAS_EXECUTE_PERMISSIONS_
              TO_(COMMIT_QUERY)
```

Fig 5. Authorization example using secure Descartes – Agent

3) Design and Implementation

The secure Descartes – Agent specifications discussed in Section IV can be transferred into UML design and then into implementation code. AGENT UML [22][23], an extension of Unified Modeling Language (UML) was proposed to facilitate developers with a smooth agent development process. The extended Descartes has already been specified using AUML in [21]. The extended Descartes – Agent security requirements can be specified using use case, sequence, and communication protocol diagrams. For instance, a sequence diagram in AUML is a diagram that describe sequence of messages between agents that exchange messages through protocols. These diagrams define the different agent roles, constraints, and the messages that are ordered according to a time axis. Sequence diagrams use the following basic components along with other components to describe a communication pattern between agents: agents and agent roles, agent lifelines and thread of interaction, connectors, messages, and conditions on messages. Authentication and

authorization requirements can be enforced via protocols every time a message transfer occurs between the agents. Since the Descartes specifications are in an executable form, with formal specification constructs close to that of programming, implementation of these Descartes – agent specifications can be done in any high-level programming language. With Descartes – Agent specifications, the transition from specification to design and then to implementation happens seamlessly.

V. Case Study Example

MAS are used to provide efficient e-commerce solutions, but different security related issues are associated with the agent solutions of e-commerce applications. A case study example of a real time MAS for e-commerce applications [19] is described for illustrating the security framework introduced in this research effort. The real time multi-agent architecture for an e-commerce application consists of four different types of agents namely: UserAgent, QuotingAgent, TrendWatchingAgent, and BuySellAgent.

The main goal of the USER_AGENT is to determine the user requirements such as the risk level, amount of money to spend, and the market sector preferences. The USER_AGENT specifies the quality threshold to ensure if the actual stock price lies within the threshold value [9]. Security requirements associated with this user agent requires authentication, authorization, and confidentiality. The following Descartes – Agent specification adds the security requirements discussed in Section IV.B to the USER_AGENT. Figure 6 illustrates the specification of USER_AGENT that includes authentication and confidentially security requirements.

VI. Lessons Learnt

Following were the lessons learnt out of this research effort. Security is a major issue when it comes to addressing requirements for MAS. The existing Descartes - Agent specification language constructs along with few newly added ones were successfully used to specify the security specifications of MAS. Case study examples similar to but not limited to the one described in the paper can be used to illustrate the extensions made to the Descartes - Agent specification language. The formal executable specification demonstrated the possibility of converting the security specification into design and then into implementation.

Some of the drawbacks identified out of this research effort are as follows: general framework for understanding the security requirements of MAS is not available; automated design and code generation techniques from the formal specification languages used to specify secure MAS is also scarce in the literature; efficient ways to rank and specify security requirements according to importance is not adequately discussed.

```
agent
    USER_AGENT_(RISK_LEVEL)_AND_(AM
    NCE)
goal
   *!to_determine_user_requirements_securily*
attributes
   RISK_LEVEL
          INTEGER
   AMT
          FLOATING_POINT
   PREFERENCE
          STRING
   stock_price
          'value_read_from_knowledge_base'
   quality_threshold
          'value_read_from_knowledge_base'
   get_price
          'message_sent'
user_authentication
   public+
          true
          false
       resources
          files
              'list_of_files_that_can_be_accessed_
          servers
               'list_of_servers_that_can_be_acces
          protected
                password*
                     alphanumericstring
user_confidentiality
    registered+
           true
           false
    actions_authorized
         check_stock_price
             'check_stock_price'
         check_risk_level
             'check_risk_level'
         decide_on_stock
              'make_decision_based_on_price_val
         confirm_transaction
             'confirm_the_completion_of_a_transa
     privileges
         (USER_AGENT)_HAS_READ_PERMI
         ROFILE)
         (USER_AGENT)_HAS_WRITE_PERM
         ECORD)
         (USER_AGENT)_HAS_EXECUTE_PE
         ESHOLD_QUERY)
```

Fig 6. Case study example using secure Descartes - Agent

VII. Summary and Future Work

A security framework that allows developers to formally specify the security requirements of MAS has been discussed in this paper. The security framework has been built as a part of the Descartes – Agent formal specification language. The key point on the developed security framework is that it can be applied early on in the development process of MAS. The identification of these security requirements early during the development of agent systems reduces the security patching cost involved with MAS development. One of the main benefits of using Descartes – Agent is that it allows partial specifications to be developed and executed. This feature allows one to specify security requirements with a high-level of abstraction.

Three important security issues with MAS, namely authentication, authorization, and confidentiality, were taken into study. The security framework built in this research effort allows for the specifications of security requirements that would implement these security solutions in MAS. The challenging aspect of incorporating a formal executable specification language to specify security requirements for MAS has been accomplished in this research effort. A case study example has also been discussed to illustrate the use of the security framework built in this research effort. The case study discussed in this paper serves as a basis for formally specifying security requirements for MAS and can be applied to different applications in similar fields.

As future work, the security framework developed will be extended to provide solutions to other security issues, such as trust, integrity, availability, accountability, and non-repudiation. Extending the security framework to enforce security with distributed MAS will be a future challenging effort.

References

[1] N. Borselius. "Security in multi-agent systems," Proceedings of the International Conference on Security and Management (SAM'02). 2002.

[2] Y. Jung, M. Kim, A. Masoumzadeh, and J. B. D. Joshi, "A survey of security issue in multi-agent systems," Artificial Intelligence Review, vol. 37, no. 3, pp. 239–260, Apr. 2011.

[3] R. C. Cavalcante, I. I. Bittencourt, A. P. D. Silva, M. Silva, E. Costa, and R. Santos, "A survey of security in multi-agent systems," Expert Systems with Applications, vol. 39, no. 5, pp. 4835–4846, 2012.

[4] Y. Hedin and E. Moradian, "Security in Multi-Agent Systems," Procedia Computer Science, vol. 60, pp. 1604–1612, 2015.

[5] S. Chong, J. Guttman, A. Datta, A. C. Myers, B. Pierce, P. Schaumont, T. Sherwood, N. Zeldovich, "Report on the NSF workshop on formal methods for security," CoRR, vol. abs/1608.00678, 2016.

[6] J. Jürjens, "UMLsec: Extending UML for Secure Systems Development," ≪UML≫ 2002 — The Unified Modeling Language Lecture Notes in Computer Science, pp. 412–425, 2002.

[7] J. Wing, "A symbiotic relationship between formal methods and security," Proceedings Computer Security, Dependability, and Assurance: From Needs to Solutions (Cat. No.98EX358).

[8] Cernuooi, L., et al., "The gaia methodology: basic concepts and extensions," in Multiagent systems, Artificial Societies and Simulated Organizations. 2004. 11(2). P. 69-88.

[9] V. H. Subburaj and J. E. Urban, "Formal Specification Language and Agent Applications," Studies in Big Data Intelligent Agents in Data-intensive Computing, pp. 99–122, 2015.

[10] Urban, J. E., "A Specification Language and its Processor," Computer Science Department. University of Southwestern Louisiana. 1977.

[11] V. H. Subburaj and J. E. Urban, "A formal specification language for modeling agent systems," 2013 Second International Conference on Informatics & Applications (ICIA), 2013.

[12] H. Mouratidis and P. Giorgini, "Secure Tropos: A Security-Oriented Extension Of The Tropos Methodology," International Journal of Software Engineering and Knowledge Engineering, vol. 17, no. 02, pp. 285–309, 2007.

[13] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.

[14] Hussain, Shafiq, Peter Dunne, and Ghulam Rasool. "Formal Specification of Security Properties using Z Notation," Research Journal of Applied Sciences, Engineering and Technology 5.19 (2013): 4664-4670

[15] Wooldridge, M., Jennings, N.R.: Intelligent agents: Theories, Architectures and Languages, January 1995. Lecture Notes in Artificial Intelligence, vol. 890, ISBN 3-540-58855-8

[16] O. Etzioni and D. Weld, "Intelligent agents on the Internet: Fact, fiction, and forecast," IEEE Expert, vol. 10, no. 4, pp. 44–49, 1995.

[17] S. Franklin and A. Graesser, "Is It an agent, or just a program?: A taxonomy for autonomous agents," Intelligent Agents III Agent Theories, Architectures, and Languages Lecture Notes in Computer Science, pp. 21–35, 1997.W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *1987 Proc. INTERMAG Conf.*, pp. 2.2-1–2.2-6.

[18] N. Borselius, "Mobile agent security," Electronics & Communication Engineering Journal, vol. 14, no. 5, pp. 211–218, Jan. 2002

[19] L. C. Dipippo, V. Fay-Wolfe, L. Nair, E. Hodys, and O. Uvarov, "A Real-Time Multi-Agent System Architecture for E-Commerce Applications," Jan. 2000.

[20] H. Mouratidis, P. Giorgini, and G. Manson, "Modelling secure multiagent systems," Proceedings of the second international joint conference on Autonomous agents and multiagent systems - AAMAS 03, 2003.

[21] V. H. Subburaj, J. E. Urban, "Intelligent Agent Software Development Using AUML and the Descartes Specification Language," Proceedings of the 2nd IEEE International Workshop on Object / component/service-oriented Real-time Networked Ultra-dependable Systems (WORNUS 2011), pp. 297-305, March 28, 2011.

[22] B. Bauer, J. Muller, and J. Odell, "An extension of UML by protocols for multi-agent interaction," Proceedings Fourth International Conference on Multi Agent Systems, pp. 207-214, 2000.

[23] M. P. Huget and J. Odell, "Representing Agent Interaction Protocols with Agent UML," Agent-Oriented Software Engineering V Lecture Notes in Computer Science, pp. 16–30, 2005.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 715–724
DOI: 10.15439/2018F235
ISSN 2300-5963 ACSIS, Vol. 15

An Approach to Transforming Requirements into Evaluable UI Design for Contextual Practice - A Design Science Research Perspective

Matthias Walter
Chair of Information Systems, esp. IS in Manufacturing and Commerce, Technische Universität Dresden
Dresden, Germany
matthias.walter3@tu-dresden.de

***Abstract*—We contribute a methodical approach in the context of IS design science research to develop UI prototypes for evaluations in practice-oriented research. Based on previous research on improving IS support for early product cost optimization, we present and discuss our methodical approach to derive UI prototypes based on an evaluated requirements model. The objective of the outlined approach comprising different steps is to derive a clickable UI prototype that is feasible for further artifact evaluation within institutional environments. Together with experts from the practice of software engineering we iterated through the working steps of the elaborated approach to determine its feasibility to derive a prototype and moreover, generate visual examples for each step to improve the approach's comprehensibility. In addition to the description of the approach itself we point to significant hurdles that have arisen with the application of it in order to generate learnings for other research projects.**

I. Motivation

In times of globalization, demand rises for agility, innovation, and quality. Furthermore, shortened product life cycles and an amplified variety of product models have increased pressure on product manufacturers [1], [2]. In order to keep up with the global competition as such, optimizing product costs throughout a product's life cycle has become a major driver for economic success. To ensure the long-lasting economic success of products in the upcoming decades, organizations have been attempting to optimize product costs for the overall product life cycle. This is especially true for the discrete manufacturing industry, where products like cars, trucks, airplanes, and high-tech machinery are assembled out of thousands of globally sourced components [1].

Figure 1 describes the cost situation in the life cycle of such products. Surprisingly, 90% of all product costs are determined before production starts – and, thus, in the phase of product development. Linking this fact to the idea of product cost optimization to ensure economic success, it becomes obvious that product development phases offer the most potential to optimize product costs. Despite this immense potential, there is a lack of information system (IS) support for product cost optimization during product development [3], [4], [5].

In our long-term research project, we aim to improve IS support for product cost optimization during product development and, therefore, aid the industry in making use of the cost optimization potential. Due to the practical relevance of the research problem and the demand for new IS approaches within the industry [4], [5], we initiated a research project together with the software corporation SAP SE. The resulting research collaboration, which includes various international companies, is framed by a design science research (DSR) approach [6]. First, we worked out a detailed problem identification together with companies among the discrete manufacturing industry [4], [5]. To overcome the identified problems, we elaborated major implementation challenges and requirements, which were further transferred into a requirements model [1]. After an industry evaluation of the requirements model [1], we now aim to design possible solutions to improve the support of IS for early product cost optimization and, thus, use the potential of product development phases (Figure 1).

In order to exploit the full potential of our industry research collaboration, it is recommended to evaluate DSR artifacts in their natural setting [7]. For this purpose, it is necessary to elaborate instantiations of potential problem solutions that can be evaluated by industry experts in practice. The question is: how can this be done?

Although DSR is generally gaining popularity in IS research [8], there are only a few research contributions that provide guidance for the design of at least partially instantiated artifacts in the context of DSR [9]. Further literature analyses have shown that the majority of the designed artifacts in DSR research are of type method or model [8] and can, therefore, only provide limited guidance for our practice-oriented approach to design potential solutions. In addition, the need to develop further research methods that immerse researchers in practice environments has been highlighted in the literature [10].

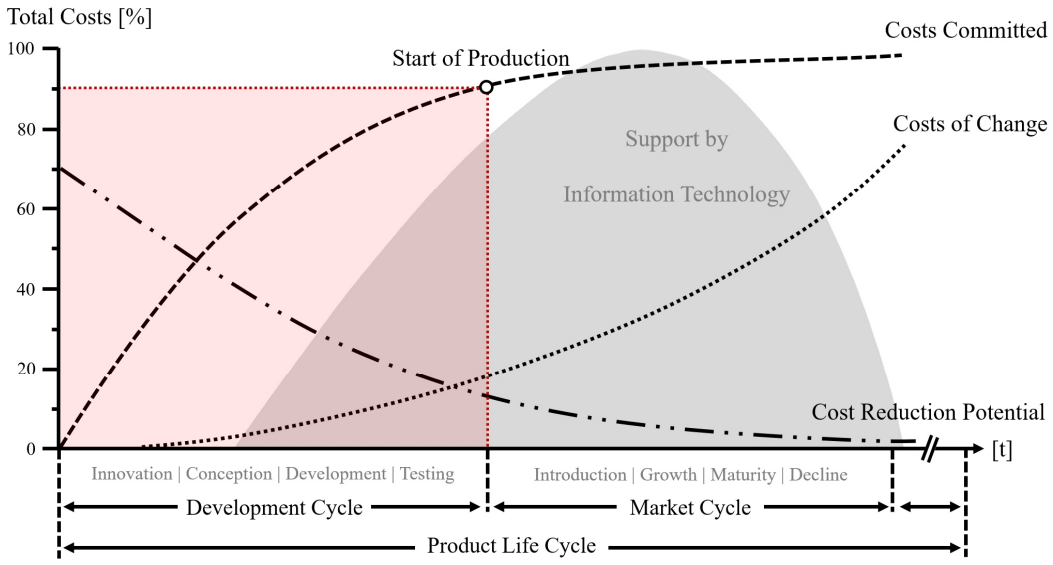


Fig. 1 Cost commitments and reduction potential along product life cycle [4], [11]

To improve this area of research, we would like to depict and discuss a methodical approach to develop user interface (UI) prototypes based on validated requirements. This approach has helped us in our industry collaboration research project, because such prototypes can be evaluated intuitively by experts without requiring a dedicated theoretical knowledge base. In addition, the development of UI prototypes requires considerably less efforts than regular software-based prototypes. Driven by the question of how to elaborate such evaluable UI designs, we focus the following research questions:

RQ1: Based on the elaborated requirements model, what could a methodical approach look like to transform requirements into UI prototypes that are suitable for iterative practical evaluations?

RQ2: What are the challenges of such an approach, and how can possible hurdles be mitigated?

To answer these research questions, we briefly introduce the research project and the research results to date. We then explain the elaborated approach to develop the UI prototype and explain the individual working steps in detail with the support of visual examples. The paper concludes with a discussion of results and an elaboration of lessons learned based on the application of this approach in the context of our research project, including possible pitfalls to provide valuable insights for other researchers.

II. Background

Though there are IS available that aim at supporting product cost calculation during product development [12], a lack of functionality and the demand for enhancements has been identified [3]. Building on this, in our initial research steps, we identified missing support for early product cost optimization and the resulting drawbacks for the industry [4], [5]. To establish a substantiated foundation for further research activities in a DSR context [13], we elaborated and evaluated a requirements model with experts from international companies within the discrete manufacturing industry [1]. Moreover, we identified major implementation challenges that have an impact on the solution acceptance in the industry and, therefore, will have an impact on the current phase of our research: solution design.

In theory, the DSR artifact design and development process is described as a rather individual and creative engineering process [9], [14]. Though there are general approaches available in the literature describing this process, the lack of guidance for artifact design in IS literature is evident [9].

Seeking such processual guidance for developing a potential solution, we had to consider an important circumstance in our research domain: Product cost optimization measures are derived in a rather unpredictable context based on deliberations of various stakeholders (e.g., product developers, production process engineers, production planers, cost controllers, and purchasers) and their collaboration [4]. Such deliberations do not follow best-practice patterns, but originate from an evolving knowledge-base along the product development phase, and therefore, result in dynamic, context-driven product cost optimization processes. Such processes can be classified as one type of emergent knowledge processes, as argued in our requirements elaboration [1].

Hence, potential solutions must support dynamically changing processes with a bandwidth of deliberations and tradeoffs based on complex, evolving expert knowledge bases within the organization, which is exclusive to product development [15].

Fig. 2 Degree of process specification [1], [16]

With this justificatory knowledge in mind, developing potential solutions becomes a challenge: First of all, it is important to determine the right degree of process support for experts using the IS without being too restrictive to limit their capabilities while carrying out product cost optimization across their organization. In addition, following the argumentation in [7] and [16], the developed solutions must be evaluated in the specific environment of their application. In our case, this requires an evaluation where business experts try to use our prototypes as support for their product cost optimization processes.

Due to the strong tie of our research problem to an organizational context, it is necessary to reflect whether the DSR methods to develop such a UI prototype fit our research context. This is especially important since it has been revealed that DSR methods consider organizational intervention to be of lesser importance [17]. In contrast to this, [10] argued to further develop specific methods aiming at the co-constitutional character of user contexts and institutional environments.

Such co-constitutional aspects can be approached in multiple ways. [18] criticized the strong sequencing of DSR, which separates developing artifacts from evaluating artifacts. To overcome this separation, action design research (ADR) has been recommended as a DSR method to closely link development to evaluation [18]. Further recommend-dations have been proposed in the literature [19]. Based on a comparative analysis of DSR with the constructive research approach, [19] highlighted the potential to improve DSR methods by developing best practices for collaborative development.

Beyond the DSR discourse, it is worth examining how IS prototypes are being used in practice. One area of regular prototype development and evaluation is agile software development (e.g., rapid prototyping) [20]. One specific example within the area of development research is mockup-driven development [21]. This approach makes use of UI prototypes (mockups) to receive early and continuous feedback to guide modeling and, thus, align further application development. This coincides with the intention for ADR as a method [18] to combine development and evaluation of artifacts into iterative steps.

We would like to take advantage of the experiences that have been made with mockup-driven development approaches to build a bridge between our scientific research and the evaluating group of practitioners. By this attempt, we intend to enable business experts within the discrete manufacturing industry to quickly grasp core intentions of our prototypes without preceding knowledge transfer. Furthermore, less effort would be exerted to create UI mockups instead of software prototypes, allowing shorter artifact iterations. The use of UI prototypes as a feasible option for DSR artifact evaluation has already been shown in other DSR projects [22].

III. Methodical Approach

The purpose of this paper is to draft and discuss a processual approach to develop evaluable UI prototypes to enhance practical research collaborations. This practical research collaboration is important for the problem solution, as product costing has a strong focus on expert knowledge [5], [23]. Moreover, industrial practice is the most important source of information for cost optimization projects [24]. Therefore, the access to knowledge from practitioner communities is fundamental for our research [1]. At the same time, such knowledge ensures that our research is relevant to practice [25].

Within our long-term design science research project, we elaborated and evaluated a requirements model [1] that included 30 detailed requirements in combination with implementation challenges on the basis of industry expert interviews and focus groups. In addition, the knowledge exchange with domain experts helped us to gain a holistic understanding of applied product cost optimization processes among different companies in an applied environment [5]. This was essential to develop a rather abstract approach – with respect to emergent circumstances toward the degree of process specification (see Background) – toward early product cost optimization, which has been agreed upon by industry experts (Figure 3).

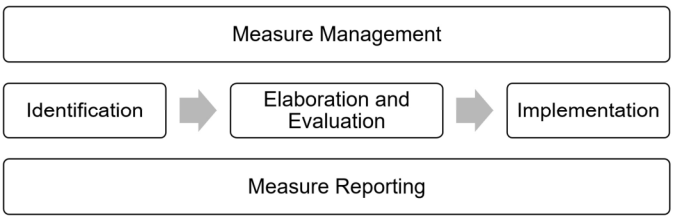


Fig. 3 Approach to support early product cost optimization [1]

As this approach provides support for the maturation of individual optimization measures (e.g., make-or-buy decisions, evaluation of alternative product concept or production processes, or optimization of logistic costs) [1], [5], our UI prototype development should be guided by such scenarios in which optimization measures mature from their identification through evaluation to implementation.

Therefore, we initiated our prototype development with the derivation of example optimization processes (Figure 4). These process descriptions were then used to deduce different user scenarios, each representing one step within the exemplary optimization process. Such a user scenario

describes the sequence of working tasks from the perspective of a specific expert user involved in the optimization processes (e.g., product cost controller, purchaser, engineer).

Such user scenarios are designed in a way that they consider aspects of the previously evaluated requirements. This is important because we used these scenarios for a focus group with five UI experts from our research partner SAP SE who were not familiar with our research domain of product cost optimization. Since the UI experts were not familiar with the industry processes, we were able to validate and adjust our user scenarios in terms of comprehensibility. After the establishment of a common understanding, it was the task of the UI experts to transfer the user scenarios independently into UI drafts without further assistance. This was done using a paper-based approach.

Afterwards, each UI expert presented his UI drafts for each user scenario to the whole group. During this presentation, each focus group participant was encouraged to provide feedback for the different UI drafts. This feedback was helpful for us to then transfer the various paper-based designs into a digital "best-of-breed" UI prototype combining the most valuable concepts. This prototype was created with Balsamiq mockup software [26], which allows individual UI screens (and their elements) to be linked into a "clickable UI prototype."

As the last step of this development approach, we internally evaluated the clickable UI prototype first against our user scenarios and example processes, and later against the evaluated requirements in detail. This iteration allowed us to adjust the prototype design and the user scenarios.

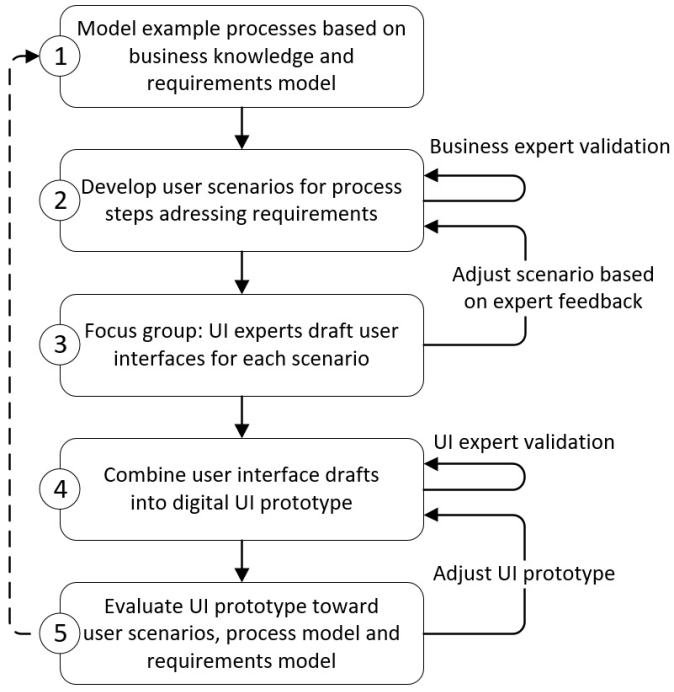


Fig. 4 Approach to transform requirements into UI prototype

This iterative approach to design a rather mature clickable UI prototype is necessary, because practitioners within the industry are usually limited toward their buy-in into research activities (especially timewise) [27]. For this reason, it makes sense to integrate these important knowledge resources in less granular and well-prepared research increments, being feasible for practice-oriented evaluation to maintain collaboration motivation.

In the next section, we will discuss the single steps of the development approach in detail and support them with examples from our research project.

IV. Process Details

A. Processes Modeling (1) and User Scenario Design (2)

According to [10], the process of designing a DSR artifact always has its point of departure in the current world. This is especially true for our practical problem of lacking IS support during early product cost optimization. Therefore, we first established an understanding of relevant optimization approaches [5] and transformed these into exemplary process flows. We started with a rather informal descriptive approach, and later used model-based approaches to establish a basis for a collective understanding among the stakeholders. In detail, we modeled the exemplary process flows with BPMN 2.0 (Business Process Model and Notation) [28].

Due to the emergent process characteristics of our research domain, process modeling is a challenge (see Background). This is because of emergent process adaptions in practice that can neither be foreseen nor fully defined by us; yet, at the same time, it is immanent that there is no chance to reach full process coverage. Therefore, we decided to transform those optimization approaches into process models, which have been evaluated as the most relevant for early product cost optimization within the discrete manufacturing industry (Table 1) [5].

Table I.
TOP 5 Optimization Approaches (Rating Scale 0 - Not Important to 10 - Very Important) [5]

Optimization Approach	Avg. Rating
Make-or-buy analysis	7.83
Material price optimization	7.78
Alternative concept and product designs	7.33
Alternative production plants	7.29
Alternative reference components, assemblies, materials, and recipe ingredients	7.06

It is important to emphasize that this set of modeled exemplary processes does not claim to be complete. Nevertheless, it provides valuable insights into where and

how the support of practitioners needs to be improved with IS.

Furthermore, process modeling helped us to identify the various process stakeholders. By differentiating between different user roles and their individual tasks within the optimization process, we were able to identify each business expert's contribution to mature a cost optimization measure from a vague idea to its implementation into a cost calculation. Building on this, we derived specific user scenarios, each representing one step within the modeled optimization process. These user scenarios form a descriptive sequence of granular working tasks from the perspective of the specific user role. These user scenarios were then enriched by transforming evaluated requirements into dedicated user actions along the processual working tasks. This was done similarly to the approach of formulating user stories in agile software development to describe functional requirements [29].

User scenario: Create a new proposal for a make-or-buy analysis for a specific component in your product

The cross-functional "cost-optimization workshop" for the development of product "Pump P-100" is in progress. You have identified a target cost deviation for the component "Casing" in your current cost calculation. Together with experts from different specialties, like purchasing, engineering, and production planning, you have identified and discussed potential optimization measures. As product cost controller, Peter, it is your task to initiate the most promising optimization measure "Make-or-buy analysis for component Casing" for further evaluation.

To do so, the following working steps must be accomplished:

- Create the measure for component "Casing" in your costing structure.
- Create an achievement plan for the measure. The cost calculation of "Pump P-100" is part of a customer quotation and, therefore, must be done by 2017-11-30.
- In order to reach the target costs, the measure must achieve savings of 300€. This targeted impact was agreed on during the cost optimization workshop. Maintain this targeted impact for the measure to enable further evaluations.
- Assign the responsibility for next evaluation steps to user Joe from purchasing department.

Fig. 5 User scenario to create a new optimization measure

Figure 5 shows a rather simple scenario from the beginning of the optimization process for a measure of type "make-or-buy analysis" (Table 1) that is being used in our prototype development. The scenario is written from the perspective of the product cost controller – who is responsible for the management of cost optimization measures at most companies [4] – and addresses requirements from our evaluated requirements model in [1]. For example, we approach the requirement toward a functionality that supports "Target costing" during the identification of an optimization measure ("Identification," Figure 3). In addition, we address requirements from the area of measure management ("Measure Management," Figure 3). In detail, this is the basic requirement of a centralized platform to manage the optimization measure among cost calculation projects, the assignment of responsibilities, the ability to create achievement plans, and the need to link the measure to components in existing cost calculations. To prioritize requirements toward our prototype, the requirements were not only validated toward their general relevance, but were prioritized on a scale from 0 (not important) to 10 (very important). Table 2 shows selected requirements and their evaluation results.

TABLE II.
EXCERPT OF EVALUATED REQUIREMENTS FROM REQUIREMENTS MODEL [1]

	Avg. Rating	Std. Dev.
Measure Management		
Collect cost-optimization measures	7.78	1.72
Select measures for […], and cost items	8.06	1.43
Define responsibilities	6.78	2.44
Create achievement plans	7.06	2.29
Estimate measure impact	8.29	1.45
Measure Identification		
Target costing	8.72	1.45

The user scenario in Figure 5 implies that there is an IS for performing early product cost calculation in place which our prototype can integrate with (see Background). As identified in previous research, this is not always the case in practice [4]. Nonetheless, the elaborated implementation challenges in previous research demand an integrated approach to ensure artifact acceptance by the end-users [1]. Moreover, this scenario implication seems to be valid because such IS are available on the market [12], although their functionality cannot cover all requirements among the discrete manufacturing industry [5].

Moreover, we introduced avatars for each user role (Figure 5) in our optimization processes. Through this idea adopted from a gamification concept for agile software development [30], we aim at making the descriptive, mostly text-based user scenarios, more attractive to people being involved in our prototype development and evaluation process (e.g., UI experts or evaluating domain experts). In addition, this should improve transparency about available user roles and their dedicated tasks in practical environments. These avatars remained consistent among the different optimization scenarios and the prototype development iterations (Table 1).

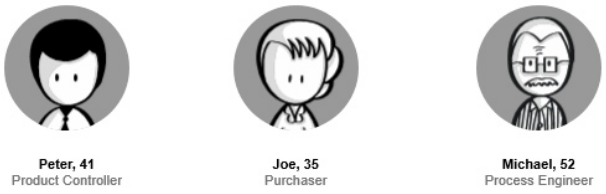


Fig. 6 Avatars that were used for the make-or-buy optimization process

B. Focus Groups with UI Experts (3)

The elaborated user scenarios build the foundation for the next step in our prototype development process. In contrast to software engineering practice where mockups are being used to enrich descriptive requirements [31], our research focuses on the derivation of design principles to improve IS support for emergent cost-optimization processes. Therefore, we decided to use the process of transforming requirements into visual mockups as an initial step to identify possible success criteria and design principles to support ex-post evaluations [32].

Due to our research collaboration with SAP SE, we had the chance to involve experts for user interfaces design in business software from SAP Innovation Center Network [33]. In a focus group involving five senior experts, who have focused on UI conception and development, with more than 7 years of work experience each, we firstly introduced the research topic driven by our problem identification [4]. Afterwards, we introduced the research approach (Figure 4) and stated the objective to create UI drafts aiming at the task fulfillment described in the set of user scenarios. As the last step of our introduction, we outlined the selected optimization process based on our BPMN models with the support of user role avatars.

Since the role of creativity to suggest solutions was highlighted in [9], we chose a pure paper-based approach to initially draft UI proposals. Nevertheless, this deliberation was not made without intention: Research has shown that paper-based approaches have certain advantages over digital prototyping approaches. Paper-based prototypes are particularly preferable when different design solutions need to be negotiated and stakeholder feedback is considered important [34]. Since we planned to have open feedback discussions among experts and, moreover, wanted to combine different solutions into a "best-of-breed" UI prototype later in the process, the paper-based approach seemed more valuable from a research perspective.

To create the UI drafts, we iterated through the optimization process based on the user scenarios as an iteration increment. The user scenario was presented to the focus group following the opportunity to clarify questions regarding its comprehensibility. By challenging the comprehensibility with the experts, we were able to further enhance the user scenarios. After reaching a collective understanding of the user scenario, the experts were asked to draft the UI individually. To underpin our claim to exemplary action, selected results of this focus group session are presented in Figure 7.

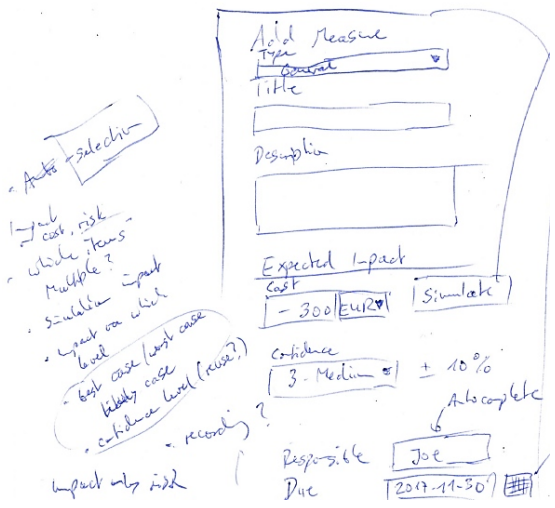


Fig. 7 Paper-based draft implementing user scenario from Figure 5

At the end of each iteration cycle, each focus group participant presented his UI draft for the specific user scenario, supported by an argumentation for the chosen UI concept. Afterwards, focus group members could provide feedback and clarify open questions. In total, we derived 47 individual, paper-based UI screen proposals for the optimization process of type "make-or-buy analysis" distributed over two separate sessions with a total duration of five working hours. The challenge in the following development step is to combine the various drafts to an evaluable prototype.

C. Development (4) and Validation (5) of a UI prototype

To transform the various paper-based UI drafts into a UI prototype that is suitable for evaluation with industry experts, we combined the most promising (also, in regard to UI expert feedback during the focus group session) drafts into digital mockups. As mentioned earlier (see Development Approach), we used the dedicated mockup software Balsamiq [26] to create a digital UI prototype.

Such a digital UI prototype consists of multiple mockups, each representing a certain UI screen (Figure 8). The major advantage of such a digital prototype over a paper-based prototype is the ability to link the different mockups (UI screens and their elements, such as buttons or text fields) to each other. The result is a UI prototype that a user can click through similarly to a real, implemented prototype.

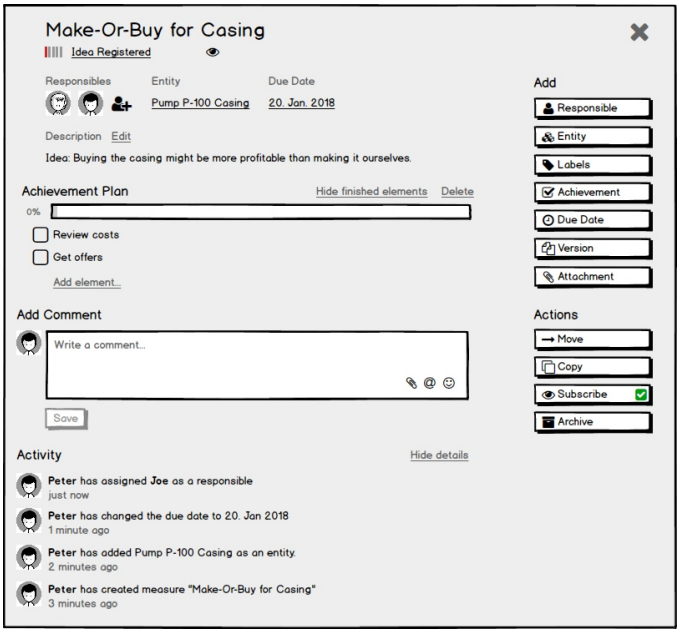


Fig. 8 UI mockup for user scenario from Figure 5

As was to be expected, not all expert UI drafts could be combined with each other. This was not due to specific UI elements or their arrangement in screens, but rather in terms of different conceptual approaches to address user scenarios. With the help of the digital UI prototype, we can easily exchange a series of UI screens to provide alternative concepts for parallel evaluation. This is extremely helpful to simulate different approaches toward process support (see *Background*).

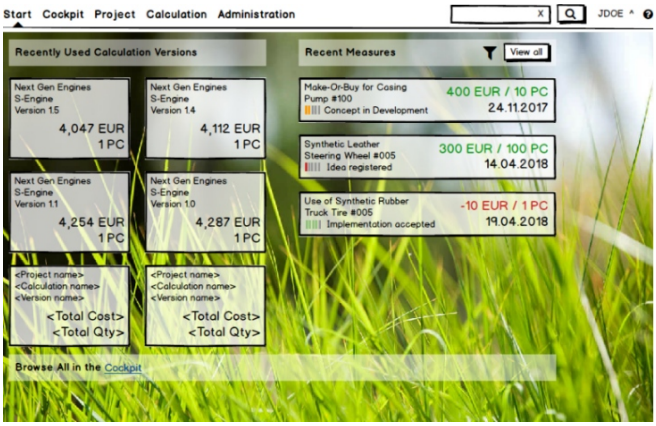


Fig. 9 Mockup of SAP Product Lifecycle's homescreen extended by functionality toward early product cost optimization

To stress the argumentation of [10] once more (see *Processes Modeling (1) and User Scenario Design (2)*) and, thus, enable an evaluation in institutional environments, we designed the clickable UI prototype to integrate with mockups of SAP Product Lifecycle Costing (Figure 9) [35]. This software is a dedicated solution to support early product cost calculation (see *Background*) and is one possible software solution available on the market that could be integrated with our UI prototype [12]. At the same time, this visual integration was necessary to address one of our elaborated implementation challenges demanding an integrated approach for optimization measure management [1].

The further development of this clickable UI prototype is linked to its internal evaluation. This internal evaluation uses three previously elaborated elements for evaluation: process models, user scenarios, and the requirements model, including a requirement description (Figure 10). Initially, we verified that the prototype addressed the tasks outlined in the user scenarios. Furthermore, we checked whether the prototype implemented by the user scenarios still corresponded to the requirements description in the validated requirements model. Finally, the prototype's click sequence needed to support the order of the tasks according to the process models.

Define responsibilities

Due to highly interdisciplinary activities, measures' responsibilities, such as functional organizations or individual task owners, must be distinctly assigned to each measure. It ensures clear organizational responsibilities for the cross-functional optimization processes.

Create achievement plans

Due to time pressure during product development, product cost optimization must be accomplished in time. Therefore, measures must contain either due dates or product development-related gates, enabling ongoing reporting.

Fig. 10 Exemplary descriptions for the evaluated requirements (Table 2) [1]

After the successful internal evaluation, we completed the development of the UI prototype that is intended for further evaluation by domain experts. This evaluation is planned in two parts: First, we let domain experts, such as consultants, product owners, and solution owners for product costing solutions, from our research partner SAP SE do the evaluation. Based on their feedback, we iterate through the outlined development process (Figure 4) once more. Business experts from the industry will then take over further evaluation.

V. Lessons Learned

In addition to the outline of our approach to transform requirements into a UI prototype, we want to contribute lessons learned from the application of this approach for our research project. Although the approach with its examples from our research project seems easy to implement, there are still some hurdles to overcome. We want to show these in the following, as they can also be relevant for other researchers in similar research contexts.

During the focus group sessions with UI experts (see Focus Groups with UI Experts (3)) we underestimated the duration of the iterations per user scenario. Looking back,

the time required for each scenario can be estimated to be one hour. Contrary to our expectations, the initial scenario explanation with subsequent questions toward comprehensibility was very time-consuming. This was mainly due to the UI expert's lack of knowledge in the specific research domain of early product cost optimization and its application in practice. Due to the unexpected delay, we had to schedule a second appointment with the UI experts to finish the draft of the optimization process of type "make-or-buy" (see Processes Modeling (1) and User Scenario Design (2)). Due to time constraints, we had to consult another UI expert and, thus, employed initial ramp up efforts once more. In order to counter such timing problems early, sufficient time should be planned for the appointments.

From our perspective, there is a chance to speed up the drafting process. Our assumption of providing the user scenarios for each step of the process in terms of process comprehensibility led to a certain redundancy across several user scenarios (e.g., different user roles were accessing the collection of optimization measures in a similar way). By removing such redundant scenarios and, moreover, focusing on less self-explaining scenarios (e.g., the cockpit screen, Figure 9), we could have been more efficient without losing relevant findings for our research problem (see Background).

For the next prototype development, we aim at a focus group approach that further enhances the creativity of involved UI experts. Though our idea of a paper-based approach was appropriate in our research context to develop potential approaches solving identified problems, we strongly believe that user scenarios provide too much guidance to elaborate ground-breaking findings. Nonetheless, the approach to draw paper-based prototypes is less time-consuming than drafting digital prototypes with dedicated software (See Development (4) and Validation (5) of a UI prototype).

In general, the approach to aim at such "clickable UI prototype" for evaluation is valuable for us. Instead of the need to implement a full-stack prototype, which, in addition, could have been limited by the extensibility options of available product costing software (see Processes Modeling (1) and User Scenario Design (2)), we could efficiently iterate through the development process to further enhance the prototype. This would simplify the testing and evaluation of several solutions in parallel. It would also motivate stakeholders to participate in the research (e.g., with new ideas or concepts), as development and its output would be progressing rapidly.

In addition, it should be mentioned that difficulties may arise when deriving and interpreting the process models and their user scenarios from the industry context into the research context. Therefore, we highly recommend evaluating the single process step results (Figure 4), like process models or user scenarios with domain experts to prevent the derivation of (partly) incorrect scenarios. To easily communicate and evaluate such results, model-based approaches like BPMN 2.0 (see Processes Modeling (1) and User Scenario Design (2)) should be used to establish a collective understanding for all stakeholders.

Furthermore, it must be ensured that the transfer of research domain content to the UI experts has been successful. Though we thoroughly introduced our research domain, including the core of our problem identification [4], [5] and the approach to derive a UI prototype (Figure 4), multiple questions for each scenario were raised. This was time consuming, but necessary for a collective understanding. The transfer from paper-based UI drafts and their underlying conceptual ideas to the digital UI prototype involves similar transition difficulties, which we recommend verifying with UI expert consultations, at least briefly, after the mockups have been created.

VI. Discussion

First, the question must be asked regarding whether the outlined approach is a valid contribution to practice-oriented DSR since we utilize a variety of well-established elements and approaches from software engineering (e.g., paper-based prototypes or mockup-driven development). According to [14], the development of tentative designs to solve identified problems is a rather pedestrian process in which no novelty beyond the state-of-art is required. Rather, the novelty should be part of the solution design itself. Following this argumentation, we do not state the UI prototype development approach as our research artifact, but as a valuable contribution to the research community.

This is especially true because, for example, the rare process guidance for artifact design has been criticized [9]. In detail, recent literature reviews have indicated only a small amount of DSR dealing with the development of software-based artifacts [8]. This is underpinned by [10], who requested methods to enable IS researchers to become immersed into institutional environments. As our approach results in UI prototypes for evaluation in practical contexts, we are convinced that it is a valuable and engaging approach toward practice-oriented DSR for all stakeholders.

What is indeed beneficial for our contextual prototyping are insights from works on action design research (ADR) as proposed in the context of DSR [18]. The interference with ADR and, therefore, the strong practical context opens room for discussions about the methodological rigor of our proposal to design a potential solution for further evaluation in the context of our problem to improve IS support for early product cost optimization. As argued by [36], there are certain conflicts in the discipline of IS research when it comes to the influence of methodical rigor. With this

contribution, we want to respond to and later solve a practical problem, which IS routine problem solving [37] has failed to address for many decades [3], [4], [5], [23], and now must be tackled by IS research.

Though we have not yet demonstrated the capability of the results to solve our identified problem [4], [5] as recommended by [6], the sole idea of designing an evaluable UI prototype has advanced our research project. The iterative validations and adjustments during the execution helped us to sharpen the understanding of the research; furthermore, the approach enabled new and improved existing collaborations with research stakeholders. According to [38], participation has the potential to enrich descriptions of the research process as well as increase the understanding of the artifact and its instantiation. Though the participation of experts in our approach led to unexpected efforts (see Lessons Learned), we can confirm and highly recommend a broad participation – especially if it concerns participants from other disciplines. Hence, the presented approach provides thought-provoking impulses and prevents tunnel vision among researchers.

This links to the contextual nature of our UI prototype for future evaluation. Earlier in this paper, we highlighted the emergent character of product cost optimization processes and the need to provide a context-integrative solution (see Background). The integration into the context of an industry application was possible and, thus, addresses one major implementation challenge to improve IS support in early product cost optimization [1]. Moreover, the evaluation of such UI prototype in practical contexts has already been shown in other research projects [22]. Therefore, we are convinced that the presented approach can contribute to finding answers to the research questions associated with our long-term research project. In addition, the concept to design the prototype iteratively improves result quality and, hence, helps to better use the already difficult-to-reach experts within the industry.

However, researchers who want to adapt this approach for their research must bear in mind that some hurdles may arise, especially regarding the transitions between the individual steps of the approach (Figure 4):

- Derivation of misguiding processes and user scenarios
- Knowledge transfer to UI experts regarding the research domain and the objective to think beyond boundaries
- Transformation of UI drafts back into the research domain-oriented context

To avoid such issues, validation and participation are key contributions to successfully derive a ready-to-evaluate UI prototype.

VII. Conclusion

Overall, we outlined our approach to develop a clickable UI prototype based on an evaluated requirements model with 30 individual requirements in the context of our long-term research project. This research project follows a DSR process recommendation [6] to improve early product cost optimization in the discrete manufacturing industry. In conjunction with this practical problem and its emergent character [1], we need to design and propose a IS solution that can be evaluated in the context of institutional environments as part of an iterative evaluation. This evaluation is part of the next research step in our long-term project, and has started with the UI prototype derived out of the outlined approach.

The approach of transforming requirements into a UI prototype has proven its value for us: Well-established techniques and state-of-the-art approaches from the software engineering discipline helped to improve research quality and strengthen relations with stakeholders with reasonable development efforts (compared to software-based prototypes). The collaboration with and the contribution of UI experts to our research were especially appreciated.

Moreover, we address the evident need within the literature to provide further guidance for the design phase of DSR. To provide meaningful guidance to IS researchers, the presented approach is enriched with examples and increments from our research project. In addition, the *Lessons Learned* section should help to adopt the process more easily. By this, we hope to further encourage researchers in practice-oriented research, and, at the same time, motivate experts among industry to join scientific research projects to solve relevant problems.

Applying the recommendation from [6] to our research project, we further concentrate on the improvement of our UI prototype and its ability to solve the problem identified in [4] and [5]. This prototype improvement is mainly driven by iterative evaluations of experts from the industry.

Acknowledgment

User role avatars were created with Scenes™ by SAP AppHaus (http://experience.sap.com/designservices/scenes).

References

[1] M. Walter, C. Leyh, and S. Strahringer, “Toward early product cost optimization: requirements for an integrated measure management approach,” in *Proc. of the Multikonferenz Wirtschaftsinformatik 2018 (MKWI 2018)*. Lueneburg, 2018, pp. 2057–2068.

[2] I. Roda and M. Garetti, “TCO evaluation in physical asset management: benefits and limitations for industrial adoption,” in *Proc. on APMS 2014: Advances in Production Management Systems*, B. Grabot, B. Vallespir, S. Gomes, A. Bouras, and D. Kiritsis, Eds. Berlin: Springer Berlin Heidelberg, 2014, pp. 216–223, https://doi.org/10.1007/978-3-662-44733-8_27.

[3] G. Schicker, F. Mader, and F. Bodendorf, “Product lifecycle cost management (PLCM): Status quo, Trends und Entwicklungsperspektiven im PLCM – eine empirische Studie,” Arbeitspapier Wirtschaftsinformatik II (2/2008), Nürnberg: Universität Erlangen-Nürnberg, 2008.

[4] M. Walter and C. Leyh, "Knocking on industry's door: product cost optimization in the early stages requires better software support," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, IEEE: Thessaloniki, 2017, pp. 330–338, https://doi.org/10.1109/cbi.2017.33.

[5] M. Walter, C. Leyh, and S. Strahringer, "Knocking on industry's door: needs in product-cost optimization in the early product life cycle stages," *Complex Systems Informatics and Modeling Quarterly (CSIMQ)*, Issue 13, pp. 43–60, 2017, https://doi.org/10.7250/csimq.2017-13.03.

[6] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge, "The design science research process: a model for producing and presenting information systems research," in *Proc. of the 1st International Conference on Design Science in Information Systems and Technology*, Claremont, 2006, pp. 83-106.

[7] R. Baskerville, "What design science is not," *European Journal of Information Systems*, vol. 17, no. 5, pp. 441–443, 2008, https://doi.org/10.1057/ejis.2008.45.

[8] R. Thakurta, B. Müller, F. Ahlemann, and D. Hoffmann, "The state of design – a comprehensive literature review to chart the design science research discourse," in *Proc. of the 50th Hawaii International Conference on System Sciences (HICSS)*. Waikoloa Village, Hawaii, 2017, pp. 4685–4694, https://doi.org/10.24251/hicss.2017.571.

[9] P. Offermann, O. Levina, M. Schönherr, and U. Bub, "Outline of a design science research process", in *Proc. of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST '09)*, New York, NY: ACM, 2009, pp. 1–11, https://doi.org/10.1145/1555619.1555629.

[10] K. Riemer and S. Seidel, "Design and design research as contextual practice," *Information Systems and e-Business Management*, vol. 11, no. 3, pp. 331–334, 2013, https://doi.org/10.1007/s10257-013-0223-2.

[11] M. Eigner and R. Stelzer, *Product Lifecycle Management: Ein Leitfaden für Product Development und Life Cycle Management*, 2nd ed. Heidelberg: Springer, 2009, https://doi.org/10.1007/b93672.

[12] S. Voelker, M. Walter, and T. Munkelt, "Improving product life-cycle cost management by the application of recommender systems," in *Proc. of the Multikonferenz Wirtschaftsinformatik 2018 (MKWI 2018)*. Lueneburg, 2018, pp. 2019–2030.

[13] R. Braun, M. Benedict, H. Wendler, and W. Esswein, "Proposal for requirements driven design science research," in *Proc. of the 10th International Conference on Design Science Research in Information Systems (DESRIST)*, B. Donnellan, M. Helfert, J. Kenneally, D. VanderMeer, M. Rothenberger, and R. Winter, Eds. Cham: Springer International Publishing, 2015, pp. 135–151, https://doi.org/10.1007/978-3-319-18714-3_9.

[14] V. K. Vaishnavi and W. Kuechler, *Design Science Research Methods and Patterns: Innovating Information and Communication Technology*, 2nd ed. Boca Raton, FL: CRC Press, 2015, https://doi.org/10.1201/b18448.

[15] M. L. Markus, A. Majchrzak, and L. Gasser, "A design theory for systems that support emergent knowledge processes," *MIS Quarterly*, vol. 26, no. 3, pp. 179–212, 2002.

[16] M. Böhringer, "Emergent case management for ad-hoc processes: a solution based on microblogging and activity streams," in *Proc. of BPM 2010: International Conference on Business Process Management*, M. Zur Muehlen and J. Su, Eds. Berlin: Springer Berlin Heidelberg, 2010, pp. 384–395, https://doi.org/10.1007/978-3-642-20511-8_36.

[17] R. Cole, S. Purao, M. Rossi, and M. K. Sein, "Being proactive: where action research meets design research," in *Proc. of 24th International Conference on Information Systems*, D. Avison, D. Galletta, and J. I. DeGross, Eds. Las Vegas, 2005, pp. 325–336.

[18] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, "Action design research," *MIS Quarterly*, vol. 35, no. 1, pp. 37–56, 2011, https://doi.org/10.2307/23043488.

[19] K. A. Piirainen and R. A. Gonzalez, "Seeking constructive synergy: design science and the constructive research approach," in *Proc. of the 8th international conference on Design Science at the Intersection of Physical and Virtual Design*, J. vom Brocke, R. Hekkala, S. Ram, and M. Rossi, Eds. Berlin: Springer Berlin Heidelberg, 2013, pp. 59–72, https://doi.org/10.1007/978-3-642-38827-9_5.

[20] A. Pranam, Product Management Essentials: Tools and Techniques for Becoming an Effective Technical Product Manager. Berkeley, CA: Apress, 2018, https://doi.org/10.1007/978-1-4842-3303-0.

[21] J. M. Rivero, J. Grigera, G. Rossi, E. R. Luna, F. Montero, and M. Gaedke, "Mockup-driven development: providing agile support for model-driven web engineering," *Information and Software Technology*, vol. 56, no.6, pp. 670–687, 2014, https://doi.org/10.1016/j.infsof.2014.01.011.

[22] D. Lück and C. Leyh, "Enabling business domain-specific e-collaboration: developing artifacts to integrate e-collaboration into product costing," in *Proc. of the 12th International Conference on Design Science Research in Information Systems*, A. Maedche, J. vom Brocke, and A. Hevner, Eds. Cham: Springer, 2017, pp. 296–312. Springer, Cham, https://doi.org/10.1007/978-3-319-59144-5_18.

[23] D. Lück and C. Leyh, "Integrated virtual cooperation in product costing in the discrete manufacturing industry: a problem identification," in *Proc. of the Multikonferenz Wirtschaftsinformatik 2016 (MKWI 2016)*. Ilmenau, 2016, pp. 279–290.

[24] M. Mörtl and C. Schmied, "Design for cost - a review of methods, tools and research directions," *Journal of the Indian Institute of Science*, vol. 95, no. 4, pp. 379–404, 2015.

[25] M. Rosemann and I. Vessey, "Toward improving the relevance of information systems research to practice: the role of applicability checks," *MIS Quarterly*, vol. 32, no. 1, pp. 1–22, 2008, https://doi.org/10.2307/25148826.

[26] Balsamiq Studios, LLC. https://balsamiq.com/, retrieved 23rd March 2018.

[27] H. Österle and B. Otto, "Consortium research," Business & Information Systems Engineering, vol. 2, pp. 283–293, 2010, https://doi.org/10.1007/s12599-010-0119-3.

[28] M. Chinosi and A. Trombetta, "BPMN: An introduction to the standard," *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, 2012, https://doi.org/10.1016/j.csi.2011.06.002.

[29] M. Cohn, *User stories applied: For agile software development*. Boston, MA: Addison-Wesley Professional, 2004.

[30] P. Lombriser, F. Dalpiaz, G. Lucassen, and S. Brinkkemper, "Gamified requirements engineering: model and experimentation," in *Proc. of the 22nd International Working Conference on Requirements Engineering*, M. Daneva and O. Pastor, Eds. Cham; Springer, 2016, pp. 171–187, https://doi.org/10.1007/978-3-319-30282-9_12.

[31] G. Reggio, M. Leotta, and R. Ricca, "A method for requirements capture and specification based on disciplined use cases and screen mockups," in *Proc. of the 16th International Conference on Product-Focused Software Process Improvement*, P. Abrahamsson, L. Corral, M. Oivo, and B. Russo, Eds. Cham: Springer International Publishing, 2015, pp. 105–113, https://doi.org/10.1007/978-3-319-26844-6_8.

[32] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: a framework for evaluation in design science research," *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016, https://doi.org/10.1057/ejis.2014.36.

[33] SAP Innovation Center Network, SAP SE. https://icn.sap.com/home.html, retrieved 26th March 2018.

[34] R. Sefelin, M. Tscheligi, and V. Giller, "Paper prototyping - what is it good for?: A comparison of paper- and computer-based low-fidelity prototyping," in *Proc. of CHI'03: Conference on Human Factors in Computing Systems*. Ft. Lauderdale, FL: ACM, 2003, pp. 778–779, https://doi.org/10.1145/765985.765986.

[35] SAP Product Lifecycle Costing, SAP SE. https://www.sap.com/products/product-lifecycle-costing.html, retrieved 23rd March 2018.

[36] R. B. Gallupe, "The tyranny of methodologies in information systems research," SIGMIS Database, vol. 38, no. 3, pp. 20–28, 2007, https://doi.org/10.1145/1278253.1278258.

[37] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013, https://doi.org/10.25300/misq/2013/37.2.01.

[38] S. Jönsson and K. Lukka, "There and back again: Doing interventionist research in management accounting," in *Handbook of Management Accounting Research*, C. S. Chapman, A. G. Hopwood, and H. G. Shields, Eds. Elsevier, 2006, pp. 373–397, https://doi.org/10.1016/s1751-3243(06)01015-7.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 725–733
DOI: 10.15439/2018F116
ISSN 2300-5963 ACSIS, Vol. 15

The ICT Adoption in Government Units in the Context of the Sustainable Information Society

Ewa Ziemba
Faculty of Finance and Insurance
University of Economics in Katowice
1 Maja 50, 40-287 Katowice, Poland
ewa.ziemba@ue.katowice.pl

***Abstract*— This study aims to advance information society research and practice by examining and understanding the information and communication technologies (ICT) adoption in government units in the context of the sustainable information society (SIS). A quantitative approach was employed to identify the levels of ICT adoption and sustainability in government units as well as investigate the correlation between the ICT adoption and sustainability. The survey questionnaires were used and data collected from Polish government units were analyzed. The research findings reveal that there are significant statistical differences between the lowest level of information culture and the highest one, namely the levels of ICT management and ICT outlay. Moreover, such differences are also identified between the lowest level of ecological sustainability and higher levels of economic, socio-cultural, and political sustainability. Finally, it is investigated that the ICT quality and information culture have a significant impact on sustainability in government units, whereas the ICT outlay and ICT management do not have such an impact.**

I. Introduction

The sustainable information society (SIS) is a new phase of information society development in which information and communication technologies (ICT) are becoming key enablers of sustainability [1]-[10]. Some areas, where the information society, sustainable development, and ICT come together, are identified and described in many studies [11-17]. All in all, the SIS is a multidimensional concept comprising environmental, economic, cultural, social, and political issues. Society stakeholders, mainly within government units, households, and public administration, could strongly affect the above issues by adopting ICT [10].

In general terms, ICT potential can be approached from two angles: ICT as an industry and ICT as a tool, and viewed as an inherent aspect for the SIS development [10]. As an industry, ICT have become a major economic driver in the hardware, software, telecommunications, and consulting services sectors. ICT as a tool can be used to transform and improve business, everyday life of people, and public governance.

ICT used as a tool to revolutionize public administration is examined in this study. Some researchers identified ICT as one of the most important tools in building sustainable government practices. ICT are expected to have a significant impact on enhancing public information and service provided for all SIS stakeholders and improving the liability and functioning of the government units [18], as well as the transparency and responsiveness of government units [19]. It is contended that ICT enable government units to improve productivity, support innovation, reduce costs, increase the effectiveness of services, and enhance the efficiency of government decision-making [7], [14], [17], [20]. Moreover, the ICT adoption in government units can yield benefits in environmental preservation by increasing energy efficiency and equipment utilization [4] as well as it can influence social development by making information and services available to all stakeholders at a faster rate [7]. All these possibilities make ICT enablers of sustainability in several respects, i.e. environmental protection (ecological sustainability), economic growth (economic sustainability), socio-cultural development (socio-cultural sustainability), and governance (political sustainability) [7], [10].

Ziemba [21], [22], [23] merely proposed a SIS model composed of the ICT adoption and sustainability, and evaluated the contribution of ICT adoption to sustainability in Polish enterprises. Following an extensive review of the literature, it can be stated that it did not uncover any deep studies identifying levels of ICT adoption and sustainability in government units nor there were any findings that interpret how the ICT adoption in government units improves sustainability. Moreover, there is a lack of research on the SIS and correlations between the ICT adoption and sustainability in less developed European countries, which are called transition economies, i.e. the former Eastern Bloc countries, which, since the early 1990s, have been undergoing transition from the command economy model to the free market model [24].

Thus, the motivation for the research is to address this under-investigated topic. The focus of this research is therefore to explore the ICT adoption and sustainability in government units. Its purposes are to identify the levels of ICT adoption and sustainability in government units, and to investigate the impact of ICT adoption on improving sustainability.

The paper is organized as follows. The next section reviews the theoretical foundation for this work and poses research questions. Subsequently, the employed research methods are discussed. The later sections describe the research findings and conclusions.

II. Theoretical Background And Research Question

A. ICT adoption

ICT are defined as a diverse set of software and hardware, to perform together various functions of information creation, storing, processing, preservation and delivery, in a growing diversity of ways [25]. They cover computers, the Internet, and mobile technologies, and mainly applications that can be used to support government units' performance and their relations with the SIS stakeholders. Based on Ross's & Vitale's study [26], the adoption of ICT is defined as ICT design, implementation, stabilization, and continuous improvement. It embraces the whole spectrum of activities from the period when government units justify the need for adopting ICT until the period when government units experience the full potential of ICT and derive benefits from them.

Based on a stream of research, Ziemba [27] advanced a model of SIS in which the ICT adoption construct is composed of four sub-constructs: ICT outlay (Out), information culture (Cul), ICT management (Man), and ICT quality (Qua). A detailed analysis allows for specifying primary variables which can be used to measure identified sub-constructs and the ICT adoption as a whole. These variables are presented in Table 1. The sub-construct of ICT outlay includes the government units' financial capabilities and expenditure on the ICT adoption, as well as funding from the European funds. The information culture sub-construct embraces digital and socio-cultural competences of government units' employees and managers, constant improvement of these competences, personal mastery and creativity of employees, and incentive systems encouraging employees to adopt ICT. The ICT management sub-construct comprises the alignment between information society strategy and ICT adoption, top management support for ICT projects, as well as the adoption of newest management concepts and standard ICT solutions developed at the national level. It also includes the implementation of legal regulations associated with the ICT adoption, regulations on ICT and information security and protection. The ICT quality sub-construct consists of the quality, interoperability, and security of back- and front-office information systems, quality of hardware, maturity of e-public services, and the adoption of electronic document management system, electronic delivery box, as well as ERP and BI systems. The construct asserted that the four sub-constructs were interrelated and critical to the design of the ICT adoption in government units in the context of the SIS.

B. Sustainability

The definition of sustainable development employed throughout this paper relates to a development in which the

TABLE I.
PRIMARY VARIABLES OF ICT ADOPTION AND SUSTAINABILITY CONSTRUCTS

Primary variables of the ICT adoption construct				**Primary variables of the sustainability construct**	
Out1	Financial capabilities	**Man18**	Management concepts adoption	**Ecl1**	Sustainability in ICT
Out2	Expenditure on ICT	**Man19**	Information security regulations	**Ecl2**	Sustainability by ICT
Out3	Funding acquired from the European funds	**Man20**	ICT regulations	**Eco3**	Reducing cost
Cul4	Managers' ICT competences	**Man21**	ICT public project	**Eco4**	Developing and increasing in the number and maturity of e-public services
Cul5	Employees' ICT competences	**Man22**	Adoption of standard ICT solutions developed at the national level	**Eco5**	Increasing effective and efficient management and decision-making
Cul6	Managers' permanent education	**Man23**	Competitive ICT market	**Eco6**	Increasing efficiency and effectiveness of customer service
Cul7	Employees' permanent education	**Qua24**	ICT infrastructure quality	**Eco7**	Increasing transparency of operations and employee responsibility
Cul8	Employees' personal mastery	**Qua25**	Back-office system quality	**Eco8**	Increasing efficiency and effectiveness of work organization
Cul9	Managers' socio-cultural competences	**Qua26**	Front-office system quality	**Eco9**	Increasing satisfaction with public services
Cul10	Employees' socio-cultural competences	**Qua27**	Interoperability of back- and front-office system	**Soc10**	Improving competences
Cul11	Employees' creativity	**Qua28**	Back-office system security	**Soc11**	Improving working environment
Cul12	Incentive systems	**Qua29**	Front-office system security	**Soc12**	Increasing safety of society members
Man13	Alignment between information society strategy and ICT adoption	**Qua30**	E-public service maturity	**Soc13**	Reducing social exclusion
Man14	Supporting management models with ICT	**Qua31**	ERP system adoption	**Pol14**	Increasing e-democracy
Man15	ICT management procedure	**Qua32**	EDMS (electronic document management system) adoption	**Pol15**	Increasing and facilitating access to public services
Man16	ICT project team	**Qua33**	Adoption of an electronic delivery box	---	---
Man17	Top management support	**Qua34**	BI (business intelligence) system adoption	---	---

Source: own elaboration.

needs of present generations are met without compromising the chances of future generations to meet their own needs [28]. According to Schauer [7], sustainable development has four dimensions which are ecological, social, economic, and cultural sustainability. Fuchs's [29] conceptualization of sustainability in the information society resonates with the Schauer's approach. He examined five dimensions of sustainability, i.e. ecological sustainability (enhancement of the natural environment), technological sustainability (usability of technologies), economic sustainability (wealth for all), political sustainability (participation), and cultural sustainability (wisdom). It is therefore expected that sustainability within government units comprises four kinds of sustainability: ecological, economic, socio-culture, and political. Such an approach was verified and confirmed by Ziemba [27].

Regarding government units, the ecological sustainability (Ecl) is the ability of government units to maintain rates of renewable resource harvest, pollution creation, and non-renewable resource depletion by means of conservation and proper use of air, water, and land resources [30], [31]. Economic sustainability (Eco) means that the government units gain competitive edge, reduce costs, organize work in a better and more efficient way, increase the number and maturity of public services delivered electronically, and boost government shareholders' value by adopting sustainable practices and improved public decision-making [14], [15], [31]. Socio-cultural sustainability (Soc) is based on the socio-cultural aspects that need to be sustained, e.g. trust, common meaning, diversity, capacity for learning and capacity for self-organization [28]. It is seen as dependent on social networks, making community contributions, creating a sense of place and offering community stability and security [32]. Political sustainability (Pol) rests on the basic values of democracy and partnership relations between government units and other SIS stakeholders. It is related to government openness, transparency and responsiveness, as well as democratic public decision-making [32], [33], [34]. Table 1 presents the description of all specified sustainability sub-constructs and variables measuring them.

C. The impact of ICT adoption on sustainability

Some studies show that ICT adoption affects sustainability. Schauer [7] stated that ICT contribute to ecological, social, cultural, and economic sustainability. Hilty and others [35] asserted that ICT can facilitate sustainability by creating the kind of economic activity that harmonizes nature with human and social welfare in the long term. Johnston [36] referred to the ICT impact on sustainability, pointing out to the need for greater investment in more effective public services and public administration as well as more active promotion of 'eco-efficient' technologies and their use.

In general, ICT for ecological sustainability comprises sustainability in ICT and sustainability by ICT [4]. Sustainability in ICT is related to greater sustainability of ICT goods and services over their whole life cycle, which is achieved by limiting energy and material flows connected with them. Then, sustainability by ICT manifests itself by creating, enabling, and encouraging sustainable patterns of production and consumption by means of ICT. ICT are crucial driving force in achieving durable, harmonious and flexible economic benefits in government units [31], [37]. The adoption of ICT can improve their efficiency and effectiveness. ICT are bound to play an increasingly prominent role in enabling socio-cultural sustainability, e.g. they can increase employment, facilitate learning and work, promote culture, reduce social exclusion [38]. ICT are used to deliver public services electronically to SIS stakeholders [38], [39]. The work by Grunwald [40] pointed out that ICT can play an important role in supporting cooperation, networking and partnership relations between government units and households. Furthermore, ICT can also allow for strengthening democracy by their adoption for improving political transparency and citizen's participation in democratic decision-making [39].

D. Research questions

Motivated by these above concerns, the SIS is a multidimensional concept encompassing two constructs: the ICT adoption and sustainability, as well as correlations between them. The sustainability construct composed of ecological, economic, socio-cultural, and political sustainability can be strongly influenced by the ICT adoption that encompasses ICT outlay, information culture, ICT management, and ICT quality.

In other study Ziemba [41] assessed the quality of the two constructs by examining the construct reliability [42], convergent and discriminant validity [43], [44]. The following measures were calculated: the loadings of each item of each component, composite reliability (CR) of all sub-constructs, average variance extracted (AVE) of all sub-constructs, Cronabch's Alpha of all sub-constructs**,** correlations between the sub-constructs, the square root of AVE for each component. Overall, the results successfully established the reliability, convergent validity, and discriminant validity of the proposed constructs and their sub-constructs [41].

The present study focuses on addressing the following research question:

RQ1: What is the level of ICT adoption in Polish government units?

RQ2: What is the level of sustainability in Polish government units?

RQ3: Does the ICT adoption influence sustainability in Polish government units?

III. Research Methodology

A. Research instrument

The Likert-type instrument (a questionnaire) was developed. The task of respondents was to assess the primary variables describing:

- The four sub-constructs of the ICT adoption construct, i.e. the ICT outlay (Out), information culture (Cul), ICT management (Man), and ICT quality (Qua) (Table 1). The respondents answered the question: *Using a scale of 1 to 5, state to what extent do you agree that the following situations and phenomena result in the efficient and effective ICT adoption in your government unit?* The scale's descriptions were: 5 – strongly agree, 4 – rather agree, 3 – neither agree nor disagree, 2 – rather disagree, 1 – strongly disagree; and
- The four sub-constructs of the sustainability construct, i.e. ecological (Ecl), economic (Eco), socio-cultural (Soc), and political sustainability (Pol) (Table 1). The respondents answered the question: *Using a scale of 1 to 5, evaluate the following benefits for your government units resulting from the efficient and effective ICT adoption?* The scale's descriptions were: 5 – strongly large, 4 – rather large, 3 – neither large nor disagree, 2 – rather small, 1 – strongly small.

B. Research subjects and procedures

In April 2016, the pilot study was conducted to verify the draft of survey questionnaire. Seven experts participated in the pilot study, i.e. five researchers from an information society and business informatics, and two employees of the Silesian Centre of Information Society in Katowice (ŚCSI). ŚCSI is a government unit that is responsible for information society development in the Silesian Province in Poland. Finishing touches were put into the questionnaire, especially of a formal and technical nature. No substantive amendments were required.

The study examined government units from the Silesian Province in Poland. The region was chosen due to its continuous and creative transformations related to restructuring and reducing the role of heavy industry in the development of research and science, supporting innovation, using *know-how* and transferring new technologies, as well as increasing importance of services. In response to the changing socio-economic and technological environment intensive work on the development of the information society has been carried out in the region for several years. In the next development strategies of the information society it was and is assumed that the potential of the region, especially in the design, provision and use of advanced information and communication technologies will be increased [ŚCSI]. All this means that the results of this research can be reflected in innovative efforts to build a SIS in the region and, at the same time, constitute *a modus operandi* for other regions throughout the country and in other countries.

Selecting a sample is a fundamental element of a positivistic study [45]. A random sample was used for statistical consideration to provide representative data. A survey questionnaire was submitted to all 185 government units in the Silesian Province.

The subjects were advised that their participation in completing the survey was voluntary. At the same time, they were assured anonymity and guaranteed that their responses would be kept confidential.

C. Data collection

Having applied the Computer Assisted Web Interview and employed the ŚCSI platform, the survey questionnaire was uploaded to the website. The data were collected between 30 May 2016 and 15 July 2016. After screening the responses and excluding outliers, there was a final sample of 118 usable, correct, and complete responses. It means that 64% of all government units from the Silesian Province completed their responses fairly, in all respects. The sample ensured that the error margin for the 95% confidence interval was 5%.

Table 2 provides details about government units' size, and their participation in SKEAP project. This project was carried out by the municipal and district authorities of the Silesian Province in 2005-2008. The project's result was the Electronic Communication System for Public Administration called SEKAP [46]. It enables government units to provide e-public services at different levels of maturity to all society stakeholders. It could be presumed that these government units which participated in SEKAP more skillfully entered into the ICT adoption than those which did not.

TABLE II.
ANALYSIS OF GOVERNMENT UNITS' PROFILES (N=118)

Characteristics	Frequency	Percentage
Number of employees		
less than 50 (small)	51	43.22%
50 and above (large)	67	56.78%
SEKAP partner		
yes	91	77.12%
no	27	22.88%

Source: own elaboration.

D. Data analysis

The data were stored in Microsoft Excel format and subsequently analyzed using Statistica package and Microsoft Excel in two stages. The first stage assessed the levels of the ICT adoption and sustainability, and the second

stage examined the significance of construct correlations and provided regression analysis.

In the first stage, the descriptive statistical analysis was employed to describe the levels of the ICT adoption and sustainability in government units. The following statistics were calculated: mean, median (MDN), first quartile (Q25), third quartile (Q75), mode, variance (VAR), standard deviation (SD), coefficient of variation (CV), skewness (SK), and coefficient of kurtosis (CK). Further, the analysis of variance (Anova Kruskala-Wallisa) was used to determine if there were statistically significant differences between distributions of scores for the ICT adoption and sustainability sub-constructs. Additionally, the Pearson Chi-square test of independence was employed to determine whether there is an association between the sub-constructs of ICT adoption/sustainability, and the size/the SEKAP participation of government units (i.e. whether the sub-constructs and the size are independent or related as well as the sub-constructs and the participation of government units are independent or related).

In the second stage, the correlation and regression analysis [47] were used to estimate the correlations between a dependent variable (sustainability and its various kinds) and one or more independent variables (ICT outlay, information culture, ICT management, and ICT quality). The coefficient of determination, denoted R^2 and advanced R^2, determines the productiveness of the proposed theoretical model. Falk and Miller [47] recommended that R^2 values should be equal to or greater than 0.10 in order to be deemed adequate for the variance explained of a particular endogenous sub-construct.

IV. Research Findings

A. The level of ICT adoption in government units

In order to answer the research question *RQ1: What is the level of ICT adoption in Polish government units?*, a detailed descriptive analysis was conducted. The results are presented in Table 3.

It has been found that the average levels of ICT adoption sub-constructs ranged from 3.29 to 3.62 (on a 5-point scale from 1.00 to 5.00). Median values were in the range between 3.22 and 3.67. On average, the level of ICT management was the highest, followed by the level of ICT outlay. The level of information culture was the lowest. The levels of the ICT adoption sub-constructs were above their average levels in most government units.

TABLE III.
THE LEVELS OF ICT ADOPTION AND SUSTAINABILITY IN GOVERNMENT UNITS

Sub-constructs	Mean	Q25	MDN	Q75	VAR	SD	CV in %	SK	CK
ICT adoption sub-constructs									
Out	3.50	3.00	3.67	4.00	0.68	0.82	23.50	-0.38	-0.20
Cul	3.29	2.89	3.22	3.89	0.41	0.64	19.46	-0.27	-0.33
Man	3.62	3.27	3.59	4.09	0.29	0.54	14.99	-0.35	0.09
Qua	3.45	3.09	3.55	3.82	0.25	0.50	14.38	-0.21	-0.59
Sustainability sub-constructs									
Ecl	2.83	2.50	3.00	3.00	0.63	0.79	27.98	0.06	0.91
Eco	3.13	2.71	3.14	3.57	0.46	0.68	21.70	-0.25	0.17
Soc	3.15	2.75	3.25	3.50	0.43	0.66	20.88	-0.64	0.70
Pol	3.19	3.00	3.00	4.00	0.60	0.77	24.21	0.06	-0.17

Source: own elaboration.

The values of h-Kruskala-Wallisa H(3, N=472)=17.861 and p=0.001) and Chi-square statistic (Chi-square=10.112, df=3, p=0.018), and finally *post-hoc* analysis confirmed significant differences between the distribution of scores for the information culture and the distributions of scores for the ICT management (p=0.000) and ICT outlay (p =0.024). In addition, the Pearson Chi-square test of independence (α=0.05) allowed for indicating that there were not statistically significant relations between the size of government units and the levels of ICT outlay, information culture, ICT management, and ICT quality. Such a relation was also not confirmed between the SEKAP participation of government units and the levels of all ICT adoption sub-constructs within them.

B. The level of sustainability in government units

In order to answer the research question *RQ2: What is the level of sustainability in Polish government units?*, a detailed descriptive analysis was conducted. The results are presented in Table 3.

It has been found that the average levels of sustainability sub-constructs ranged from 2.83 to 3.19 (on a 5-point scale from 1.00 to 5.00). Median values were in the range between 3.00 and 3.25. On average, the level of political sustainability was the highest, followed by socio-cultural sustainability. The level of ecological sustainability was the lowest. The levels of economic and socio-cultural sustainability were above their average levels in most government units, whereas the levels of ecological and political – below their average levels.

The values of h-Kruskala-Wallisa (H(3, N=472)=20.256, p=0.000) and (Chi-square=43.034, df=3, p=0.000), and finally *post-hoc* analysis confirmed significant differences between the distributions of scores for ecological sustainability and the distributions of scores for economic (p=0.003), socio-cultural (p=0.001, and political (p=0.002) sustainability. In addition, the Pearson Chi-square test of independence (α=0.05) allowed for indicating that there was statistically significant relation between socio-cultural sustainability and the size of government, as well the SEKAP participation of government units. The average level of socio-cultural sustainability was higher in small government units and in those government units which were not a SEKAP partner.

C. The contribution of ICT adoption to sustainability

Table 4 shows the results of the correlations between:

- the ICT adoption sub-constructs and the sub-constructs of sustainability; and
- the ICT adoption sub-constructs and the total sustainability construct (Y).

TABLE IV.
CORRELATIONS AMONG SUB-CONSTRUCTS OF ICT ADOPTION AND TOTAL SUSTAINABILITY AND ITS SUB-CONSTRUCTS

Sub-constructs	Ecl	Eco	Soc	Pol	Y
Out	0.055 p=0.553	0.146 p=0.115	0.192	0.064 p=0.494	0.154 p=0.960
Cul	0.174 p=0.059	0.416	0.467	0.338	0.446
Man	0.157 p=0.089	0.467	0.439	0.371	0.467
Qua	0.208	0.542	0.406	0.321	0.498

Source: own elaboration.

The correlation coefficients for the sub-constructs of ICT adoption and the sub-constructs of sustainability were significantly different from zero (p=0.000 < 0.05=α), with the exception of one correlation between the ICT outlay and ecological (p=0.553), economic (p=0.115), and political (p = 0.494) sustainability, between information culture and ecological sustainability (p=0.059), as well as between the ICT management and ecological sustainability (p=0.089). In other cases there was a positive linear correlation. In addition, the three sub-constructs of the ICT adoption construct, i.e. the ICT quality, information culture, and ICT management had a significant association with the total sustainability construct (Y). Such an association was not indicated between the ICT outlay and total sustainability (p=0.960).

The correlations analysis was sought into the linear regression. In the first and second steps of building the regression model, the following results were established. For the sub-constructs of ICT outlay (p=0.125) and ICT management (p=205), p-values were higher than the accepted level of significance (α=0.05). There is not enough evidence at the 0.05 significance level to conclude that there is a linear relationship between the level of ICT outlay and the level of sustainability as well as between the level of ICT management and the level of sustainability in the examined population. Therefore, these sub-constructs were removed from the regression model and then the correct model was received (Table 5).

TABLE V.
CORRELATIONS AMONG COMPONENTS OF ICT ADOPTION AND THE TOTAL SUSTAINABILITY AND ITS COMPONENTS

Sub-constructs	Standardized coefficients		Unstandardized coefficients		t(115)	Signif. p
	Beta	Stand. error	Beta	Stand. error		
R=0.552; R^2=0.305; Adv.R^2=0.293 (F(2.115)=25.247; p<0.0000);Standard error of estimation:0.504;N=118						
Constant			0.726	0.342	2.121	0.036
Cul	0.272	0.088	0.255	0.083	3.083	0.003
Qua	0.370	0.088	0.448	0.107	4.200	0.000

Source: own elaboration.

Two sub-constructs of the ICT adoption construct, i.e. the ICT quality and information culture explained 31% of the variance in sustainability (F2.115=25.247; p<0.0000). These sub-constructs predicted sustainability significantly well. The examination of coefficients indicated that the sub-constructs had a positive significant impact on sustainability. The effect of ICT quality was stronger than of the information culture.

Then, the relationship between the ICT adoption and sustainability may be written:

$$\hat{Y} = 0.726 + 0.255*Cul + 0.448*Qua$$

where:
$\hat{Y}$p – the theoretical level of sustainability in government units, including ecological, economic, socio-cultural, and political sustainability;
Cul – the level of information culture in government units;
Qua – the level of ICT quality in government units.

Generally, the estimated model is correct and there is no reason to reject it. It allows for understanding of the ICT adoption contribution to sustainability in government units. It gives an answer to the question – whether the growth in

levels of ICT outlay, information culture, ICT management, and ICT quality in government units determines an increase in the level of sustainability of the information society. The above results successfully established the significant and positive contribution of information culture and ICT quality to sustainability. Such a contribution was not confirmed between the ICT outlay and sustainability and between ICT management and sustainability.

V. Conclusions

A. Research contribution

This work contributes to existing research on the SIS, in particular the contribution of ICT adoption to sustainability by:

- indicating and describing the level of ICT adoption in government units, especially the levels of ICT outlay, information culture, ICT management, and ICT quality;
- indicating and describing the level of sustainability in government units, especially the levels of ecological, economic, socio-cultural, and political sustainability; and
- investigating how the ICT adoption in government units, i.e. the ICT outlay, information culture, ICT management, and ICT quality contribute to sustainability comprising its four types, i.e. ecological, economic, socio-cultural, and political.

Firstly, this study indicated significant statistical differences in the level of information culture and the level of ICT management in the Polish surveyed government units. On average, ICT management was at the highest level, whereas the lowest level was specific to information culture, followed by the ICT quality. It means that government units should focus their efforts particularly on improving information culture and ICT quality. Based on the detailed analysis of primary variables, it can be pointed out that the ICT management in government units is rather the result of top-down regulations than the efficient management, e.g. through the implementation of modern management concepts already employed in business. There were also indicated no statistically significant relations between the size of government units and the levels of ICT outlay, information culture, ICT management, and ICT quality.

Secondly, the outcomes showed significant statistical differences in the level of ecological and the levels of economic, socio-cultural, and political sustainability in the Polish surveyed government units. On average, ecological sustainability was at the lowest level, whereas the highest level was specific to political sustainability. The levels of socio-cultural and economic were only minimally lower to the political sustainability. Generally, the levels of all kinds of sustainability are low. It means that government units reap insignificant ecological, economic, socio-cultural, and political benefits from adopting ICT.

Thirdly, it was indicated that the ICT quality and information culture significantly and positively contribute to sustainability in government units. However, the effect of the ICT quality was stronger than of information culture.

With regard to the presented results, it is reasonable to conclude that this study expands the existing research on the SIS provided by Schauer [7], Fuchs [1], [2], Hilty et al. [4], [5], Guillemette, Paré [14, [15], and Curry and Donnellan [12], [13] by presenting the levels of ICT adoption and sustainability as well as identifying how the ICT adoption influence sustainability. The research outputs are also complementary with findings related to the effect of ICT adoption on sustainability in enterprises [21], [22], [23] and households [48]. Summarizing up-to-date research findings, it can therefore be concluded that the ICT quality, ICT management, and information culture within enterprises and households have a significant impact on sustainability, whereas in government units only the ICT quality and information culture influence sustainability. In addition, the ICT outlay does not have any impact on sustainability both in enterprises, households, and government units. On average, the levels of ICT adoption and sustainability in government units were lower than such levels in enterprises and households.

B. Research implication for research and practice

The research findings of this study can be used by scholars to improve and expand the research on the SIS. Researchers may use the proposed methodology to do similar analyses with different sample groups in other countries, and many comparisons between different countries can be drawn. Moreover, the methodology constitutes a very comprehensive basis for identifying the levels of ICT adoption and sustainability, as well as the correlations between the two constructs, but researchers may develop, verify and improve this methodology and its implementation.

This study offers several implications for government units. They may find the results appealing and useful in enhancing the adoption of ICT, experiencing the full potential of ICT and deriving various benefits from the ICT adoption. The findings suggest some framework comprising various kinds of benefits like ecological, economic, socio-cultural, and political that can be obtained thanks to the ICT adoption. In addition, they recommend some guidelines on how to effectively and efficiently adopt ICT in order to obtain those benefits. It is evident from the findings that government units should pay utmost attention to the improvement of information culture and ICT quality. In particular, this research can be largely useful for the transition economies in Central and Eastern Europe. This is because the countries are similar with regard to analogous geopolitical situation, their joint history, traditions, culture and values, the quality of ICT infrastructure, as well as building democratic state structures and a free-market

economy, and participating in the European integration process.

All in all, the research results might provide a partial explanation to the issue of how government units can participate in the creation of sustainable development and sustainable information society.

C. Research limitations and future works

As with many other studies, this study has its limitations. First, the ICT adoption and sustainability constructs are new constructs that have yet to be further explored and exposed to repeated empirical validation. Second, the sample included Polish government units only, especially from the Silesian Province. The study sample precludes statistical generalization of the results from the Silesian government units to government units in other Polish provinces. After all, caution should be taken when generalizing the findings to other regions and countries. Finally, the research subjects were limited to government units and it is therefore only the viewpoint of government units toward the ICT adoption for achieving sustainability in the information society.

Additional research must be performed to better understand the SIS, the ICT adoption and sustainability construct, and the correlations between the ICT adoption and sustainability. First, further validation of the levels of ICT adoption and sustainability should be carried out for a larger sample comprising government units from different Polish provinces. Second, the methodology of the ICT adoption, sustainability, and SIS measurement should be explored in greater depth. A composite index for the SIS with sub-indexes of ICT adoption and sustainability in government units should be explored. In addition comparisons between government units and enterprises [21], [22], [23] may be made.

REFERENCES

[1] Ch. Fuchs, "Sustainable information society as ideology (part I)," *Informacion Tarsadalom*, vol. 9, no 2, pp. 7–19, 2009.

[2] Ch Fuchs, "Sustainable information society as ideology (part II)," *Informacion Tarsadalom*, vol. 9, no 3, pp. 27–52, 2009.

[3] Ch. Fuchs, "Theoretical foundations of defining the participatory, co-operative, sustainable information society," *Communication & Society*, vol. 13, no 1, pp. 23–47, 2010. https://doi.org/10.1080/13691180902801585

[4] L.M. Hilty and B. Aebischer, "ICT for sustainability: An emerging research field," *Advances in Intelligent Systems and Computing*, vol. 310, pp. 1–34, 2015. https://doi.org/10.1007/978-3-319-09228-7_1

[5] L.M. Hilty and M.D. Hercheui, "ICT and sustainable development," in *What kind of information society? Governance, virtuality, surveillance, sustainability, resilience*, *Proceedings of 9th IFIP TC 9 International Conference, HCC9, and 1st IFIP TC 11 International Conference,* J. Berleur, M.D. Hercheui, and L.M. Hilty, Eds. Brisbane, September 20-23, 2010, p. 227–235. https://doi.org/10.1007/978-3-642-15479-9_22

[6] J.W. Houghton, "ICT and the environment in developing countries: A review of opportunities and developments," in *What kind of information society? Governance, virtuality, surveillance, sustainability, resilience*, *Proceedings of 9th IFIP TC 9 International Conference, HCC9, and 1st IFIP TC 11 International Conference,* J. Berleur, M.D. Hercheui, and L.M. Hilty, Eds. Brisbane, September 20-23, 2010, p. 236–247. https://doi.org/10.1007/978-3-642-15479-9_23

[7] T. Schauer, *The sustainable information society – vision and risks.* Vienna: The Club of Rome – European Support Centre, 2003.

[8] J. Servaes and N. Carpentier, Eds. *Towards a sustainable information society. Deconstructing WSIS*. Portland: Intellect, 2006.

[9] E. Ziemba, "The holistic and systems approach to a sustainable information society," *Journal of Computer Information Systems*, vol. 54, no 1, pp. 106–116, 2013. https://doi.org/10.1080/08874417.2013.11645676

[10] E. Ziemba, Eds. *Towards a sustainable information society: People, business and public administration perspectives*. Newcastle upon Tyne: Cambridge Scholars Publishing, 2016.

[11] C. Avgerou, "Discourses on ICT and development," *Information Technologies and International Development*, vol. 6, no 3, pp. 1–18, 2010.

[12] E. Curry and B. Donnellan, "Understanding the maturity of sustainable ICT," in *Green business process management – Towards the sustainable enterprise*, J. vom Brocke, S. Seidel, and J. Recker, Eds. Berlin: Springer, 2012, pp. 203–216. https://doi.org/10.1007/978-3-642-27488-6_12

[13] B. Donnellan, C. Sheridan, and E. Curry, "A capability maturity framework for sustainable information and communication technology," *IT Professional*, vol. 13, no 1, pp. 33–40, 2011. https://doi.org/10.1109/MITP.2011.2

[14] M.G. Guillemette and G. Paré, "Toward a new theory of the contribution of the IT function in organizations," *MIS Q*, (36:2), 2012, pp. 529–551.

[15] M.G. Guillemette and G. Paré, "Transformation of the information technology function in organizations: A Case study in the manufacturing sector," *Canadian Journal of Administrative Sciences*, vol. 29, 2012, pp. 177–190. https://doi.org/doi:10.1002/cjas.224

[16] S. Seidel, J. Recker, and J. vom Brocke, "Sensemaking and sustainable practicing: functional affordances of information systems in green transformations," *MIS Q*, vol. 37, no 4, pp. 1275–1299, 2013.

[17] R.T. Watson, M.C. Boudreau, A.J. Chen, and M. Huber, "Green IS: Building sustainable business practices," in *Information systems*, R.T. Watson, Ed. Athens: Global Text Project, 2008, pp. 247–261.

[18] N.P. Rana, Y.K. Dwivedi, and M.D. Williams, "Analysing challenges, barriers and CSF of egov adoption," *Transforming Government: People, Process and Policy*, vol. 7, no 2, pp. 177–198, 2013. https://doi.org/10.1108/17506161311325350

[19] P. Palvia, N. Baqir, and H. Nemati, "ICT for socio-economic development: A citizens' perspective," *Information &* Management, vol.55, pp. 160–176, 2018. https://doi.org/10.1016/j.im.2017.05.003

[20] J. Jurado-González and J.L. Gómez-Barroso, "What became of the information society and development? Assessing the information society's relevance in the context of an economic crisis," *Information Technology for Development*, vol. 22, no 3, pp. 436–463, 2016. https://doi.org/10.1080/02681102.2016.1155143

[21] E. Ziemba, "The contribution of ICT adoption to the sustainable information society," *Journal of Computer Information Systems*, 2017. http://dx.doi.org/10.1080/08874417.2017.1312635

[22] E. Ziemba, "The ICT adoption in enterprises in the context of the sustainable information society," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems FedCSIS*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., Czech Technical University in Prague, Prague, September 3-6, 2017, p. 1031–1038. https://doi.org/10.15439/2017F89

[23] E. Ziemba, "Synthetic indexes for a sustainable information society: Measuring ICT adoption and sustainability in Polish enterprises," in *Information technology for management: Ongoing Research and Development*, E. Ziemba, Ed. *Lecture Notes in Business Information Processing*, vol. 311, pp. 151-169, 2018. https://doi.org/10.1007/978-3-319-77721-4_9

[24] N. Roztocki and H.R. Weistroffer, "Information and communication technology in transition economies: An assessment of research trends," *Information Technology for Development*," vol. 21, no 3, pp. 330–364, 2015. https://doi.org/10.1080/02681102.2014.891498

[25] J.H. Nord, M.T. Riggio, and J. Paliszkiewicz, "Social and economic development through information and communications technologies: Italy," *Journal of Computer Information System*, vol. 57, no (3), pp.

278–285, 2017. https://doi.org/10.1080/08874417.2016.1213621
[26] J.W. Ross and M.R. Vitale, "The ERP revolution: surviving vs thriving," *Information Systems Frontiers*, vol. 2, no 2, pp. 233–241, 2000. https://doi.org/10.1023/A:1026500224101
[27] E. Ziemba, *Zrównoważone społeczeństwo informacyjne [Sustainable information society]*, Publishing House of the University of Economics in Katowice, Katowice, 2017.
[28] M. Missimer, K.H. Robèrt, and G. Broman, "A strategic approach to social sustainability-Part 2: A principle-based definitions," *Journal of Cleaner Production*, vol. 149, no 1, pp. 42–52, 2017. https://doi.org/10.1016/j.jclepro.2016.04.059
[29] Ch. Fuchs, "Sustainability and the information society," in *Social informatics: An information society for all? In remembrance of Rob Klin*g, T. Berleur, M.I. Numinen, and T. Impagliazzo, Eds. Boston: Springer, p. 219–230, 2006. https://doi.org/10.1007/978-0-387-37876-3_18
[30] J. Nicolette, S. Burr, and M. Rockel, "A practical approach for demonstrating environmental sustainability and stewardship through a net ecosystem service analysis," *Sustainability*, vol. 5, pp. 2152–2177, 2013. https://doi.org/10.3390/su5052152
[31] T. Nyström and M.M. Mustaquim, "Finding sustainability indicators for information system assessment," in *Proceedings of the 19th International Academic Mindtrek Conference*, Tampere, Finland, September 22-24, 2015, p. 106–113. https://doi.org/10.1145/2818187.2818278
[32] R. Khan, "How frugal innovation promotes social sustainability," *Sustainability*, vol. 8, no 10, paper 1034, 2016. https://doi.org/10.3390/su8101034.
[33] V. Mani, R. Agarwal, A. Gunasekaran, T. Papadopoulos, R. Dubey, and S.J. Childe, "Social sustainability in the supply chain: Construct development and measurement validation," *Ecological Indicators,* vol. 71, pp. 270–279, 2016. https://doi.org/10.1016/j.ecolind.2016.07.007
[34] B. Ngwenya, "Realigning governance: From e-government to e-democracy for social and economic development," in *Digital solutions for contemporary democracy and governme*nt, K.J. Bwalya and S. Mutula, Eds. Hershey: IGI Global, pp. 21–45, 2015. https://doi.org/10.4018/978-1-4666-8430-0.ch002
[35] L.M. Hilty, E.K Seifert., and R. Treibert, Eds. *Information systems for sustainable development*. Idea Group Publishing, Hershey, 2005.
[36] P. Johnston, "Towards a knowledge society and sustainable development: deconstructing the WSIS in the European policy context," in *Towards a sustainable information society. Deconstructing WSIS*, J, Servaes and N. Carpentier, Eds. Portland: Intellect, pp. 203–206, 2006.
[37] R. Isenmann, "Sustainable information society," in *Encyclopedia of information ethics and security*, M. Quigley, Ed. Hershey: IGI Global, pp. 622–630, 2008.
[38] Hameed, T. *ICT as an enabler of socio-economic development*. Daejeon: Information & Communications University, 2015, http://www.itu.int/osg/spu/digitalbridges/materials/hameed-paper.pdf, (accessed: 12th June 2016).
[39] A.K. Srivastava and S. Sharma, "Social justice through Aadhaar: An e-policy initiative," in *Technology, society and sustainability. Selected concepts, issues and cases*, L.W. Zacher, Ed. Cham: Springer, pp. 83–97, 2017. https://doi.org/10.1007/978-3-319-47164-8_5
[40] A. Grunwald, "Technology assessment and policy advice in the field of sustainable development," in *Technology, society and sustainability. Selected concepts, issues and cases*, L.W. Zacher, Ed. Cham: Springer, pp. 203–221, 2017. https://doi.org/10.1007/978-3-319-47164-8_14
[41] E. Ziemba, "The contribution of ICT adoption within local government to sustainability – Empirical evidence from Poland," (in the preparation).
[42] P.R. Hinton, C. Brownlow, I. McMurvay, and B. Cozens, *SPSS Explained*. East Sussex: Routledge, 2004.
[43] D. Gefen and D. Straub, "A practical guide to factorial validity using PLS-graph: Tutorial and annotated example," *Communications of the Association for Information Systems*, vol. 16, no 5, pp. 91–109, 2005.
[44] J.F. Hair, C.L. Hollingsworth, A.B. Randolph and A.Y.L. Chong, "An updated and expanded assessment of PLS-SEM in information systems research," *Industrial Management & Data Systems*, vol. 117, no 3, pp. 442–458, 2017. https://doi.org/10.1108/IMDS-04-2016-0130
[45] D. Crowther and G. Lancaster, *Research methods: A concise introduction to research in management and business consultancy*. New York: Routledge, 2008.
[46] E. Ziemba and T. Papaj, "A pragmatic approach to e-government maturity in Poland – implementation and usage of SEKAP," in *Proceedings of 13th European Conference on eGovernment ECEG 2013*, E. Ferrari and W. Castelnovo, Eds. University of Insubria, Varese, Como, Italy, June 13-14, 2013, p. 560-570.
[47] R.F. Falk and N.B. Miller, *A primer for soft modeling*. Akron: The University of Akron Press, 1992.
[48] E. Ziemba, "The contribution of ICT adoption to sustainability: households' perspective," *Information Technology & People*, 2018. https://doi.org/10.1108/ITP-02-2018-0090

13th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

TOPICS

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles
- Information Systems for Sustainability
  - sustainable business models, financial sustainability, sustainable marketing
  - qualitative and quantitative approaches to digital sustainability
  - decision support methods for sustainable management

EVENT CHAIRS

- **Arogyaswami, Bernard,** Le Moyne University, USA
- **Chmielarz, Witold,** University of Warsaw, Poland
- **Karagiannis, Dimitris,** University of Vienna, Austria
- **Kisielnicki, Jerzy,** University of Warsaw, Poland
- **Ziemba, Ewa,** University of Economics in Katowice, Poland

PROGRAM COMMITTEE

- **Aguillar, Daniel,** Instituto de Pesquisas Tecnológicas de São Paulo, Brazil
- **Alghamdi, Saleh,** King Abdulaziz City for Science and Technology, Saudi Arabia
- **Bontchev, Boyan,** Sofia University St Kliment Ohridski, Bulgaria
- **Cingula, Domagoj,** Economic and Social Development Conference, Croatia
- **Czarnacka-Chrobot, Beata,** Warsaw School of Economics, Poland
- **Damasevicius, Robertas,** Kaunas University of Technology, Lithuania
- **Duan, Yanqing,** University of Bedfordshire, United Kingdom
- **El Emary, Ibrahim,** King Abdulaziz Univetrsity, Saudi Arabia
- **Espinosa, Susana de Juana,** University of Alicante, Spain
- **Feltus, Christophe,** Luxembourg Institute of Science and Technology, Luxembourg
- **Gaweł, Aleksandra,** Poznan University of Economics and Business, Poland
- **Geri, Nitza,** The Open University of Israel, Israel
- **Halawi, Leila,** Embry-Riddle Aeronautical University, United States
- **Jankowski, Jarosław,** West Pomeranian University of Technology in Szczecin, Poland
- **Kania, Krzysztof,** University of Economics in Katowice, Poland

- **Kobyliński, Andrzej,** Warsaw School of Economics, Poland
- **Leyh, Christian,** Technische Universität Dresden, Germany
- **Michalik, Krzysztof,** University of Economics in Katowice, Poland
- **Mullins, Roisin,** University of Wales Trinity Saint David, United Kingdom
- **Muszyńska, Karolina,** University of Szczecin, Poland
- **Nuninger, Walter,** Polytech'Lille, Université de Lille, France
- **Ohira, Shigeki,** Nagoya University, Japan
- **Popescu, Elvira,** University of Craiova, Romania
- **Queirós, Ricardo,** Escola Superior de Media Artes e Design, Politécnico do Porto, Portugal
- **Rizun, NIna,** Alfred Nobel University, Dnipropetrovs'k, Ukraine
- **Rozevskis, Uldis,** University of Latvia, Latvia
- **Schroeder, Marcin Jan,** Akita International University, Japan
- **Sobczak, Andrzej,** Warsaw School of Economics, Poland
- **Swacha, Jakub,** University of Szczecin, Poland
- **Symeonidis, Symeon,** Democritus University of Thrace, Greece
- **Szczerbicki, Edward,** University of Newcastle, Australia
- **Travica, Bob,** University of Manitoba, Canada
- **Wątróbski, Jarosław,** University of Szczecin, Poland
- **Wielki, Janusz,** Opole University of Technology, Poland
- **Žemlička, Michal,** Charles University in Prague, Czech Republic

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 737–746
DOI: 10.15439/2018F238
ISSN 2300-5963 ACSIS, Vol. 15

A Geofencing Algorithm Fit for Supply Chain Management

Paolo Walter Modica, Mark Phillip Loria and Marco Toja
See Your Box
2 Cormont Road
London, England
SE5 9RA (UK)
Email: {pmodica, mloria, mtoja}@seeyourbox.com

Vincenza Carchiolo and Michele Malgeri
Dip. Ingegneria Elettrica Elettronica e Informatica,
Universitá di Catania,
Viale Andrea Doria 6,
95125 Catania, Italy
Email: {vincenza.carchiolo, michele.malgeri}@dieei.unict.it

Index Terms—**IIoT, Geofencing, Supply chain, algorithm**

***Abstract*— Location Based Services play an important role in decision-making processes, company activities or in any control and policy system in modern computer organizations. Usually LBS applications provide location-specific information only when user requests it. However, Supply Chain Management applications require to push geolocalized information directly to the user. The most discussed and requested application is *Geofencing*, which allows to determine the topological relation between a moving object and a set of delimited geographical areas. This paper describes the design of an innovative solution for implementing proactive location-based services suitable for application scenarios with strong time constraints, such as real-time systems, called Proactive Fast and Low Resource Geofencing Algorithm.**

I. Introduction

LOCATION-BASED services (LBS), software-level functionalities that use location data to control features, have recently become a hot topic for both consumer and industrial applications, evolving from simple synchronization-based service models to authenticated and complex tools thanks to advancements in telecommunication technologies and localization services.

Nowadays, Location Based Services are *crucial for many businesses*, as well as for government organizations, as they could play an important role in decision-making processes, company activities or in any control and policy system in modern computer organizations.

The majority of applications exploiting LBSs are based on the idea to present location-specific information in case the user asks for it. A relatively small amount of new applications act *proactively*, delivering *enter*, *exit* and *cross* geo-notifications directly to the end user. The most discussed and requested proactive LBS nowadays is *Geofencing*, which allows to determine the topological relation between a moving object and a set of delimited geographical areas.

This paper describes the design of a new *geofencing algorithm*, Proactive Fast and Low Resource Geofencing Algorithm (PFLGA) proposed as an innovative solution for implementing proactive location-based services suitable for application scenarios with strong time constraints, such as real-time systems. PFLGA is then exploited in a real application scenario to implement a geofencing service within the See Your Box system, an IT company offering Business-to-Business services which allow early detection of logistic issues happening in Supply Chain Management across several industry verticals.

In particular, PFLGA aims to resolve trajectory-based topological join queries to catch the occurrence of topological events, related to the movement of mobile IoT/M2M tracking devices, within strict time constraints, in the context of real-time supporting services for Supply Chain Management.

PFLGA proposes a centralized, thin-client solution to the geofencing problem, exploiting the GeoJSON format and tree-based index structures for the representation, collection and indexing of geospatial geometric shapes. These features allow the proposed solution to face two classical challenges of Geofencing: reducing the energy consumption at the mobile device, and allowing the matching process within the centralized solution to scale [1].

Classic geofencing solutions require the use of GPS locators continuously connected and therefore in need of continuous power supply.

Several different algorithms have been proposed in the technical literature for the implementation of geofencing functionalities and this section analyzes some of the latest and most interesting designs of the recent years. One of the main problem is the *point-in-polygon problem (PIP)*, which, in computational geometry, solves the question about the position of a point with respect to the boundary of a polygon in the same plane. The paper compares PFLGA with *Parallel In-memory Spatio-temporal TOpological joiN* (PISTON) [2], *Scan-Line Algorithm and Grid Compression (SLGC-1)* [3] and *Geofencing via Hybrid Hashing* [4]. The features and performance of PFLGA are far better than that of the algorithms mentioned above, making it a good solution for the Supply Chain Management context.

Section II examines location-based services, addressing their components, their technical characteristics, focusing on geofencing and route-matching services. Section III describes

This work was funded in part by University of Catania, Dip. Ingegneria Elettrica Elettronica Informatica (DIEEI), under DEDuCE project

features and requirements of the application scenario considered in this work, which relates to *Machine-to-Machine (M2M)* data processing systems based on infrastructure-assisted mobile devices, potentially exploitable in various areas such as Supply Chain Management, then considers many existing geofencing solutions, taken from the technical literature. Section IV deeply discuss PFLGA and one of its possible implementation and provides a comparison between the algorithm mentioned above and those existing geofencing designs which better fit with the service requirements of the considered application scenario. Section IV-B provides comparison both in terms of features and performance for achieving the geofencing result. Finally, conclusions and ideas for future developments are discussed in Section V.

II. Related Works

A. Location Based Service and Geofencing

A location-based service is a software-level service that uses location data to control features [5]. LBS is a part of virtually all control and policy systems which work in computers today [6] and can be used in a variety of contexts, such as healthcare [7], entertainment, indoor object localization and work. LBSs have rapidly evolved from simple synchronization-based service models to authenticated and complex tools for implementing virtually any location-based service model or facility, becoming crucial for many businesses, as well as for government organizations, as they could play an important role in any control and policy system in modern computer organizations.

Together with the Internet of Things (IoT) paradigm, it constitutes an enabling technology for advanced Machine-to-Machine services, useful for companies across diverse industries where often the efficiency plays an important role [8].

According to the definition given by the international Open Geospatial Consortium [5] a LBS is an intersection of three technologies:

1) wireless and mobile telecommunication technologies, which play an important role for the development of many new location-based services for both business and consumer applications, providing applications with an infrastructure able to manage the communications between mobile terminals and Providers of the service [5];
2) Geographic Information Systems (GIS) [9] that provide a strong framework to build database management systems and object extensions, to store and maintain geographical records to monitor the status and changing of the world's geography, and application software such as interactive maps and analysis instruments;
3) New Information and Communication Technologies (NICT), which encompass all those technologies and smart assets, with embedded processing and communication capabilities, which enable human actors to access, store, manipulate, transmit and share information wherever and whenever they want.

LBSs differ from the common Internet services because they must be aware of the context in which they are being used and they must adapt their contents' actions accordingly [5]. LBSs must be aware of any information that could be used to describe the context, such as a place, its features, the objects and people standing in it and anything that is relevant to the interaction between an user and the LBS application. Systems that can dynamically change their behavior according to the context are defined *context-sensitive or responsive*. Service adaption can take place at different levels, from the information level, where the information provided by the service is adapted according to the context, to the user interface and presentation level.

According to literature [10] LBSs can be classified in the following two categories:

- *Pull LBSs*, which deliver information directly requested from the user/customer. Pull LBSs can be further classified [5] in functional services, which facilitate the user in the acquisition of goods and services related with his/her position, and information services, which retrieve information about a specific subject depending on the user's position;
- *Push LBSs*, which deliver information not requested or indirectly requested from the user, although the user may have originally subscribed to the service at an earlier time. Push services are activated by events and are usually more complex to establish.

In the last years the attention moved from Pull LBSs to a more advanced, proactive type of location-based services, where environmental information is pushed to the user depending on the geographical position of a mobile device [11]. The most discussed and requested Push LBS in recent time is geofencing.

In this paper we focus on *Geofencing* that is is a location-based service which enables to detect and monitor when a mobile IoT/M2M device enters, leaves, crosses or bypasses a precise geographical area delimited by a virtual perimeter, called geofence [10], [12] providing alerts or notifications, usually referred to as *geo-notifications*. A geofence can be dynamically generated, like a circular area surrounding the current position of a mobile device, or can be made of a predefined set of boundaries, which may be arbitrarily drawn by the user or specific for a place or a building. Geofencing services can be classified, depending on the geographical references used to check device's position, in [13] *static* that checks the geographical position of a mobile device with respect to a fixed area, *dynamic* that operates according to the position of a mobile device with respect to a changing area and *peer-to-peer* that uses the geographical position of a mobile device with respect to other mobile devices.

A geofencing service can be characterized according to the following features [13]:

- *location accuracy*: geofencing accuracy is strictly related to the accuracy of the geographical position provided by the service used to track the location of the mobile device, either satellite/GPS or GSM-based.
- *Tracking Rate*: expresses the frequency at which the

device provides a location update to the server of the proactive LBS.
- *Device Speed*: the speed of a device determines the time period within which the device must provide a location update to be evaluated against eligible events.
- *Device Route*: the path a device takes across a geofenced area which affects the time period within which location update must occur.
- *Geo-notification delivery*: geo-notifications can be delivered to the user only once or every time the mobile device successfully enters, leaves, crosses or bypasses a geofenced area.

The spread of location-based services applied to IoT technologies, especially for mobile-based solutions, makes it necessary to add to the features listed above the *power efficiency*, since the signal is triggered by small, battery powered mobile devices.

The behavior of a moving mobile IoT/M2M device relative to a set of one or more geofences can be easily defined using the spatial predicates enter, leave, bypass and cross, proposed for the first time by Erwig, Schneider et al. in [14]. Whether a moving object enters, leaves, crosses or bypasses a given geofenced area can be determined by examining one or more segments of its trajectory, checking if they intersect with the before mentioned geofence and evaluating the intersections found. The whole process is often referred to as trajectory-based topological join query.

Trajectory-based topological join queries are really powerful instruments for spatiotemporal analysis, but they are also rather compute-intensive. For their resolution they require a description of the route traveled by the moving object, which could be expressed as a set of segments or as a polyline. This requirement raise the issue of how the route should be determined, which will be discussed later on when it comes to route-matching service.

B. Geofencing implementation

One of the most important component in a geofencing system is the Location Monitoring Unit (LMU), which is the component inside the geofencing system infrastructure which is responsible for location processing of the positions of a mobile device and for keeping the geofence scenarios secret. Technically, a geofencing system can be implemented in two different solutions [1]: mobile-based and centralized system.

In a mobile-based geofencing system, the device positioning, determined with satellite-GPS technology, together with the matching of the position with a set of geofences is executed at the mobile device. This type of geofencing systems represents a thick client solution which is mainly used in case a trustful position of nodes is needed, although it requires high battery consumption due to the geospatial processing executed at the mobile node.

In a centralized geofencing system, a mobile device is being tracked by the surrounding network infrastructure, while the matching of the retrieved position with a set of geofences is executed by the servers which make up the geofencing system infrastructure. Centralized geofencing systems represent thin client solutions [15] and have several advantages over mobile-based counterparts, such as:
- mobile devices get rid of the CPU-intensive geospatial processing necessary to determine the current state of a mobile client regarding the geofence scenarios.
- Centralized geofencing systems use network-based positioning methods within the infrastructure, such as satellite-GPS positioning or GSM Cell ID positioning, which relieve the mobile clients from the energy-draining positioning process.
- Since the LMU monitors all the mobile clients of the system, centralized systems allow for collaborative geofencing and monitoring of the current number of clients within a particular geofence.

On the other side, in centralized geofencing systems the communication between the LMU and mobile clients increases, accompanied by all the weaknesses of mobile communications like loss of connection, unpredictable latency and an energy consumption tightly depending on the location update frequency.

C. Geofencing challenges

Geofencing is associated with two main technical challenges: reducing the energy consumption at the mobile device, in particular within the mobile-based solution, and allowing the matching process within the centralized solution to scale [1]. The high energy consumption of the mobile device is mainly caused by the positioning modules (satellite-GPS or GSM-based) nodes are equipped with and, in case of mobile-based geofencing systems, by the geospatial processing necessary to determine node status.

As for the high energy consumption caused by the positioning modules inside the device, this is tackled by selecting the positioning method based on:
- accuracy need: satellite-GPS technologies allow for a more precise localization than GSM Cell ID technology;
- current environment: in case GSM signal doesn't reach the area the mobile device is currently in, positioning technique is switched to satellite-GPS;
- current position/speed towards a geofence: in case the mobile device is far from a geofence boundary, at a distance which is greater than a specific safety radius, then no location update is needed. Otherwise, if the mobile device is close to a geofence boundary, it will transmit periodic location updates to make the system check its position relative to the geofence.

In the context of geofencing systems, scalability is required in two dimensions [16]:
1) amount of geofences set by the user;
2) number of location updates processed per time unit (throughput).

To allow the customer using the geofencing system to store as many geofences as he wants, without compromising the efficiency of the service and aiming at the maximum scalability, geospatial objects need to be indexed using spatial indexes.

The use of spatial indexes allows to reduce greatly the time needed to resolve geospatial queries, which results in more location updates processed per time unit, which increases system scalability.

D. Route-matching

Intelligent Transport Systems (ITS) and Location-Based Services (LBS) require location information about mobile IoT/M2M telemetry devices. In the last few years Global Positioning System (GPS) has established itself as the major positioning technology for providing location data. This information can be used with spatial road network data to determine the spatial reference of device location via a process known as map-matching or route-matching [17].

Route-matching techniques integrate positioning data, coming from satellite-GPS or GSM positioning technologies, with spatial road network data to provide enhanced positioning capabilities, with the aim of identifying the most plausible route segment traveled by a mobile object between two or more location points [18] and determining the device location inside the calculated road segment [17].

The quality of the result returned by a map-matching service depends on:

- the quality of the spatial road map used by the algorithm, which must always been up to date and checked in-depth in order to identify and correct flaws in the available road network data;
- devices' sampling frequency, which in turn depends on the precision requirements of the localization service and on the performance of the transceiver the mobile device is equipped with. Since the positioning technology chosen for tracking the mobile device is characterized by a known measurement error the desired accuracy for the route-matching algorithm can be achieved by adjusting the sampling rate of the tracking service, according to the lens-shaped probability distribution function describing the sampling error, depicted in [19];
- the result of the initial map matching process, which selects a set of road segments falling within an error ellipse, representing the area in which the current position of the mobile device may be, according to the error of the localization service. In case the vehicle's initial position is further from roads junctions, the ellipse produced by the initial matching process won't contain any junction point nor shape point assuming the vehicle is outside of the known road network;
- the implementation of the route-matching service. The mobile-based approach requires high battery consumption, due to the route-matching algorithm executed at the mobile node whilst it reduces communications between the device and the central server. The centralized implementation relieves mobile devices of the CPU-intensive map-matching processing, although communications between devices and the server increase, accompanied by all the weaknesses of mobile communications, such as loss of connection, unpredictable latency and energy consumption.
- The route detail level required by the route-matching service application. Simplified routes are less accurate than fully detailed routes, but their computation time is lower .

E. Geofencing algorithms in literature

Custom virtual fences surrounding specific areas of interest have been used for more than a decade for on-line mapping applications, proximity-based digital coupon distribution and many other application software. Since its first appearance in research and technical literature, geofencing has evolved into a powerful geospatial analysis tool, becoming one of the most cutting-edge feature in application software and systems used in different fields.

Several different algorithms have been proposed in the technical literature for the implementation of geofencing functionalities and this section analyzes some of the latest and most interesting designs of the recent years.

One of the main problem is the *point-in-polygon problem (PIP)*, which, in computational geometry, solves the question about the position of a point with respect to the boundary of a polygon in the same plane. The PIP test finds application in areas dealing with geometrical data processing, such as computer graphics, computer vision, geographical information systems and many more. One of the first approach is the ray-casting algorithm, proposed in the early description of the point-in-polygon problem [20], but this method doesn't work in case the point is on the edge of the polygon.

Many of the discussed examples define simplified geofencing features by solving, with different approaches, the point-in-polygon test.

PISTON: Parallel In-memory Spatio-temporal TOpological joiN (PISTON) is a geofencing algorithm, designed by the research team of the Department of Computer Science of the University of Toronto, which implements a parallel, main memory, query execution infrastructure designed specifically to address spatiotemporal join [2]. PISTON, which was initially designed as an optimization of the INLJ2I geofencing algorithm, introduces a novel parallel, in-memory trajectory index I_R, designed to handle a high rate of location data updates, and a novel in-memory spatial index I_S, organized with a two level grid approach and specifically optimized for point-in-polygon test. PISTON delivers low query response times acceptable for real-time use-cases, even with large geofence datasets.

SLGC-1: Scan-Line Algorithm and Grid Compression (SLGC-1) is a geofencing algorithm, designed by the development team of the Software School of the Xiamen University of China to solve regional limited problems in Internet of Vehicles (IoV) systems with restricted time and storage requirements [3]. It works in two separate steps. In the preprocessing step the algorithm imposes a spherical grid on the geofence area in input, matching the shape of the real region, then a scan conversion algorithm is used to determine the location attribute

of each cell of the grid. Finally the grid is compressed using a QuadTree compression algorithm, which provides a memory-efficient index structure (storage requirements is less than O(n)) for the geofence area to analyze and calculates the Morton Code (MD code) to identify each node inside the QuadTree structure.

Geofencing via Hybrid Hashing: Geofencing via Hybrid Hashing was selected as one of the three best geofencing algorithms, out of the 29 submitted ones, proposed as a solution for the task posed by the ACM SIGSPATIAL GIS CUP 2013 contest [4]. It builds and updates the in-memory hash tables used to index polygons during system spare time, shifting some computation cost from the "point-in-polygon test" stage to non-time-critical processing stage [4]. taking advantage that, in typical geofencing applications, points position are changed much more frequently than those of polygons.

On the basis of the results obtained by testing the algorithm on the dataset provided for the ACM GIS CUP 2013 contest, the algorithm provides low response times with respect to many other algorithms.

III. Application Scenario

This section focuses on the application of geofencing algorithms in services related to Supply Chain Management (SCM), which involve the movement and storage of raw materials and unfinished products from the point of origin to the point of destination and/or consumption. SCM was traditionally driven by Enterprise Resource Planning systems, which provided plans and estimation regarding the different aspects of the business activity. In recent times, a quiet revolution has been taking place thanks to the use of Location-based technologies and innovative solutions to track and trace transportation equipment, materials and drivers across all the step of the supply chain [21]. LBSs allow the enterprise to dynamically tender and dispatch shipments in real time, divert a route because of weather conditions or a severe accident that is causing major delays in the transportation route, or transmit notification messages to the stakeholders whenever a shipment arrives at a warehouse [21] making the whole process more efficient and less expensive.

For all these reasons, location-based services can be considered a disruptive technology for the supply chain that will bring great opportunity for logistics innovation. The evolution in mobile telecommunication technology, together with the advances in electronics and the introduction of the IoT paradigm, has enabled the networking of portable wireless devices and wearable computers that can provide new types of usable knowledge to all the members and stakeholders of a globally dispersed supply chain [22].

These devices, equipped with sensors and actuators, exploit their connection capability to transmit to the Service Provider's servers, in a Machine-to-Machine communication, all the data regarding the status of the shipment which are important for the service's functioning.

The installation of M2M nodes in pallets, containers, vehicles and warehouses, along with new types of inference algorithms and techniques, will enable seamless, efficient, and transparent movement of raw materials and products through the global supply chain [22], allowing the business's customers to look at all the critical points of the chain.

M2M communication, together with geofencing, map-matching and localization services, represent the enabling technologies for developing and deploying a location-based service.

Developing a Location-Based Service, founded on M2M Communication, to support the Supply Chain Management raises several technical challenges, the most important are: geographical diversity and telecommunication coverage, location awareness, response time, accuracy of the result, power conservation, security and privacy, meet customer expectations.

Geographical diversity and telecommunication coverage: along their trip, from the moment goods are packaged for shipping to the moment they arrive at destination, containers and cargoes go to many places where GSM coverage is poor to non-existent. Thus, the use of dual mode GSM-satellite M2M devices is crucial to provide uninterrupted service to customers thus satellite communication is always available as a back-up technology to transmit the device's position. Furthermore, mathematical statistical interpolation may be used to fill the missing data.

Location awareness: in some application scenarios, such as air transport, the ability of the LBS to switch operative mode depending on devices' position could be an important feature both to meet customer needs and legislative restrictions (e.g. IATA restrictions on network-enabled electronic devices [23]).

Response time: the algorithm behind the LBS services should return the result of the computation within a specific deadline from the moment the packet, transmitted by the device, is received. This is important in order to guarantee the responsiveness of the application which uses the service, and is crucial for time-critical applications.

Accuracy of the result: the precision of devices' location depends on the hardware and software used in the mobile communication system, as well as on the positioning service [24]. The accuracy level requested to the Location-Based Service, both for position tracking or route-matching, influence the service's response time and varies depending on the application scenario in which the service will be used.

Power management: energy efficiency and power consumption are critical aspect when developing a LBS using battery-powered M2M devices. Containers and cargoes trips from source to destination may last 75 days in average, so the device attached to them should work properly for long period of time, often without the possibility to recharge the battery.

Security and privacy: customer concerns about security and privacy are another challenge for location-based technologies applied to Supply Chain Management. With regards to shipment security, it is desirable that the LBS integrates a priority function which immediately alerts the customer in case of illicit manipulation of the container holding the goods. On the other hand, the M2M devices should transmit the shipment status and location data using data security

instruments, such as cryptography, to keep them confidential and avoid interception of sensible information.

Finally, the developed services have to satisfy customers' expectations, in terms of expected results, perceived Quality of Service, reliability, availability and more importantly cost.

Among all available tools which contribute to the establishment of applications and systems for Supply Chain Management, geofencing plays an important role in the context detection of proactive applications, which can automatically adapt business and industrial operations to the geospatial context a user, or a mobile device, is currently in.

A. Geofencing service for SCM systems

Geofencing allows to detect and monitor the changing in the topological relation between a mobile device and a bounded geographical area (the geofence). The aforementioned topological relation can be expressed in terms of the spatio-temporal predicates enter, leave, cross and bypass, which are obtained as the result of trajectory-based topological join queries. These queries test the intersection between the whole or part of the trajectory of a moving object and a geofence, returning the spatio-temporal predicate describing their relationship based on the intersections found. These queries are really powerful instruments for geospatial analysis, but they are also computationally intensive.

This represents a challenge, as it requires to identify or construct the most efficient algorithm or method which, under the operating conditions of the specific system and application scenario, resolves geofencing problems, returning the result of trajectory-based topological join queries in a period of time that is acceptable for interactive, ad hoc geospatial analysis services.

The application scenario of the geofencing service proposed in this paper is that of IoT-based industrial services supporting Supply Chain Management and logistics for remote monitoring of goods and assets using mobile devices, smart cards, tags or similar technologies. This kind of services are placed in the context of Industry 4.0. In particular, the application scenario presented in this paper provides for a system using uniquely identifiable mobile objects, from here onwards called trackers, which transmit real-time location data with a precise, configurable frequency, which may change over time. The aforementioned system is centralized and thin-client, meaning that the trackers have limited resources and processing capabilities, in order to save battery power to provide a long-lasting monitoring service.

The geofencing service to be implemented should make it possible to detect whether one of the aforementioned mobile devices, capable of transmitting real-time location data to the system, enters, leaves or crosses one or more specific areas of interest, the geofences, and, whenever this occurs, it should notify the system of the event, depending on the specific service configuration assigned to the specific device. Each geofence should be statically defined by a geometric shape or by indicating a location identifier, such as an address. In order to reduce the processing within the tracker, all computing related to geofencing should be performed within server computers. Furthermore, to meet the real-time requirements, the geofencing routine must be non-blocking and the service must return a correct result within specific time constraints, often referred to as deadlines, failing which the result should be invalid.

In order to determine the topological relation between the moving object and the set of assigned geofences, the service should be coupled with a utility capable of reconstructing the path traveled by the device. In addition, if it is not possible to estimate all or part of the route traveled by the tracker, or in case a low level geospatial analysis is requested, the service should be able to work with the simpler geospatial data available at the time of the request. Following the reception of a notification from the geofencing service, the system should notify the user about the event and/or switch the device configuration depending on the event occurred.

According to the above features, the requirements of a good geofencing for SCM are:

- static, meaning that the spatio-temporal predicates are verified by checking the trajectory of the moving object with respect to fixed, bounded areas;
- geometric and symbolic addressed, so that the geofences could be defined with both geometric shapes or symbols, such as words and alphanumeric codes, which identifies precise locations;
- centralized, so that the matching between the trajectory of the moving object with the set of associated geofences is executed by the servers, which are the main part of the system;
- capable of operating effectively with different and variable location accuracy, tracking rate and device speed.

IV. Proactive Fast and Low Resource Geofencing Algorithm

In order to satisfy all the requirements discussed in the previous section, we propose a new geofencing design inspired by the ray-casting algorithm called Proactive Fast and Low Resource Geofencing Algorithm (PFLGA). The developed solution exploits geofences drawn over the WGS84 (or EPSG:4326) world geodetic coordinate system [25], which is a mathematical model of the Earth from a geometric, geodetic and gravitational point of view and is used by GPS navigation system and for aviation as a mandatory standard.

The proposed design provides every tracker for which geofencing service is enabled with a set of one or more geofences, i.e. geospatial objects, such as polygons and circles, whose boundaries are drawn over a specific geodetic coordinate system.

The data periodically transmitted by the tracker carries its geographical position, expressed in terms of latitude and longitude coordinates, enabling the location update for each tracker. These geographical points are used to determine the most plausible *path traveled* by the device between location

updates, exploiting route-matching services with different levels of detail. The *traveled path* is then used to determine the topological relation between the moving object and the set of geofences assigned to the shipment the device is attached to.

Although the algorithm is designed to determine the topological relation between a moving object and a set of geofences using its trajectory, it can also perform the geofence inclusion test using other geospatial objects, from a geographical location point to a single segment of the whole complex trajectory traveled by the tracker, depending on the available geospatial information regarding the moving object and the complexity required for the geofencing analysis. This allows the algorithm to be potentially applied to different use cases, from those which require an examination with low level of detail, in favor of a low query response time, to those that require a detailed geospatial analysis regardless of the query response time.

Since the application scenario in which the geofencing service will be used provides that each tracker can be assigned a set of one or more geofences, and the trajectory-based topological join queries are rather compute-intensive, it is important that data structures containing geospatial data support the retrieval of elements of an arbitrarily large size in an efficient way, therefore the proposed geofencing algorithm uses an in-memory, tree-based index structure for indexing the set of geofences assigned to each tracker for which the geofencing service is required. The insertion strategy for these structures has a computational complexity of $O(n)$, while the search operation has a computational complexity of $O(\log n)$, which permits a fast object retrieving in time critical applications.

The use of this kind of spatial index enables the application of an efficient filtering strategy on the set of geofences on which the intersection test with the trajectory will be performed.

PFLGA searches for any intersection between the route traveled by the tracker and progressively smaller bounding areas, called *Minimum Bounding Rectangles* (MBRs), which contain one or more geofences within them. In case the trajectory doesn't have any intersection with those bounding areas, the test ends without checking the set of geofences, otherwise the test continues with smaller bounding areas, until a precise geofence is found and tested.

PFLGA is based on the theory behind trajectory-based topological join queries. Given a set of geofences, bounded geographical areas represented as polygons or circular shapes, and the whole or a part of a trajectory defining the path traveled by the moving object, represented as a polyline geometric object, the algorithm verifies the existence of intersections between this polyline and the set of MBRs containing the geofences to be analyzed. If the polyline defining the trajectory intersects one or more MBRs, the algorithm performs the following steps on each geofence contained within the MBRs of interest:

- gets the previous position of the device and checks whether it was inside or outside the current geofence;
- calculates the intersections between the geofence and the trajectory traveled by the device, if there is any;
- analyzes the result obtained above and returns a composite topological predicate [19], which tells whether the object entered, crossed or left that precise delimited area.

This algorithm implies the a-priori construction of the in-memory, tree-based index, which will use a time interval proportional to the dimension of the set of geofences in exam (since the insertion algorithm for this tree-based index structure has a computational complexity of $O(n)$). Since the set of geofences is quite static and it is updated rarely, compared to the location of the moving object and its trajectory, the additional processing required for the index is bearable, especially if it is compared to the query processing speed up offered by the use of this index structure. The algorithm, whose possible implementation in pseudo-code is shown below, has a computational complexity of $O(\log n)$.

input : $\mathcal{F}$: set of geospatial objs making geo-fences
idx : index of the set of geofences
$route$: trajectory traveled by the object
output: $predicate$, position with respect to the fence

```
// list of FeatureIDs of the MBRs in
// the index intersecting the route
I = getIntersection(idx, route)
foreach pos ∈ I do
    prev_pos = getPrevPosition(route);
    // checks if the prev_pos was
    // inside the current MBR
    if (getIntersection(F[i],prev_pos)!=null) then
        wasInside = true;
    else
        wasInside = false;
    end
    Intersections = getIntersections(pos, route)
    if (wasInside) then
        if Intersections number is odd then
            the object left the fence
        else
            the object is still inside
        end
    else
        if Intersections is empty then
            the object is still outside
        else
            if Intersections number is odd then
                the object entered the fence
            else
                the object crossed the fence and it is
                outside
            end
        end
    end
end
```

Algorithm 1: PFLGA description

A. Comparison between geofencing algorithms

In order to verify the performance of the proposed geofencing algorithm, the developed solution is compared with other geospatial analysis algorithms which resolve the same geofencing problem.

First, the algorithms are compared on the basis of the set of features requested by the application scenario the proposed solution has been developed for. Subsequently, the developed design is tested using the same dataset and setup utilized for testing the performance of the other solutions considered in this comparison, which vary depending on the algorithm in exam.

All the features of PFLGA are used to study the similarities between the proposed design and the other geofencing solutions considered in the previous paragraph. Table I shows the comparison among the following features:

1) In-Memory Spatial Index (IMSP);
2) Spatio-Temporal Topological Join Predicates (STTJP);
3) Use of Trajectory-Filtering / Filtering Strategy (TFFS);
4) Geofencing via evaluation of point-in-polygon (PIP);
5) Geofencing via evaluation of spatial intersection with trajectory segments (SEGS);
6) Geofencing via evaluation of spatial intersection with complex trajectories (TRAJ).

TABLE I
COMPARISON OF FEATURES

	IMSP	STTJP	TFFS	PIP	SEGS	TRAJ
Hybrid Hashing	Yes	Yes	Yes	Yes		
SLGC-1	Yes		Yes	Yes		
PISTON	Yes	Yes	Yes		Yes	Yes
PFLGA	Yes	Yes	Yes	Yes	Yes	Yes

Next we compare the algorithms with respect to the benefits and/or their problems in the context of the application scenario the algorithm has been developed for.

Hybrid hashing: The Hybrid Hashing adopts a very efficient filtering strategy, based on the use of two in-memory hash tables, to reduce the time spent for the point-in-polygon test. Moreover, it builds or updates the in-memory hash tables for the geofences during system spare time, shifting some computation from the "point-in-polygon test" stage (which is often time critical) to non-time-critical processing stage. This algorithms shows a low response time. CONs. The main drawbacks are the high storage requirements, the time required for the construction and update of index and data structures for the geofences. Finally Hybrid Hashing is not able to return trajectory-based spatio-temporal topological join predicates.

SLGC-1: This algorithms uses QuadTree compression algorithm to store data regarding geofences, which reduces the amount of memory requested for the index structure, the complexity of the storage is less then $O(n)$ and the approach to geofencing is a simple and straightforward point-in-polygon based solution. Lastly the point-in-polygon test time does not increase with the number of edges of the analyzed geofence. As the previous algorithm, SLGC-1 is not able to return spatio-temporal topological join predicates, so the geospatial analysis is not deep.

PISTON: PISTON adopts a parallel in-memory indexing for trajectories and spatial geofences, which is a very scalable approach. Moreover, it evaluates the spatio-temporal topological join predicates with a sequence of topological relations that may hold between the trajectory of the moving object and the geofences at different time units, The trajectory index IR is optimized for high rate of location updates and can handle both coordinate-based and trajectory-based queries. PISTON adopts an efficient trajectory-filtering strategy and it is scalable due to its native multi-threaded setup. Main problems deal with the time required for the construction and update of the R-Tree index for the geofences, which is high, and the high memory requirements.

PFLGA: PFLGA adopts an in-memory, tree-based indexing for the spatial polygons representing the geofences of interest, the time required for the construction and update of the in-memory index for the geofences is low. Moreover, it adopts an efficient trajectory-filtering strategy which checks for intersections between the MBR of the analyzed trajectory and the MBRs inside the in-memory index and is able to return the trajectory-based spatio-temporal topological join predicates describing the relation between the moving object and each area in the set of geofences. Lastly, the algorithm can also evaluate the simple point-in-polygon test for the current position of the moving object if a simple geospatial analysis is requested. The main drawback is the response time of the trajectory-based topological join query, which is not the shortest among the other algorithms.

B. Performance Comparison

The performance of PFLGA is compared against the ones of the above discussed algorithms. In order to allow a fair comparison and avoid data dependencies, every test uses the same dataset used by the algorithm's author to determine the performance. The test has been executed on a laptop computer loaded with Intel Core i7 4720HQ CPU (Quad Core, 2.60 GHz up to 3.60 GHz) and 8 GB DDR3 RAM, running Linux Ubuntu 14.04 LTS using Apache Bench (also indicated as ab), a tool designed for benchmarking Apache installations and any HTTP server in general [26].

This section reports the dataset characteristics for each algorithms and the results in term of performance.

PISTON vs PFLGA: We used the following datasets to compare these algorithms:

1) Geofences Dataset: US TIGER® Texas Arealm, a real-world spatial-objects dataset which contains 6694 geofences drawn in the Texas area [52];
2) Trajectories Dataset: 10000 trajectories between couples of random location point inside the Texas area, generated according to the specifications in the PISTON paper [48].

The test requires the execution of a geospatial query for each of the trajectories contained in the dataset.

The following table compares the performance of PFLGA against PISTON [2].

TABLE II
PISTON VS. PFLGA

	PISTON	PFLGA
Average Spatial Index Creation Time (ms)	537000	860
Average Query Execution Time (ms)	340	330

Hybrid Hashing vs. PFLGA: For this performance test we used two datasets , according to the paper [4]. The first dataset is provided by the ACM Open GIS Cup 2013 [4], which includes two location point files and two polygons files, for this comparison we selected two of these files, one for each category:

1) Geofences Dataset: Poly10 file, which contains 32 instances of 10 different polygons;
2) Locations Dataset: Point500, which contains 39289 instances of 500 different points.

The test requires the execution of 10000 geospatial queries for each of the trajectories contained in the dataset, according to the experiment specification in the paper. The table III contains the result obtained at the end of the test. According with

TABLE III
HYBRID VS PFLGA

	Hybrid Hashing	Proposed
Average Spatial Index Creation Time (ms)	5	15
Average Query Execution Time (ms)	7	10

Hybrid Hashing's Authors we used another dataset constructed as follows:

1) Geofences Dataset: Poly-OSM1, which contains 200 instances of 20 different polygons, selected from the Land Polygon dataset.
2) Locations Dataset: Point-OSM1, which contains 80000 instances of location points randomly selected from the MBR area of each polygon.

The test requires the execution of 10000 geospatial queries for each of the trajectories contained in the dataset. Table IV shows the result obtained.

TABLE IV
HYBRID HASHING VS PFLGA

	Hybrid Hashing	PFLGA
Average Spatial Index Creation Time (ms)	40.9	45.0
Average Query Execution Time (ms)	76.0	46.0

SLGC-1 vs. PFLGA: The two datasets used for this performance comparison are constructed according to the SLGC-1 paper [3]. This geofences dataset contains 5 different polygons, respectively having 5, 10, 50, 100 and 223 vertexes, and a circular area. Each of these geospatial objects covers a geographical area almost equal to 4000 km^2, and Locations Dataset, 125000 different GPS location points limited in the geofences' areas. For each couple of GPS points a trajectory connecting them is calculated. The test requires the execution of a geospatial query for each of the trajectories contained in the dataset. Table V shows the results of the comparison.

TABLE V
SLGC-1 VS. PFLGA

	SLGC-1	PFLGA
Average Spatial Index Creation Time (ms)	19.0	1.5
Average Query Execution Time (ms)	3255	3

The comparison between PFLGA and the other solutions highlights the proposed algorithm has a wider set of features compared to the other solutions. In particular PFLGA is designed to perform the geofencing test with different geospatial objects, depending on the available information regarding the moving object, which offers a great flexibility depending on the use cases it is applied to. Moreover, in the inclusion test stage, PFLGA performances are comparable to, and in some case better than, those of the other designs, both in terms of average spatial index creation time, in the preprocessing stage, and average query execution time.

V. CONCLUSIONS

This work has described the design process of a geofencing algorithm to be used to implement proactive, context aware functionalities in Supply Chain Management systems with real-time requirements.

PFLGA has been extensively set side by side to other existing geofencing designs, presented in the last few years in technical literature. PFLGA shows more features and comparable performance to the other geofencing solutions, making it more flexible to the different service requirements which may occur depending on available data or service requests.

Subsequently, PFLGA has been implemented in a back end software module within See Your Box test environment. The algorithm has been tested using data collected by a set of trackers along their journeys on different routes in Europe, proving its correctness and its ability to provide the result of the requested analysis within precise time constraints. Then the algorithm has been implemented as a service within See Your Box system. As for future developments, there is still room for optimization of the search strategy inside the set of geofences assigned to a specific mobile device, referred to as the tracker. Upcoming implementations may adopt a new indexing structure for the set of geofences, in order to further reduce the computational complexity and the response time of the algorithm. Another great opportunity for improvement concerns the route-matching service. Route-matching is used to determine the most plausible route traveled by the tracker, given as input a set of geographical points, each of which is coupled with a timestamp, which is then exploited in the proposed algorithm to perform a topological analysis to catch the occurrence of precise events. The introduction or improvement of route selection strategies based on the timestamps of the collected geographical points, or on the estimation of the cruising speed of the device, would allow to increase the accuracy of the result offered by the route-matching service, which would bring great benefit to the

geofencing service in all those application scenarios within Supply Chain Management where the estimated trajectory traveled by the device represents a critical parameter (e.g. alert service in case of crossing a specific area not allowed by company policies).

REFERENCES

[1] S. R. Garzon and B. Deva, "Infrastructure-assisted geofencing: Proactive location-based services with thin mobile clients and smart servers," in *2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, March 2015, pp. 61–70.

[2] S. Ray, A. D. Brown, N. Koudas, R. Blanco, and A. K. Goel, "Parallel in-memory trajectory-based spatiotemporal topological join," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct 2015, pp. 361–370.

[3] K. Lin, Y. Chen, M. Qiu, M. Zeng, and W. Huang, "Slgc: A fast point-in-area algorithm based on scan-line algorithm and grid compression," in *2016 11th International Conference on Computer Science Education (ICCSE)*, Aug 2016, pp. 352–356.

[4] S. Tang, Y. Yu, R. Zimmermann, and S. Obana, "Efficient geo-fencing via hybrid hashing: A combination of bucket selection and in-bucket binary search," *ACM Trans. Spatial Algorithms Syst.*, vol. 1, no. 2, pp. 5:1–5:22, Jul. 2015. [Online]. Available: http://doi.acm.org/10.1145/2774219

[5] S. Steiniger, M. Neun, and A. Edwardes, "Foundations of location based services," 01 2006.

[6] I. Junglas and R. Watson, "Location-based services," vol. 51, pp. 65–69, 03 2008.

[7] V. Carchiolo, L. Compagno, M. Malgeri, N. Trapani, M. L. Previti, M. P. Loria, and M. Toja, "An efficient real-time monitoring to manage home-based oxygen therapy," in *Trends and Advances in Information Systems and Technologies*, A. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Cham: Springer International Publishing, 2018, pp. 741–749.

[8] M. Loria, M. Toja, V. Carchiolo, and M. Malgeri, "An efficient real-time architecture for collecting iot data," 2017, pp. 1157–1166, cited By 1. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039924461&doi=10.15439%2f2017F381&partnerID=40&md5=8de00d20f17d82a86236f43f33a716f8

[9] ESRI, "What is gis?" http://www.esri.com/what-is-gis, accessed: 2018.

[10] D. M. S. Sachin W. Rahate, "Geo-fencing infrastructure: Location based service," *International Research Journal of Engineering and Technology*, vol. 3, 11 2016.

[11] ITU, "Y.2060 : Overview of the internet of things," https://www.itu.int/rec/T-REC-Y.2060-201206-I, 06 2012, accessed: 2018.

[12] M. Rouse, "geo-fencing (geofencing)," http://whatis.techtarget.com/definition/geofencing, accessed; 2016.

[13] G. Allen, "Internet of things, industrial internet of things, industry 4.0—it's all connected! (no pun intended)," https://redshift.autodesk.com/industrial-internet-of-things-iot-terms/, accessed: 2015.

[14] M. Erwig and M. Schneider, "Developments in spatio-temporal query languages," in *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*, 1999, pp. 441–449.

[15] T. SearchNetworking, "Thin client (lean client) definition," accessed: 2016.

[16] M. Bauer, D. Dobre, N. Santos, and M. Schmidt, "Scalable processing of geo-tagged data in the cloud," *Nec Technical Journal*, vol. 7, no. 2, 2012.

[17] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," 2007.

[18] A. L. B. Nagendra R. Velaga, Mohammed A. Quddus, "Developing an enhanced weight-based topological map-matchingalgorithm for intelligent transport systems," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 672–683, 12 2009.

[19] D. Pfoser and C. S. Jensen, "Capturing the uncertainty of moving-object representations," in *Proceedings of the 6th International Symposium on Advances in Spatial Databases*, ser. SSD '99. London, UK, UK: Springer-Verlag, 1999, pp. 111–132. [Online]. Available: http://dl.acm.org/citation.cfm?id=647226.719082

[20] I. E. Sutherland, R. F. Sproull, and R. A. Schumacker, "A characterization of ten hidden-surface algorithms," *ACM Comput. Surv.*, vol. 6, no. 1, pp. 1–55, Mar. 1974. [Online]. Available: http://doi.acm.org/10.1145/356625.356626

[21] G. Allen, "Harnessing the power of location based services," http://blogs.dcvelocity.com/supply_chain_innovation/2016/03/harnessing-the-power-of-location-based-services.html, accessed: 2016.

[22] B. Rao and L. Minakakis, "Evolution of mobile location-based services," *Commun. ACM*, vol. 46, no. 12, pp. 61–65, Dec. 2003. [Online]. Available: http://doi.acm.org/10.1145/953460.953490

[23] I. A. T. A. (IATA), "Guidance on the expanded use of passenger portable electronic devices (peds)," 2014.

[24] R. Ahas and Ülar Mark, "Location based services—new challenges for planning and public administration?" *Futures*, vol. 37, no. 6, pp. 547 – 561, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0016328704001521

[25] *DOC 9674/AN 946 - WGS84 Manual*, ICAO, 2002.

[26] T. A. S. Foundation, "ab - apache http server benchmarking tool," https://httpd.apache.org/docs/2.4/programs/ab.html, accessed: 2018.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 747–750
DOI: 10.15439/2018F258
ISSN 2300-5963 ACSIS, Vol. 15

Enhancing Project Management for Cyber-physical Systems Development

Filipe E. S. P. Palma, Marcelo Fantinato
University of São Paulo, São Paulo, Brazil
Email: {fpalma, m.fantinato}@usp.br

Laura Rafferty, Patrick C. K. Hung
University of Ontario Institute of Technology, Oshawa, Canada
Email: {patrick.hung, laura.rafferty}@uoit.ca

***Abstract*—In this paper, specific practices are proposed for better managing Cyber-physical Sytems (CPS) projects, called CPS-PMBOK approach. CPS-PMBOK is based on the Project Management Institute's PMBOK body of knowledge. It is focused on the integration, scope, human resource and stakeholder knowledge areas; which were chosen considering a systematic literature review conducted to identify the main CPS challenges.**

I. Introduction

CYBER-physical Systems (CPS) refer to computational systems interacting with the physical world [1], [2]. CPS gained remarkable advances in science, such as medical surgery, autonomous vehicles, energy harvesting and smart buildings. A CPS is composed of a computing platform, the physical world, sensors and actuators [1], [2], [3]. CPS merge areas from embedded systems, mechanical engineering, software, among others [3]. CPS development projects tend to be large, complex and groundbreaking, with innovative technologies [1], [2], [4]. A usual feature is multidisciplinarity, which requires good team communication skills as CPS development merges computing and physical concepts. Collaboration among practitioners from different areas (such as software engineering, civil engineering, experimental physics or natural sciences) is needed to accomplish CPS developments [3], [4].

Project management practices aim to enhance the probability of success in a product or service development [5]. Success depends on organization, application area and project goals, and priorities may vary, including: finishing within planned time, meeting agreed scope, reaching satisfactory quality, or finishing in determined budget. Managing a project consists of controlling the development and providing all resources necessary for project execution, and it is a responsibility usually assigned to a project manager. Project management may be useful for many fields in most diverse applications, such as: medicine, civil engineering, software development, advertising campaigns etc. The Project Management Institute gathers best practices in the so-called Project Management Body of Knowledge (PMBOK) [6], which presents tools and techniques for a better management considering experts' knowledge. PMBOK organizes the best practices through five process groups (initiating, planning, executing, monitoring and controlling, and closing) and ten orthogonal knowledge areas (integration, scope, time, cost, quality, human resource, communications, risk, procurement, and stakeholder).

Considering the particularities of CPS projects and the need to manage them to reach their goals according to the success factors established, this paper addresses specific practices for better managing CPS projects. These specific practices are proposed as a PMBOK extension, called CPS-PMBOK. CPS-PMBOK is focused on the integration, scope, human resource and stakeholder knowledge areas. They were chosen considering a systematic literature review conducted to identify the main CPS challenges. Thus, we expect to improve both team communication skills and understanding of the project activities. The proposed practices are based on approaches previously presented in literature as well as the authors' background. We consider that a well-managed CPS project may increase physical world comprehension, modeling and interaction, enhancing the technological advances.

The remainder of this paper presents: related work and research method, the proposed approach, and conclusion.

II. Related Work and Research Method

Although PMBOK is a general-purpose guide, specific application areas, including CPS projects, may benefit from adapted or focused project management practices, which can better drive project activities and prevent common weaknesses [5]. Some authors propose, for example, new techniques for stakeholder management in civil engineering projects and in clinical research environments [7], [8]. Taking organizational structures differences, some works address concerns on stakeholders, scope, human resources, and communications for globally distributed projects [9], [10]. Other authors propose entire revisions of PMBOK processes, knowledge areas or other project management approach adaptations, but in a general way. One example extends the knowledge areas creating the new 'project sustainability management', dealing with reuse of lessons learned and standardization of project management practices within an organization [11].

To propose our PMBOK extension, we used results from a systematic literature review, conducted to link PMBOK's knowledge areas and the CPS development. We used various technical CPS-related terms to embrace as many primary studies as possible, such as: embedded systems, system of systems, sensors network, IoT, and automation and control.

The primary studies obtained were analyzed to find which knowledge areas were subject of study. A relevance score was applied based on the number of times that keywords related

to each area were mentioned. The outcomes are that scope, human resource and stakeholder were the areas with more issues studied. Considering the outcomes of this systematic review, our work proposes project management practices, focused on the CPS context for the scope, human resource and stakeholder knowledge areas. We also propose a generic practice related to the integration area.

These practices were found to manage scope in CPS projects: software and frameworks for requirements analysis, application of international standards, estimates based on use case points and hardware points, specific modeling languages for requirements elicitation and system architecture visualization, requirements review through peer reviewing and Scrum boards, development of design models, meetings with live demonstrations, and requirements lists and model-driven design [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

As for the project human resource, these practices were found to CPS projects: use of an expert and multidisciplinary team, statistical estimates and classification of familiarity of team members, training in specific development methods, such as goal- and model-driven and extreme programming, and skill-based human resource management [12], [14], [15], [32], [17], [33], [21], [22], [23], [26], [28], [27].

Finally, considering the project stakeholders, these practices were found as suggestions to address this knowledge area in CPS projects: identification of stakeholders and assignment of tasks following systematic algorithms and norms, assignment of stakeholders within the organization, involvement of stakeholders during the transition between development phases, and workshop meetings and constructive SoS integration model [12], [14], [15], [17], [21], [23], [24], [26], [28], [34].

To propose CPS-PMBOK, we further analyzed all the practices obtained in the primary studies to find practices still not covered by PMBOK and practices already covered but with suggested specializations. The final practices chosen are those most frequently found in the primary studies as well as aligned with the primary insights of the authors of this work.

III. The CPS-PMBOK Approach

CPS-PMBOK is composed of the original set of PMBOK best practices, extending it for CPS projects. The specialization address four PMBOK's areas. For each, one or more practices are proposed: *(a)* integration – characterization model (artifact); *(b)* scope – pre-elaborated requirements lists (technique), review requirements (process), process simulation (technique); *(c)* human resource – specialized team division (technique), cross-training (technique); *(d)* stakeholder – build technical trust (technique), dynamic follow-up strategies (technique).

A. Project Integration Management

The proposed practice 'characterization model' should be used as a brainstorm driver, to equalize the comprehension and familiarization with the system being developed. It should be produced as an output of the develop project chart process, which is part of the initiating process group; and it should be used as input by all processes that use the project charter also as input, i.e.: plan scope management, collect requirements, define scope, plan schedule management, plan cost management, plan risk management, and plan stakeholder management. During the brainstorming, participants should indicate levels for some characteristics, providing estimates about project size and technical challenges besides to discussions among team members. These characteristics are divided in: *(i)* CPS environment, representing the variables present in the CPS to be developed, such as how much limited tasks are required, communication with known group of devices, interaction with known group of people, and industrial standards or norms should be followed and *(ii)* CPS complexity, based on specific technological areas, such as mechanical structures, network, sensors, actuators, data storage, user interaction, legacy systems integration, and power energy systems.

B. Project Scope Management

In terms of scope, some processes present special challenges for CPS projects due to their highly innovative and dynamic aspects [35], [2], [4], [3]. In addition, the high complexity involved for modeling the physical world and its phenomena is another challenge source. CPS project managers and team should be able to constantly look for new requirements, bringing up changes in scope as soon as possible. As a result of this scenario and needs, two practices are proposed to the scope management, as presented in this section: pre-elaborated requirements lists and review requirements.

1) Pre-elaborated Requirements Lists Technique: to gather requirements, CPM-PMBOK includes a technique called pre-elaborated requirements lists to create reusable assets by gathering common requirements in CPS projects. This technique is proposed to be used within the collect requirements process, which is part of the planning process group.

2) Review Requirements Process: CPS development may lead to unexpected results and dynamic requirements [23]. Since such scope revisions and redefinitions are highly common in CPS projects, one of our specific practices is proposing an additional process to the scope knowledge area – review requirements – as part of the monitoring and controlling process group. Review requirements results in change requests similarly to performed by the control scope process, as described in PMBOK. The difference is that, in CPS-PMBOK, review requirements is a creation-focused process, considering less the already known requirements and revisiting the highest definitions of the project looking for new requirements. In PMBOK, the control scope process focuses on ensuring the accomplishment of the defined scope and, when needed, the appropriate processing of changes is made. In this new process, techniques to collect requirements already described in PMBOK are used, as meetings, surveys and interviews.

3) Process Simulation Technique: this technique is added in support of review requirements. Simulation tools to predict environment or conditions such as mechanical simulation, radiation diagrams and thermal dissipation are useful in review requirements and are part of process simulation. Other tools

to isolate part of the CPS, to validate models or equipment, such as hardware or software in the loop may be used.

C. Project Human Resource Management

Considering multidisciplinarity, human resources can be from different specialization areas, what increases the challenge of managing relationships and technical communication [33]. As a result, two additional techniques are proposed in CPS-PMBOK for human resource management: specialized team division and cross-training.

1) Specialized Team Division Technique: specialized team division is included in CPS-PMBOK to improve the development performance and avoid inappropriate assignment of tasks. The team should be split into subteams taking different application areas or project deliverables. Some works found in literature were used as a basis to propose it, including: the application of team division based on academic profiles, such as electrical engineering, computer engineering and information technology [22], [2]. This technique is proposed to be used within two processes: the plan human resource management process, which is part of the planning process group; and the acquire team process, which is part of the executing process group. We propose an initial suggestion for a specialized team division considering the context of CPS projects and taking into account the proposed characterization model in terms of CPS complexity. According to our suggestion, the sub-teams for a CPS projects could be: *(a)* mechanical design team – responsible for physical structures and mechanical packing; *(b)* hardware design team – responsible for processing platforms, sensors and actuators specification; *(c)* electrical design team – responsible for electrical project and drawings, besides power energy design; *(d)* network design team – responsible for communication protocols and technologies specification; *(e)* information system development team – responsible for software development; *(f)* other specialized teams – power bank development team, human-computer interface team, antenna design team, specific sensors team etc. Other options for specialized team division can be used according to specific project needs, based on the context of the system application. An alternative division is based on deliverables or partial results of the project, assigning a focused team for each logical deliverable part of the developed CPS system. A specialized team division may be used to support organizational or resource breakdown structures.

2) Cross-training Technique: cross-training is a practice briefly depicted in PMBOK, proposed to reduce impact when a team member leaves the project. It consists in allocating more than one resource to a task execution. For CPS projects, we propose that the cross-training should be always used to enable some team members acting as a communication bridge between different sub-teams by allocating a team member from a given area to perform a task of some other area. This technique is proposed to be used within the develop team process, which is part of the planning process group. Considering cross-training, a software engineer may sporadically follow a mechanical engineer's work with the purposed of understanding and even positively contributing with potential ideas and insights emerged from another outlook. Cross-training can be used as a facilitator in the identification and development of multidisciplinary practitioners.

D. Project Stakeholder Management

Project stakeholders in CPS projects are usually highly technical or very close to the system's final users. This occurs mainly in joint projects of research with universities, involving researchers and students. Also in industrial projects aiming to improve production performance, where many stakeholders are production leaders experts in many technologies of the area [4]. Consequently, two additional techniques are proposed in CPS-PMBOK for stakeholder management: build technical trust and dynamic follow-up strategies.

1) Build Technical Trust Technique: CPS projects tend to involve academic researchers or experts to support the development of CPS physical elements. They may represent technical stakeholders who know both the application and engineering areas. PMBOK describes a practice of trust building for stakeholder engagement management, showing that the company, team and the manager have competencies to accomplish project's requirements in time and cost. Accordingly, when involving technical stakeholders in CPS projects is to build technical trust between them and the team. In this context, CPS-PMBOK proposes a specialization of the trust building, adding the technical aspect to this practice. Build technical trust is proposed to be used within the manage stakeholder engagement process, which is part of the executing process group. Build technical trust means to pass technical confidence regarding project accomplishment conditions, considering the team and project manager. Accordingly, the team should get close to the stakeholders, mainly in situations in which the stakeholders are highly technical. For CPS-PMBOK, an internal expert or an external consultant should be put in charge of following up the project management activities allowing more technical stakeholders to be more comfortable with the project progress. This person has the role of translating technical stakeholders concerns. The technical trust may improve stakeholders' satisfaction due to their proximity and understanding of technical issues. Besides that, the developers may feel more comfortable as well, due to the understanding of terms and concerns provided by a expert or consultant.

2) Dynamic Follow-up Strategies Technique: some approaches found to improve communication with CPS projects' stakeholders are: face-to-face meetings to update the project status to stakeholders [23], stakeholders' participation in every last weekly follow-up meeting of development iterations [21], and weekly workshops for system demonstrating – to update stakeholders [24]. Most of these approaches are based on agile methods, which has the communication with stakeholders as one of their most important concerns. To meet the different levels of demand and satisfaction of stakeholders, we propose dynamic follow-up strategies as part of CPS-PMBOK. This technique is proposed to be used within two processes: the manage stakeholder engagement process, which is part of the

executing process group; and the control stakeholder engagement process, which is part of the monitoring and controlling process group. According to different aspects of a given specific CPS project, the project manager should adapt the follow-up strategy aiming to enhance stakeholder engagement and reach their expectations. The following suggested strategies are proposed: *(i)* during the project initiation and planning stages, which involve, for example, discovering of requirements and stakeholders, understanding of highly engaged stakeholders, and understanding of stakeholders' application area – regular face-to-face meetings should be adopted as follow-up strategy; *(ii)* during the project execution and monitoring stages, which involve, for example, resolution of requirements conflicts and alignment between technical demands from stakeholders and project documents – only sporadic participation of stakeholders could be included during planning and technical meetings; and *(iii)* during the closing stage, which involves, resource scarcity, time re-planning and stakeholder staff updating – the stakeholders should be able to follow up on the final results through workshops with live CPS demonstrations.

IV. Conclusion

This work proposed project management practices driven to CPS projects. The approach is based on PMI's PMBOK best practices and focused on integration, scope, human resource and stakeholder. CPS-PMBOK relies on the following requirements for CPS projects: multidisciplinary teams, high level of innovation and unpredictable requirements. Our challenge is to be able to evolve the proposed practices considering two needs that can be seen as antagonistic ones: on the one hand, being specific to the CPS project domain; but, on the other hand, being not too specific in order to allow adjustments as required for specific contexts and organizations.

References

[1] R. R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: The next computing revolution," in *47th Des. Aut. Conf.*, 2010, pp. 731–736.

[2] L. Sha, S. Gopalakrishnan, X. Liu, and Q. Wang, *Cyber-physical systems: A new frontier*. Springer, 2009, pp. 3–13.

[3] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems: A cyber-physical Systems Approach*, 2nd ed. Lee & Seshia, 2017.

[4] R. Baheti and H. Gill, "Cyber-physical systems," *The Impact of Control Technology*, vol. 12, pp. 161–166, 2011.

[5] A. Lester, *Project Management, Planning and Control: Managing Engineering, Construction and Manufacturing Projects to PMI, APM and BSI Standards*, 6th ed. Butterworth-Heinemann, 2014.

[6] PMI, *Guide to the Project Management Body of Knowledge*. Project Management Institute, 2013.

[7] Y. Shen, M. M. Tuuli, B. Xia, T. Y. Koh, and S. Rowlinson, "Toward a model for forming psychological safety climate in construction project management," *Int. J. of Proj. Manag.*, vol. 33, no. 1, pp. 223–235, 2015.

[8] S. R. Pandi-Perumal, S. Akhter, F. Zizi, G. Jean-Louis, C. Ramasubramanian, R. E. Freeman, and M. Narasimhan, "Project stakeholder management in the clinical research environment: How to do it right," *Frontiers in Psychiatry*, vol. 6, pp. 71.1–71.18, 2015.

[9] S. Deshpande, S. Beecham, and I. Richardson, "Using the PMBOK guide to frame GSD coordination strategies," in *8th Int. Conf. on Global Software Engineering*. IEEE, 2013, pp. 188–196.

[10] R. Golini and P. Landoni, "International development projects by non-governmental organizations: An evaluation of the need for specific project management and appraisal tools," *Impact Assessment and Project Appraisal*, vol. 32, no. 2, pp. 121–135, 2014.

[11] P. J. A. Reusch, "Extending project management processes," in *8th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, vol. 2. IEEE, 2015, pp. 511–514.

[12] B. Greene, "Agile methods applied to embedded firmware development," in *Agile Development Conf.* IEEE, 2004, pp. 71–77.

[13] D. Jun, L. Rui, and H. Yi-min, "Software processes improvement and specifications for embedded systems," in *5th ACIS Int. Conf. on Software Engin. Research, Management & Applications*. IEEE, 2007, pp. 13–18.

[14] R. Madachy, B. Boehm, and J. A. Lane, "Assessing hybrid incremental processes for SISOS development," *Software Process: Improvement and Practice*, vol. 12, no. 5, pp. 461–473, 2007.

[15] R. J. Madachy, "Cost modeling of distributed team processes for global development and software-intensive systems of systems," *Software Process: Improvement and Practice*, vol. 13, no. 1, pp. 51–61, 2008.

[16] C. M. B. d. Silva, D. S. Loubach, and A. M. d. Cunha, "An estimation model to measure computer systems development based on hardware and software," in *28th Dig. Avi. Sys. Conf.*, 2009, pp. 6C2.1–6C2.12.

[17] A. Shatil, O. Hazzan, and Y. Dubinsky, "Agility in a large-scale system engineering project: A case-study of an advanced communication system project," in *Int. Conf. on Soft. Sci., Tech. and Eng.*, 2010, pp. 47–54.

[18] C. Berger and B. Rumpe, "Supporting agile change management by scenario-based regression simulation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 504–509, 2010.

[19] J. R. B. Garay and S. T. Kofuji, "Architecture for sensor networks in cyber-physical system," in *Latin-Amer. Conf. on Comm.*, 2010, pp. 1–6.

[20] D. Savio, P. C. Anitha, and P. P. Iyer, "Considerations for a requirements engineering process model for the development of systems of systems," in *1st Works. on Requirements Engineering for Systems, Services and Systems-of-Systems*. IEEE, 2011, pp. 74–76.

[21] G. Rong, D. Shao, H. Zhang, and J. Li, "Goal-driven development method for managing embedded system projects: An industrial experience report," in *Int. Symp. on Empirical Software Engineering and Measurement*. IEEE, 2011, pp. 414–423.

[22] R. Helps and F. N. Mensah, "Comprehensive design of cyber physical systems," in *13th Ann. Conf. on Inf. Tech. Educ.*, 2012, pp. 233–238.

[23] P. M. Huang, A. G. Darrin, and A. Knuth, "Agile hardware and software system engineering for innovation," in *Aeros. Conf.*, 2012, pp. 1–10.

[24] B. Penzenstadler and J. Eckhardt, "A requirements engineering content model for cyber-physical systems," in *2nd Works. on Requir. Engineering for Systems, Services and Systems-of-Systems*. IEEE, 2012, pp. 20–29.

[25] C. C. Insaurralde and Y. R. Petillot, "Cyber-physical framework for early integration of autonomous maritime capabilities," in *Int. Systems Conf.* IEEE, 2013, pp. 559–566.

[26] J. Zhu and A. Mostafavi, "Towards a new paradigm for management of complex engineering projects: A system-of-systems framework," in *8th Annual IEEE Systems Conf.* IEEE, 2014, pp. 213–219.

[27] A. V. Parkhomenko and O. N. Gladkova, "Virtual tools and collaborative working environment in embedded system design," in *11th Int. Conf. on Remote Engineering and Virtual Instrument*. IEEE, 2014, pp. 90–93.

[28] T. Yue and S. Ali, "Applying search algorithms for optimizing stakeholders familiarity and balancing workload in requirements assignment," in *Conf. on Genetic and Evolut. Comput.* ACM, 2014, pp. 1295–1302.

[29] G. Sapienza, I. Crnkovic, and P. Potena, "Architectural decisions for hw/sw partitioning based on multiple extra-functional properties," in *IEEE/IFIP Conf. on Software Architecture*. IEEE, 2014, pp. 175–184.

[30] M. Faschang, F. Kupzog, E. Widl, S. Rohjans, and S. Lehnhoff, "Requirements for real-time hardware integration into cyber-physical energy system simulation," in *Works. on Modeling and Simulation of Cyber-Physical Energy Systems*. IEEE, 2015, pp. 1–6.

[31] Z. Lattmann, J. Klingler, P. Meijer, J. Scott, S. Neema, T. Bapty, and G. Karsai, "Towards an analysis-driven rapid design process for cyber-physical systems," in *Int. Symp. on Rapid Sys. Prot.*, 2015, pp. 90–96.

[32] Y. M. Chen and C.-W. Wei, "Multiagent approach to solve project team work allocation problems," *Int. J. of Production Research*, vol. 47, no. 13, pp. 3453–3470, 2009.

[33] C. Wolff, I. Gorrochategui, and M. Bücker, "Managing large hw/sw codesign projects," in *6th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems*, vol. 2. IEEE, 2011, pp. 919–922.

[34] M. P. Singh, "Norms as a basis for governing sociotechnical systems," *ACM Trans. on Intel. Systems and Technology*, vol. 5, no. 1, p. 21, 2013.

[35] E. A. Lee, "Cyber physical systems: Design challenges," in *11th Int. Symp. on Obj. Orie. Real-Time Dist. Comp.* IEEE, 2008, pp. 363–369.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 751–760
DOI: 10.15439/2018F27
ISSN 2300-5963 ACSIS, Vol. 15

Towards a Language to Support Value Cocreation: An Extension to the ArchiMate Modeling Framework

Christophe Feltus, Erik HA Proper
Luxembourg Institute of Science and Technology,
5, avenue des Hauts-Fourneaux,
L-4362 Esch-sur-Alzette, Luxembourg
christophe.feltus@list.lu, erik.proper@list.lu

Kazem Haki
University of St.Gallen,
Müller-Friedberg-Strasse 8,
CH-9000 St.Gallen, Switzerland
kazem.haki@unisg.ch

***Abstract*—Value cocreation is gaining momentum as organizations' underlying business logic and encompasses tools and techniques for discovering new valuable and necessary artefacts to support inter-organizational and network-centric business activities. To cocreate value, organizations must talk to each other using a clear and easy to use language. In the course of the ValCoLa (Value Cocreation Language) project, we aim at elaborating such language. To that end, in previous work, we developed a value cocreation metamodel based on three dimensions: the nature of the value, the object concerned by the value and the method to cocreate value. In this paper, we first extend ArchiMate to the domain of value cocreation to provide our metamodel with a dedicated modeling language. Second, we illustrate the language with a case study from the financial sector.**

I. Introduction

Business collaboration is a process that requires a considerable examination of the jointly created value among the parties involved in these exchanges. Value cocreation (VCC) is a notion mostly associated to the paradigm of service-dominant logic (S-DL), rooted in the marketing theories and whose aim is to *jointly create value during business exchanges among two or more partners* [1], [2]. A first example of VCC between a company and its customers is PowerDrive, a Swedish manufacturer of hydraulic drive systems that cocreates value with three of its customers based on the collection and analysis of data from an existing remote monitoring system [3]. Another example is Starbucks that has developed an online community platform to allow its customers, around the globe, to suggest innovative ideas and to allow the most voted ones to be deployed in practices [4]. In those examples, but also in other ones like those reported in [17], VCC is made possible thanks to the interconnections between the involved parties' information systems (IS). Accordingly, depicting value cocreation processes is paramount for IS designers but also to support the communication between IS designers and developers. Therefore, in our previous work, we designed an abstract language (metamodel) to support VCC exchanges [16], [21], [22].

To construct this abstract language, we first observed that the creation of value is an integration of three dimensions [16]: the nature of the value (e.g., financial value, quality, and security [5]-[8]), the object concerned by the value (e.g., a service, a contract, and a database [9]-[11]) and the method used to create the value (e.g., model-based, by design, chunk [12]-[15]). We also observed that, in practice, each of these dimensions is expressed using a specific language and that none of them alone allows expressing all dimensions at once. This lack of shared language is a problem when IS designers want to communicate together, especially when there is a shift from a local creation of value to a cocreation of value in a network of organizations. Indeed, in this context, communication among the IS designers from each of the involved organizations is essential. Due to the different languages that may be used by the different organizations engaged in value cocreation, however, communication can become extremely complex.

To address this problem, our approach consisted in building a value creation metamodel that simultaneously captures and abstracts all the dimensions of value cocreation. By abstracting the value propositions (originating from each organizations of the network), our goal was *to support the IS designers* from those organizations to communicate with each other using a shared language, expressed by means of common elements, having the same semantic (definitions of the concepts), the same structure (associations between concepts) and the same syntax (modelling language). Practically, and as demonstrated in [16], while being instantiable with specific languages, the VCC metamodel is suited to play the role of binding element between the modelling languages (i.e., the language has been designed at an abstraction layer appropriated to be instantiated to various types of value, like the security or the quality).

In this paper, we have exploited an enterprise architecture (EA) model to express VCC using only one language. EA consists in approaches which enable illustrating the interrelations between a company's different layers and between its different aspects such as behavior, information, or people. EA metamodels provide views that are understandable by all the stakeholders and that allow making decisions and trace the impact of such decisions. Although the concept of value exists in some EA metamodels, this concept (and its relationship with other concepts), is not appropriate to express value cocreation. As a result, we acknowledge that existing EA metamodels are not dedicated to accurately model value cocreation. However, we consider that the EA metamodels

The research is supported by the National Research Fund, Luxembourg (http://www.fnr.lu) and the Swiss National Science Foundation, Switzerland (http://www.snf.ch), and financed by the ValCoLa (VCC Language) project.

provide a good basis for modelling VCC since they model the most significant concepts of a company's information systems. To reap the benefits of the enterprise architecture metamodel for value cocreation engineering and management, we have opted for focusing our research on integrating the value cocreation metamodel with the ArchiMate EA metamodel. We have decided to focus on ArchiMate because it does not address VCC at all yet and because it is an open standard published by The Open Group[1].

All along the paper, the usage of the ArchiMate extension to the value cocreation is illustrated with a case study related to knowledge-intensive business services in the financial sector [21]. This case study concerns the cocreation of value between a bank and a datacenter. The context is that because both organizations have been collaborating for a long time, the datacenter has good knowledge of the bank's information system. For that reason, the bank has decided to outsource the improvement of the privacy of the customers' data to the datacenter. Both have hence started to cooperate in designing the privacy improvement service of the customers and therefore the bank has agreed to give information about its information system (architecture, functions, etc.) to the datacenter. In turn, the datacenter enhances its offer of services and thereby stabilizes its own business. The enhancement is possible as a result of the bank's feedback.

In the following, we first present the state of the art in VCC as well as our previous work in VCC modeling in Section II. Then, we introduce ArchiMate, its language and its extension mechanisms in Section IIIa, b, and c, and we extend it for expressing value cocreation in Section III.d. The financial case study is presented in Section IV and consists in expressing VCC metamodel through ArchiMate extension. Finally, Section V discusses the results and proposes future works and Section VI concludes the paper.

II. State of the art and previous works

This section presents the state of the art related to VCC modeling using concrete syntax and more especially using the ArchiMate metamodel.

A. Litterature review

Value cocreation is a very old topic that has been incorporated by Vargo and Lusch in the notion of service-dominant logic [1, 2]. According to the authors, a service is the basis of all exchanges and focuses on the process of value creation rather than on the creation of tangible outputs. Against this backdrop, Vargo and Lusch further elaborate on the idea that value is derived and determined in use rather than in exchange, meaning that value is proposed by a service provider and is determined by a service beneficiary. Hence, the firm is in charge of the value-creation process and the customer is invited to join in as a co-creator [2].

For Grönroos [47], this interaction is defined through situations in which the customer and the provider are involved in each other's practices. Consequently, the context (social, physical, temporal, and/or spatial) determines the value-in-use experience of the user in terms of his individual or social environment [48].

Recently, Chew [49] has argued that, in the digital world, service innovation is focused on customer value creation. Chew proposes an integrated Service Innovation Method (iSIM) that allows analyzing the interrelationships between the design process elements, including the service system. The latter being defined as an IT/operations-led, cross-disciplinary endeavor. In IS literature, Blaschke et al. [50] propose a business-model-based management method encouraging cocreation interactions by reconciling value propositions, customer relationships, and interaction channels.

Gordijn et al. [51] explain that business modeling is not about process but about value exchange between different actors. Gordijn et al. propose e3value to design models that sustain the communication between business and IT groups, particularly in the context of the development of e-business systems. In [52], Weigand extends the e3value language to consider cocreation. He defines so-called value encounters, which consist in spaces where groups of actors interact to derive value from the groups' resources. In a similar way, Razo-Zapata et al. propose visual constructs to describe the VCC process [53]. These constructs are built on requirements from the service-dominant logic and software engineering communities.

B. The VCC metamodel

In this section, the metamodel of value creation in the field of IT-related business services is defined according to three dimensions: the nature of the value, the method of value creation, and the object concerned by the value.

Provided that this research is anchored in Design Science Research [19-20], its development has followed an iterative cycle. Only the last version of it is presented in this section. The first version was presented in the conference FedCSIS 2017 [16], the second version in LNBiP [21], and the last version in AINA 2018 [22]. This metamodel is elaborated based on the analysis of value related frameworks [5]-[8], of scientific literature [1], [2], [47], [51], [52] and on a performance evaluation methodology for decision support in industrial project proposed in [23]. The aim of this methodology is to propose a benefit-cost-value-risk based approach to help decision makers in evaluating performance at any stage of an industrial project.

In the next sub-sections, each dimension of the value is successively analyzed and presented. Moreover, concepts of our VCC metamodel are illustrated using the first part of the case study.

[1] http://www.opengroup.org/subjectareas/enterprise/archimate

1) Dimension 1: Nature of the value

To understand and model the nature of the value, first we have reviewed a set of frameworks addressing the different value natures in the field of IT, including security, quality, compliance, privacy, responsibility, and others. Based on this review, we have extracted the most meaningful concepts necessary to express this nature. For example, we have analyzed the information systems security risks management (ISSRM [24]) framework, which addresses the IS security (*Nature of the value*). This framework characterizes security through integrity, confidentiality, availability, non-repudiation, and accountability (i.e., *Value component* concept of the VCC metamodel). And the security concerns business assets of the company (*Objects* concept of the metamodel). Finally, based on a further review of the literature, our own definitions of the constitutive concepts of the dimension have been provided and the concept has been integrated in the nature of value metamodel (Fig. 1).

Basically, most of the analyzed reference frameworks focus on depicting the semantic of value following a given perspective being function of the beneficiary of the value. In practice, due to the quantity of heterogeneous value natures [32], clearly defining the semantic of this nature is laborious. However, we observe that, in the same transaction, two main perspectives of value nature emerge depending on the context: value at the provider's side vs. value at the customer's side. At the provider's side, the basic rationale for all organizations entering into dyadic exchange relationships is the value capture [33] from a service exchange. This can be in the form of value-in-exchange (e.g., money given by the client), or in the form of value-in-context. In that regard, it is worth noting that considering the provider in the context of the digital society expands this narrow meaning to the consideration of other value elements. An example of them are the information collected on the customers (e.g., analyzing customer data to support the creation of new offerings) which, afterwards, contributes to economic increase [34]. On the customer's side, value generated by a transaction never refers to money but consists in other wealth, which contributes in sustaining and supporting the customer's own business.

According to [23], value is described as the degree of satisfaction of a set of stakeholder expectations or needs, expressed by the level of appreciation associated to a number of performance indicators. Li [35] explains that value can be described by the relative worth, utility, or importance of something. Value increases when the customer's degree of satisfaction increases. The concept of value becomes different depending on the point of view (stakeholder). Accordingly, the *expected value* is the value that the stakeholder would like to get and the *perceived value* is the real value that a stakeholder can finally get. The degree of satisfaction is identified through the comparison of these two elements. According to Zeithaml, value implies some form of *assessment of benefits against sacrifices* [36].

In our analyzed case, at the bank's side, the privacy of the customers' data is a legal requirement that has to be fulfilled by each entity processing private information. Having this data privacy generates the benefit of being compliant with regulations, but it is also expensive because the bank needs to deploy an appropriate mechanism such as performing privacy impact assessment. At the datacenter's side, offering 24/7 data availability to the bank is a benefit to distinguish the datacenter from its competitors, but this offering is also costly because it requires a very robust infrastructure.

According to this review, the concepts that are relevant to the metamodel for the nature of the value are:

- **Value.** This concept is defined as a degree of worth of something [23, 35] and that improves the well-being of the beneficiary after it is delivered.
- **Nature of the value.** The nature of the value **defines** the value to be delivered. Table 1 shows that the nature of the value expresses a domain of interest related to which the value will be delivered (e.g., security of the IS, the cost of a transaction, or the privacy of personal data). In the case of the datacenter that archives the data of the bank customers, the nature of the value generated by the datacenter is the *availability* of the customer's data.
- **Value component**. This concept expresses the different elements that constitute the value (e.g., availability, confidentiality, portability, etc.). Hence, the value **aggregates** value components and these components may also, as a result, themselves be other **types of** value. Regarding the case study, one component of the availability is the *accessibility in real time*.
- **Object.** The object concerned by the value is the element from the information system that has significance and is necessary for a company to achieve its goal (e.g., software, process, data). From a modeling point of view, the value is associated to an object with a relation of type **concerns** or an objective to be achieved. In the case study, the object concerned by the value is the *customers' data.*
- **Measure.** The measure corresponds to a property on which calculations can be made for determining the amount of value expected from a value cocreation method. Measure can result from different factors impacting value. As stated by [23, 35], the value components are measured by means of estimation methods. Accordingly, there exist an association named **appraises** from the concept of measure to the concept of value, an association named **is function of** between the concept of measure and the type of value, and between the concept of measure and the object concerned by the value. The first expresses that the measure is characterized by the nature of the value and the second posits that the measure also depends on the object concerned by the value. According to [35], measure may integrate qualitative and quantitative elementary performance expressions.

Based on the above definitions, the nature of the value is modeled in Fig. 1.

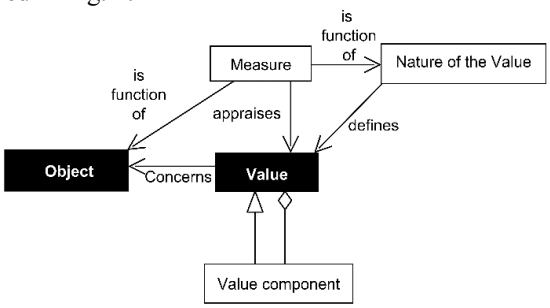


Fig. 1. Nature of the value metamodel

2) Dimension 2: Method of value creation

A method of value creation corresponds to a set of activities that contribute to the generation of value. Similar to the nature of the value, to depict the elements relevant for the creation of value, a set of IT-related frameworks on value creation methods have been reviewed. The analyzed methods include method by design [12], model driven [15], impact assessment [17], method chunk [14], risk-based [37], and process-based [38] approaches.

Traditionally, value is created through the exchange and use of goods and services [1]. Methods for value creation are the body of techniques and activities that use and generate resources [39]. These correspond, at the corporate level, to a bundle of approaches including the design of strategies, the integration of models, and the evaluation of results. By looking more closely at the analyzed methods, it has been observed that each has a dedicated goal, that they are composed of method elements, and that method elements are organized in a sequence of ordinated steps. For instance, by investigating the model-driven approach to interoperability, one can notice that it has for goal to improve interoperability of enterprises' information systems that it is composed of models, and that three steps are required for model-driven interoperability: model design, model integration, and model instantiation. Amongst the other methods reviewed, it is also interesting to highlight that one (method chunk) has for particular objective the creation of methods themselves, using, as chunk of existing methods as method elements, and as method steps the decomposition of existing methods into method chunks and the definition of new method chunks from scratch [14].

As a summary and according to this analysis, the concepts that construct the method of value creation are:

- **Method**. The method is a specific **type of** object that defines the means used by the stakeholder to **create** objects and value. A method is **composed** of a set of activities necessary to achieve a dedicated goal. In the same vein, Sein et al. [40] explain that the elementary quantitative value expressions (the value components) are aggregated by means of selected aggregation methods and quantitative weights to generate the overall value. An example of method used to create security of the IS consists for instance in performing a security risk assessment [24].
- **Activity**. The activity is an element of the method that corresponds to a unitary task (e.g., analysis, data collection, or report). The activities **compose** the method and are organized and coherently articulated with each other (e.g., if-then-else, process elements ordination, etc.). This relation is modeled using an iterative association of a type: activity **follows** activity. The articulation of activities corresponds to the aggregation from [16]. One particular type of activity consists in **generating** resources. For instance: *acquiring a backup tool, maintain the backup tool, etc.*
- **Stakeholder.** A stakeholder is a human, a machine or an organization that is involved in the creation of value at three levels. First, it **performs** the method that generates value (e.g., the risk manager performs a risk analysis); second, it **generates** resources used by the method; and third it **expresses** the value expected after the execution of the method. For example, the *datacenter* is the stakeholder that exploits the redundancy system and the *bank* expresses that it expects availability of the data.
- **Resource.** This element is **a type of** object from the IS that is generated by a stakeholder and that **is used by** an activity composing the value creation method. Resources are typically information and data (e.g., passenger location), but could also consist in computing resources, funding, manpower, etc. For instance, the *backup software* is the resource used by the exploitation of a redundancy system.

Based on the above definitions, the value creation method is modeled in Fig. 2.

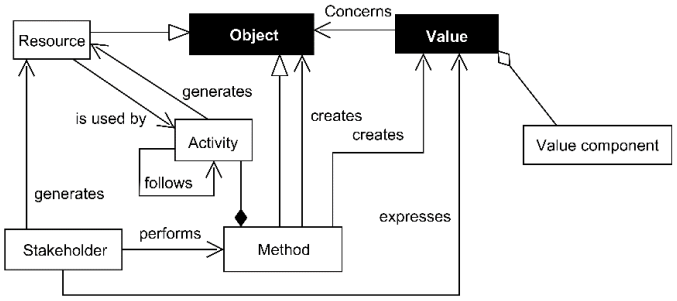


Fig. 2. Value creation method metamodel

3) Dimension 3: Object concerned by the value

The object concerned by the value corresponds to elements (e.g., information, process, tool, or actor) being part of an enterprise. These elements exist in a specific environment represented by the context. This context has an influence on the type and the amount of value associated with this object, for instance, a *customer's browsing history* is an object of a data type that has a particular pecuniary value for an airline travel agency that can estimate the value ascribed to a flight ticket for a customer. This value is calculated based on the number of times this flight ticket is viewed on the company's website by

the customer. At the opposite, this *customer's browsing history* is not an object of value on a drugstore website with fixed prices. Complementarily, it is also worth noting that the context has no impact on the nature of the value. For example, privacy in healthcare is defined in the same way with the same characteristics as in industry.

To collect and deal with the concepts that are necessary to model the object of value, it has been assumed that each sector such as manufacturing, finances, or healthcare, is associated with a specific information system. Each enterprise specific architecture models the objects composing this enterprise as well as the relationships between these objects, using a dedicated language.

Sector-specific information systems and enterprise architecture (EA) models and languages are good approaches here because they semantically define generic objects and sometimes concrete languages to express these objects. Numerous frameworks have been designed to model IS and EA of various sectors, e.g., Cimosa [41], ArchiMate® [42], DoDAF [43], and many others.Regarding the financial case study, the data of the bank's customers is the object concerned by the required privacy (generated by the bank) and concerned by the required availability (generated by the datacenter).

As a summary and according to this analysis, the concepts defining the context and the object concerned by the value are:

- **Information system**. The information system encompasses, and **is composed by,** the objects concerned by the value and the stakeholders that benefit from the value created.
- **Context**. The context represents the surrounding of the IS. It includes (1) the constraints on the system in which the value is created and (2) the definition of the borders of this system (e.g., the sector and the sector purpose of the business entity that is concerned by the IS, the rules and regulations related to the sector or the IS, the institutional arrangements, etc.). Accordingly, the context is associated to the information system with an association named characterizes. As stated in [23], the context also allows selecting the performance components […] necessary to define the scope of the performance evaluation problem. Hence, this selection defines a particular context, or viewpoint, for the evaluation of the value. To model this, the concept of context is associated to the measure with a relation named influence. Regarding the case study in the financial sector, the context is the financial regulation.

Based on the above definitions, the object concerned by the value is modeled in Fig. 3.

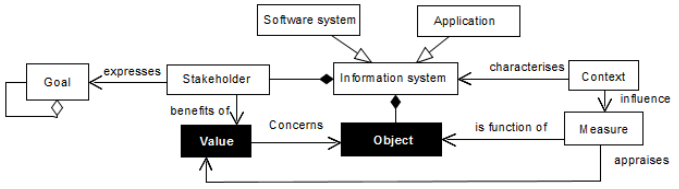


Fig. 3. Object concerned by the value metamodel

III. ARCHIMATE EXTENSION

In this section, we extend ArchiMate language to the VCC domain. Therefore, first we introduce ArchiMate's metamodel, language and extension mechanisms, and finally present the extended ArchiMate to VCC.

A. Introduction to ArchiMate

ArchiMate is a modeling language built on a thorough metamodel for enterprise architecture. It is used by IT architects to design static business and IT views and their links in enterprise architecture endeavors [42]. ArchiMate allows reducing the complexity and proposes means to model and thus better understand the enterprise, and the interconnections and interdependency between the processes, the people, the information, and the systems. Consequently, one objective of ArchiMate is to provide pictures of each enterprise architecture aspects such as the organisational structure, the business processes, the information processing system or the infrastructure. It permits to ensure uniform semantics of the instantiated models but it is not really appropriate to enable quantitative analysis.

One of the underlying assumptions of ArchiMate is to support enterprise architecture for the creation of business value. Relying on ArchiMate's metamodel, each business value is generated by business processes that are supported by applications and infrastructures.

ArchiMate's core is structured in three horizontal layers: the business layer, the application layer and the technology layer. All three layers are built with the same type of concepts and associations. They are structured according to three aspects (vertical layers). The first aspect concerns the active structure elements, which are defined as *entities that are capable of performing behaviour,* e.g., a role or an actor. The second aspect concerns the behavioural elements, which are defined as *units of activity performed by one or more active structure elements*, e.g., a process or a function. The last aspect addresses passive structure elements, which are defined as *objects on which behaviour is performed*, e.g., a contract or an object. ArchiMate metamodel is presented in Figure 4.

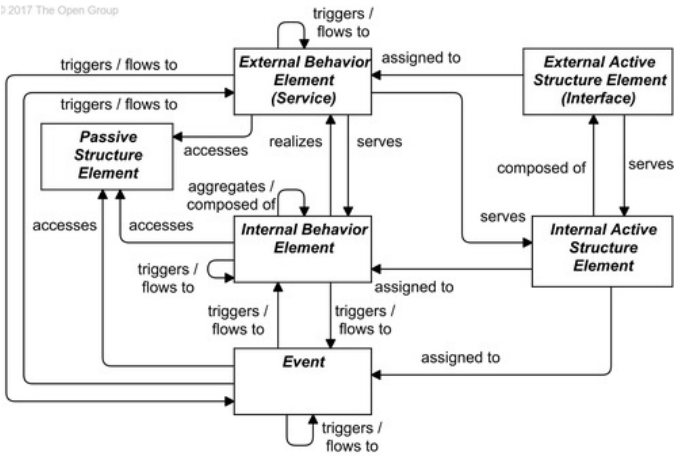
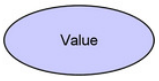
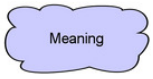
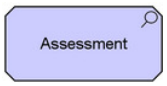


Fig. 4. ArchiMate metamodel (extracted from [10])

B. ArchiMate language

ArchiMate uses a syntax based on symbols and colors, related to the vertical and horizontal layers. Table I provides a sample of ArchiMate elements, definitions and symbols that we later use in the mapping and integration of both metamodels (i.e., ArchiMate's and our previously outlined metamodels).

TABLE I. SAMPLE OF ARCHIMATE SYMBOLS

ArchiMate 3.0 metamodel	Definition	ArchiMate 3.0 metamodel element symbol
Value	Value represents the relative worth, utility, or importance of a core element or an outcome.	
Meaning	The knowledge or expertise present in, or the interpretation given to, a core element in a particular context	
Assessment	An assessment represents the result of an analysis of the state of affairs of the enterprise with respect to some driver.	
Business function	A business function is a collection of business behavior based on a chosen set of criteria (typically required business resources and/or competencies), closely aligned to an organization, but not necessarily explicitly governed by the organization	
Business process	A business process represents a sequence of business behaviors that achieves a specific outcome such as a defined set of products or business services	
Business actor	A business actor is a business entity that is capable of performing behavior	
Resource	A resource represents an asset owned or controlled by an individual or organization	
Capability	A capability represents an ability that an active structure element, such as an organization, person, or system, possesses	
Driver	An external or internal condition that motivates an organization to define its goals and implement the changes necessary to achieve them	

C. ArchiMate extension mechanismes

ArchiMate extension is achieved by integrating its metamodel with the metamodel of the domain that extends it. According to [44], the integration of two metamodels requires resolving three types of heterogeneities: syntactic, semantic and structural. For our integration, only the semantic and the structural heterogeneities have been addressed. In effect, the syntactic heterogeneity aims at analyzing the difference between the serializations of the metamodel. As explained by [45], it addresses technical heterogeneity such as hardware platforms and operating systems, or access methods, or it addresses the interface heterogeneity such as the one which exists if different components are accessible through different access languages. The structural heterogeneity exists when the same metamodel concepts are modelled differently by each metamodel primitives. This structural heterogeneity has been addressed together with the analysis of the conceptual mapping and the definition of the integration rules. Finally, the semantic heterogeneity represents differences in the meaning of the considered metamodel' elements and must be addressed through elements mapping and integration rules. Regarding the mappings, three situations are possible: no mapping, a mapping of a type 1:1, and a mapping of a type n:m (n concepts from one metamodel are mapped with m concepts from the other).

After defining the mapping, the concepts can be integrated in a single metamodel using both ArchiMate' extensions mechanisms: the addition of attribute as well as the specialization [46]. Concretely, if no mappings are detected, the concept from extension domain is added in the ArchiMate using the first extension mechanism, which consists of adding an attribute to an existing concept. If a 1:1 mapping exists without conflict between two concepts, both concepts are merged in a unique one. The resultant concept is added into the integrated metamodel, and this concept keeps the name of the ArchiMate concept. If a mapping of type 1:1 with conflict exists between two concepts, this means that one concept from one metamodel is richer or poorer than a concept from the other metamodel and in this case, both concepts are added in the integrated metamodel using the second extension mechanism of ArchiMate i.e., the stereotype (specialization) (e.g.: [56]).

D. ArchiMate extension to VCC

In this section, the ArchiMate extension mechanisms have been applied to the field of VCC. Table II explains the mapping between elements from the VCC and from the ArchiMate metamodels. Nine VCC elements (as outlined in section B) are mapped with ArchiMate elements (as outlined in section C) and only one VCC element (i.e., the **value component**) has no corresponding ArchiMate element. In effect, although the **value component** from the VCC metamodel could have been mapped to the **value** from the ArchiMate metamodel, we have preferred to keep the semantic difference amongst the **elements of value** and the **value component** from the VCC metamodel in the ArchiMate metamodel. Accordingly, the integration rule that we have exploited to integrate the **value components** with the ArchiMate metamodel is the addition of attribute, and as a result, we have considered that the **value component** is an attribute of the **value**.

Another integration rule that we have used is the merge, i.e., the concept of **value** from the VCC metamodel has been merge with the concept of **value** from the ArchiMate metamodel. This is due the fact that both concepts are defined somewhat equivalently, respectively: *as the degree of worth that concerns something* [which] *improves the well-being of the beneficiary after it is delivered* (VCC metamodel) and as *the relative worth, utility, or importance of a core element or an outcome* (ArchiMate metamodel).

TABLE II. VCC-ARCHIMATE EXTENSION MAPPING

VCC elements	ArchiMate elements	Mapping	Integration rule	Integrated element
Value	Value	1-1	Merge	Value
Nature of the value	Meaning	1-1	Specialization	<<Nature of the value>>
Value component	-	-	Addition of attribute	<<Value>>, Value component: description
Object	Business, Application and Technology layers	1-n	Generalization	Business, Application and Technology layers
Measure	Assessment	1-1	Specialization	<<Measure>>
Activity	Business function	1-1	Specialization	<<Activity>>
Method	Business Process	1-1	Specialization	<<Method>>
Stakeholder	Business actor	1-1	Specialization	<<Stakeholder>>
Resource	Resource and Capability	1-2	Generalization	Resource
Information system	Business, Application and Technology layers	1-n	Generalization	Information system
Context	Driver	1-n	Generalization	Context

We considered four concepts of the VCC metamodel as specialization of concepts from ArchiMate: **nature of the value**, **measure**, **method**, and **stakeholder** in VCC are respectively specialization of **meaning**, **assessment**, **business function** and **business actor** in ArchiMate. For instance, the method is defined as *a property on which calculations can be made for determining the amount of value expected from a value creation method* in VCC metamodel and by *the result of an analysis of the state of affairs of the enterprise with respect to some driver* in ArchiMate metamodel. The second definition is hence more general than the first.

Finally, we considered four concepts of the VCC metamodel as generalization of concepts from ArchiMate: **Object**, **Resource**, **Information system** and **context** in VCC are respectively generalization of elements from the **Business, Application and Technology layers**, **Resource and Capability**, **Business, Application and Technology layers**, and **Motivation** in ArchiMate.

According to the ArchiMate semantic, the VCC concepts may be expressed using the corresponding symbols, as illustrated in Table II

IV. CASE STUDY

The case study presented in the introduction section is illustrated using UML at Figure 5. This figure demonstrates that without an appropriate visual language, the UML model are hardly exploitable by business people having to design new business activity and to co-create new value.

At Figure 6, which model the same case, we illustrate that using the ArchiMate extension provides a much more understandable presentation of our case in terms of clarity and readability.

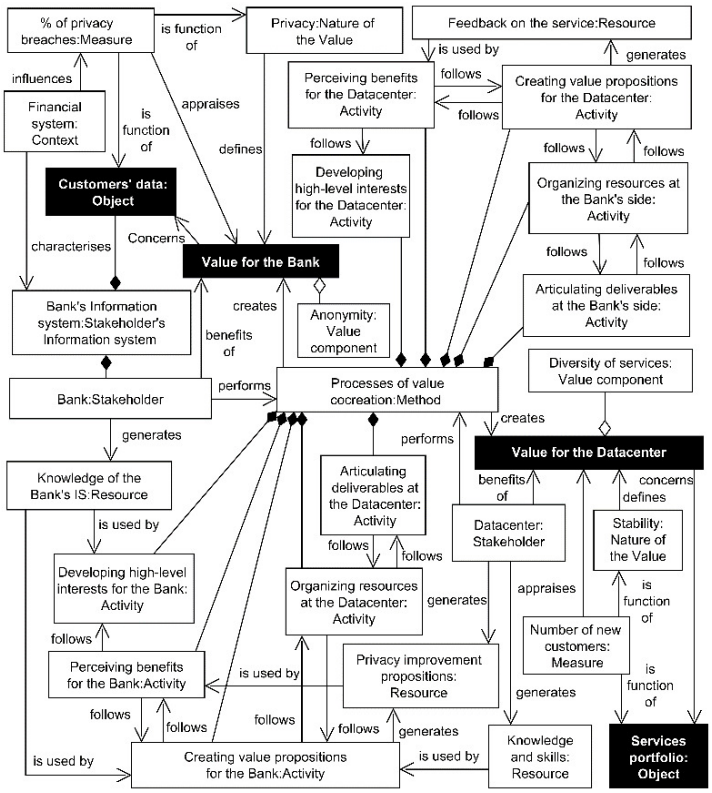


Fig. 5. Value creation perspectives

The advantages are the following:

1. The elements expressed in the model are classified using a code of colors, i.e., **business concepts** are in yellow, **resources** are in orange, **value** related concepts are in purple. These are mainly specialization from the **motivation extension** of ArchiMate, which means that the cocreation of value is something that may be perceived in addition to the **information system** and that motivates the design of elements of this IS.
2. Elements on the figure may also more easily be geometrically organized, e.g. **activity of value cocreation** is on the right-side and **value related elements** are on the left side.
3. Concept reading is facilitate using the shape of the symbols. For instance, **value elements** are rapidly detectable on the model because they are in oval. The **nature of this value** is also easily differentiated because it is presented as clouds.
4. The last advantage is that using ArchiMate also allows us to take advantage of the relationships between concepts semantic. For instance, a task that accesses a resource is illustrated using a dotted line, the association between the activity or the actor that generates the resource is illustrated in dash line, and the generic association is illustrated using a plain line. To improve the semantic of the association, we have specialized it, e.g., the association between the context and the information system has been specialized so that the context <<**characterize**>> the information system.

V. Discussions and Future Works

A. ArchiMate extension

Although ArchiMate extension has already been achieved in many areas such as security [55] and risk management [55], our study conducts such extension in the new field of value cocreation. Concretely, such extension effort resulted in the improved readability of the cocreation instances of the value cocreation metamodel and that all ambiguities have been removed regarding the conceptual semantic.

On the other hand, the most challenging issue is that ArchiMate must be adopted as a common language beforehand, and that all organizations involved in the cocreation have to understand the meaning of the symbol and the language structure, but also that they agree to invest in the usage of the framework.

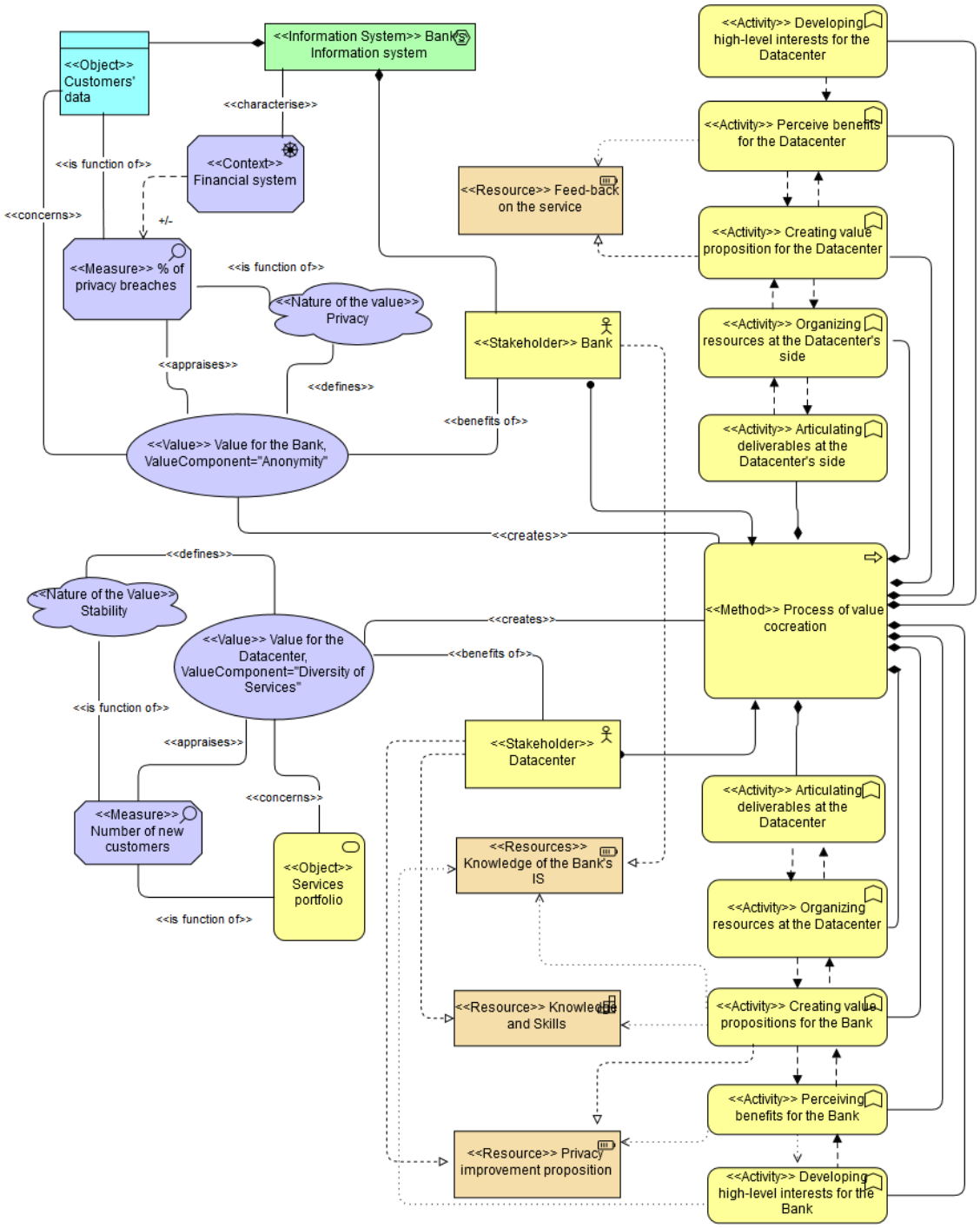


Fig. 6. Value cocreation expressed with ArchiMate

B. *Value perspectives*

ArchiMate has been extended for the field of value cocreation. However, more perspective may be addressed in that VCC domain, such as those illustrated in figure 7:

- The creation of value. This is the most basic but important one. It addresses the method used to valuate an object of the IS, e.g., a privacy impact assessment method that improves the privacy of a database, or a process based method that contributes to the repeatability of the incident management activity of a company. Accordingly, in this first perspective the creation of value is generally expressed based on the three following dimensions: the nature of the value, the object concerned by this value, and the method that creates the value. Preliminary work related to the modeling of the value with ArchiMate were achieved in [54].
- **The method of value creation**. The second perspective considers that the creation of value is a value per se for the company. Hence, the method of value creation may be view as a type of value creation. Example of contribution in this perspective is the method chunk [14] which consists in a type of method of value creation, which in turn, contributes to making an object of the company better off.
- **The value cocreation.** As explained in this paper, the creation of value results sometime to a collaboration between a provider and a client. For instance, a consultant that improves the security of its client's information system collaborates with the client to access the IS architecture, to analyze the value of the business assets to be protected, and to understand the threats. Hence, when a customer collaborates with a provider to generate value, we are in the perspective of value cocreation.
- **The method of value cocreation.** Similar to value creation, the value cocreation may also be perceived as a type of value being cocreated by more than one actor. For instance, a provider and a customer who collaborate for a long time and who analyze, together, how they could cogenerate new value for each other's businesses (like in the case of PowerDrive [3]). Example of processes to support this cocreation mechanism are proposed in [18].

In frames 1 and 3, the (co)created value concerns the creation of value of a concrete nature (e.g. security, privacy, quality,…) and therefore corresponds to a type of value that already has a benefit for company. The value created in both frames 1 and 3 also concerns a concrete object of the IS or of the company. We thus advocate that the value created in both frames corresponds to value-in-use [28].

In frames 2 and 4, we postulate that the created value is the method of value(co)creation itself. This method of value (co)creation is necessary before (co)creating concrete value. In frames 1 and 3 this method is transformed in value-in-use when it is used to (co)create value of a concrete nature. In the frame 4, the value proposition (defined by one actor) is proposed to another actor, which accepts it or not. If accepted, this proposition of value cocreation is transformed in value-in-use when the concrete value is realized through a collaboration among the actors involved.

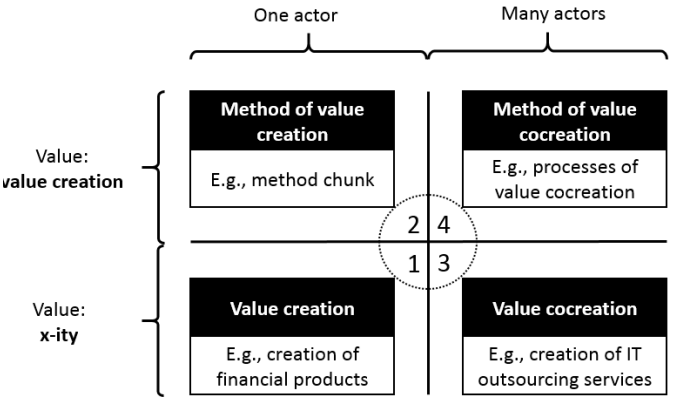


Fig. 7. Value creation perspectives

Provided the similarities among the four perspectives, we claim that perspectives 2, 3 and 4 are specializations of perspective 1. Accordingly, we claim that designing one language for many value creation perspectives is redundant and that our language designed to express the cocreation of value could be specialized to express all perspectives. Therefore, we plan for specializing the ArchiMate extension for the VCC to the four perspectives and validate the expressiveness of these specialization in our future works.

VI. CONCLUSION

This paper has defined a concrete value cocreation language based on the VCC metamodel previously presented in [16, 21, 22]. To define this language, we have extended ArchiMate using its extension mechanism, to know: the specialization and the addition of attributes as explained in [46]. Finally, we have demonstrated the usability of the language with a case in the financial domain.

REFERENCES

[1] S. L. Vargo, R. F. Lusch, "Service-dominant logic: continuing the evolution," Journal of the Academy of marketing Science, vol. 36, no. 1, pp. 1-10, Mar. 2008.

[2] S .L. Vargo, R. F. Lusch, "Evolving to a new dominant logic for marketing," Journal of marketing, vol. 68, no. 1, pp. 1-17, Jan. 2004.

[3] Westergren, U. H.: Opening up innovation: the impact of contextual factors on the co-creation of IT-enabled value adding services within the manufacturing industry. Information Systems and e-business Management, 9(2), 223–245 (2011)

[4] B. Leavy, "Collaborative innovation as the new imperative–design thinking, value co-creation and the power of "pull"." *Strategy & Leadership* 40, no. 2 (2012): 25-34.

[5] C. Calero, J. Ruiz, and M. Piattini, "Classifying web metrics using the web quality model," Online Inf. Review, vol. 29 (3), pp. 227-248, 2005.

[6] C. Feltus, E. Grandry, T. Kupper, and J. N. Colin, Model-Driven Approach for Privacy Management in Business Ecosystem, in 5th International Conference on Model-Driven Engineering and Software Development, 2017.

[7] R. M. Foorthuis, F. Hofman, S. Brinkkemper, and R. Bos, "Assessing business and IT projects on compliance with enterprise architecture," in Procs. of GRCIS, 2009.

[8] A. Dix, "Human-computer interaction: A stable discipline, a nascent science, and the growth of the long tail," Interacting with Computer, vol. 22, no. 1, Jan. 2010. 13-27.

[9] A. Josey, M. Lankhorst, I. Band, H. Jonkers, and D. Quartel, "An Introduction to the ArchiMate® 3.0 Specification," White Paper from The Open Group, Jun. 2016.

[10] G. Berio and F. Vernadat, "Enterprise modelling with CIMOSA: functional and organizational aspects," Production planning & control, vol. 12, no. 2, pp. 128-136, Jan 2001.

[11] A. W. Scheer, and M. Nüttgens, “ARIS architecture and reference models for business process management,” Business Process Management, 2000, pp. 376-389.

[12] M. Langheinrich, “Privacy by design—principles of privacy-aware ubiquitous systems,” in International Conference on Ubiquitous Computing, 2001, pp. 273-291.

[13] A. Cavoukian, "Privacy by design: The 7 foundational principles. implementation and mapping of fair information practices." *Information and Privacy Commissioner of Ontario, Canada*, 2009.

[14] J. Ralyté, “Towards situational methods for information systems development: engineering reusable method chunks, ” in Procs. of 13th International Conference on Information System Development. Advances in Theory, Practice and Education. 2004.

[15] F. Bénaben, J. Touzi, V. Rajsiri, S. Truptil, J. P. Lorré, and H. Pingaud, “Mediation information system design in a collaborative SOA context through a MDD approach,” in Procs. of MDISIS, 2008, pp. 89-103.

[16] C. Feltus, and E. H. A Proper, Conceptualization of an Abstract Language to Support Value Co-Creation, 12th Conference on Information Systems Management (ISM'17), Federated Conferences on Computer Science and Information Systems, Prague, Czech Republic

[17] H. Becker, “Social impact assessment: method and experience in Europe, North America and the developing world,” Routledge, 2014

[18] L. Lessard, C.P. Okakwu, Enablers and Mechanisms of Value Cocreation in Knowledge-Intensive Business Service Engagements: A Research Synthesis. In: 2016 49th Hawaii International Conference on System Sciences, USA. IEEE Computer Society, pp. 1624–1633 (2016)

[19] R. Hevner, S. T. March, and J. Park, “Design science in information systems research,” MIS quarterly, vol. 28, no. 1, 2004. DOI: 10.1007/978-1-4419-5653-8_2

[20] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” Journal of management information systems, vol. 24, no. 3, pp. 45-77, Dec. 2008. DOI: 10.2753/MIS0742-1222240302

[21] C. Feltus, L. Lessard, F. Vernadat, D. Amyot, Erik H.A. Proper, Conceptualization of a Value Cocreation Language for Knowledge-Intensive Business Services, In: E. Ziemba (ed.), LNBiP 311, 2018

[22] C. Feltus, E. HA Proper, A. Metzger, J. F. García López, R. C. González, Value CoCreation (VCC) Language Design in the Frame of a Smart Airport Network Case Study, 32nd IEEE Int. Conf. on Advanced Information Networking and Applications (AINA-2018), Poland.

[23] F. Li, A. Etienne, A. Siadat, F. Vernadat, A Performance Evaluation Methodology for Decision Support in Industrial Projects. In: Proc. 7th IESM conference, htw saar, Germany (2017)

[24] R. Matulevicius, N. Mayer, P. Heymans, Alignment of misuse cases with security risk management. In: 3rd Int. Conf. ARES. IEEE, pp. 1397–1404. (2008)

[25] C. Feltus, M. Petit, E. Dubois, Strengthening employee's responsibility to enhance governance of IT: COBIT RACI chart case study. In: 1stACM Workshop on Information Security Governance. ACM, pp. 23–32 (2009)

[26] C. Calero, J. Ruiz, M. Piattini, Classifying web metrics using the web quality model. Online Inf. Review, 29(3), 227–248 (2005)

[27] R. M. Foorthuis, F. Hofman, S. Brinkkemper, R. Bos, Assessing business and IT projects on compliance with enterprise architecture. In: GRCIS’09. CEUR-WS Vol-459, paper 6 (2009)

[28] S. L. Vargo, P. P. Maglio, and M. A. Akaka, 2008. On value and value co-creation: A service systems and service logic perspective. *European management journal*, *26*(3), pp.145-152.

[29] M. Langheinrich, Privacy by design—principles of privacy-aware ubiquitous systems. In: Ubicomp 2001: Ubiquitous Computing, LNCS, vol. 2201. Springer, pp. 273–291 (2001)

[30] OMG: Value Delivery Metamodel, Version 1.0. OMG Document formal/2015-10-05 (2015)

[31] A. Dix, Human-computer interaction: A stable discipline, a nascent science, and the growth of the long tail. Interacting with Computer, 22(1), 13–27 (2001)

[32] H. Alves, C. Fernandes, M. Raposo, Value co-creation: Concept and contexts of application and study. J. of Business Research, 69(5), 1626–1633 (2016)

[33] A. Cox, Business relationship alignment: on the commensurability of value capture and mutuality in buyer and supplier exchange. Supply Chain Management, 9(5), 410–420 (2004)

[34] J. Nyman, What is the value of security? Contextualising the negative/positive debate. Review of Int. Studies, 42(5), 821–839 (2016)

[35] F. Li, Performance Evaluation and Decision Support for Industrial System Management: A Benefit-Cost-Value-Risk based Methodology. PhD thesis, Arts & Mét. Paritech, France (2017)

[36] V. A. Zeithaml, “Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence,” The journal of marketing, pp. 2-22, Jul. 1988. DOI:10.2307/1251446

[37] I. Manuj, J. T. Mentzer, Global supply chain risk management. J. of Business Logistics, 29(1), pp. 133–155 (2008)

[38] M. Daneva, “Applying real options thinking to information security in networked organizations,” No. TR-CTI. Centre for Telematics and Information Technology, University of Twente, 2006.

[39] E. Ziemba, M. Eisenbardt, R. Mullins, Use of Information and Communication Technologies for Knowledge Sharing by Polish and UK-Based Prosumers. In: E. Ziemba (ed.) Information technology for management: New ideas and real solutions, Lecture Notes in Business Information Processing LNBIP, vol. 277, pp. 49–73 (2017)

[40] M. K. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren (2011) Action design research. MIS Q., 35(1), pp. 37-56, ISSN 0276-7783.

[41] G. Berio, F. Vernadat, Enterprise modelling with CIMOSA: functional and organizational aspects. Production planning & control, 12(2), 2001.

[42] M. M. Lankhorst, H. A. Proper, H. Jonkers, The Architecture of the ArchiMate Language. In: Business-Process and Information Systems Modeling, LNBIP, vol 29, Springer (2009)

[43] U.S. DoD: DoDAF framework, version 2.02 (2010).

[44] C. Parent, and S. Spaccapietra, Database integration: The key to data interoperability. Advances in Object-Oriented Data Modeling, 2000.

[45] S. Zivkovic, H. Kühn, and D. Karagiannis, Facilitate modelling using method integration: An approach using mappings and integration rules. ECIS 2007, pages 2038-2049. University of St. Gallen.

[46] The Open Group. ArchiMate® 2.1 Specification. Van Haren Publishing, The Netherlands. 2012-2013

[47] C. Grönroos, Service logic revisited: who creates value? And who co-creates? European business review, 20(4), 298–314 (2008).

[48] J. D. Chandler and S. L. Vargo, Contextualization and value-in-context: How context frames exchange. Marketing Theory, 11(1), 35–49, (2011)

[49] E. K. Chew, iSIM: An integrated design method for commercializing service innovation. Information Systems Frontiers, 18(3), 457–478 (2016)

[50] M. Blaschke, M. K. Haki, U. Riss, S. Aier, Design Principles for Business-Model-based Management Methods – A Service-dominant Logic Perspective. In: Designing the Digital Transformation (DESRIST 2017), LNCS, vol. 10243. Springer, pp. 179–198 (2017)

[51] J. Gordijn, H. Akkermans, H. Van Vliet, Business modelling is not process modelling. In: ER 2000, LNCS, vol. 1921. Springer, pp. 40–51.

[52] H. Weigand, Value encounters–modeling and analyzing co-creation of value. In: I3E 2009. IFIP Advances in ICT, vol. 305, Springer, Berlin, Heidelberg, pp. 51–64 (2009)

[53] I. S. Razo-Zapata, E. K. Chew, E. Proper, Visual Modeling for Value (Co-)Creation. In: 10th Int. W. on Value Modeling and Business Ontologies, Trento, Italy, paper 6 (2016)

[54] S. de Kinderen, K. Gaaloul, and E. Proper, 2012, February. Integrating value modelling into archimate. In International Conference on Exploring Services Science (pp. 125-139). Springer, Berlin, Heidelberg.

[55] E. Grandry, C. Feltus, E. Dubois, 2013, Conceptual integration of enterprise architecture management and security risk management. In Enterprise Distributed Object Computing Conference Workshops *(EDOCW), 2013 17th IEEE International* (pp. 114-123). IEEE.

[56] C. Feltus, E. Dubois, E. Proper, I. Band, M. Petit, Enhancing the ArchiMate® Standard with a Responsibility Modeling Language for Access Rights Management, 5th ACM International Conference on Security of Information and Networks (ACM SIN 2012), Jaipur, Rajastan, India. ISBN: 978-1-4503-1668-2

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 761–768
DOI: 10.15439/2018F211
ISSN 2300-5963 ACSIS, Vol. 15

An Exploration of BPM Adoption Factors: Initial Steps for Model Development

Renata Gabryelczyk
University of Warsaw
Faculty of Economic Sciences
ul. Długa 44/50,
00-241 Warszawa, Poland
Email: r.gabryelczyk@wne.uw.edu.pl

Abstract — **The main aim of the proposed research is to identify factors that create an environment conducive to successful Business Process Management (BPM) adoption. Factors predicting successful BPM adoption have been identified within the TOE (Technology-Organization-Environment) framework using a literature review and methodology for constructing conceptual frameworks. The following factors are proposed: top management support for previous projects of organizational change, complexity of BPM system and notation, satisfaction with existing systems, business-IT alignment level, perceived strategic benefits of using BPM, extent of coordination, organizational readiness, performance measurement, culture conducive to organizational change, and, perceived environmental pressure. Study results have the potential to fill the research gap by contributing to the development of a theoretical model of BPM adoption that has not been proposed in studies thus far. In practical aspects, the proposed study can influence the understanding of the factors predicting successful BPM adoption.**

I. Introduction

THE main aim of the proposed research is to identify factors that create an environment receptive for Business Process Management (BPM) adoption and allow the prediction of successful adoption and use of this management concept.

BPM has been developing for over 25 years in Information Systems (IS) research [17], [14] and also in management practice [44]. BPM combines the identification, modeling, automation, implementation, control, measurement and improvement of business processes to support organizational goals and increase its effectiveness and efficiency [39]. In each of these BPM areas there are a number of studies, which, however, lack a coherent, theoretical adoption model [23], [16]. This evident gap in BPM research and expected contribution for theory and practice are the main motivations for this proposed study.

The term "adoption", in the context of this research, is defined as the use and acceptance of BPM assumptions in an organization relating to: process-based organizational structures, employee communication, process documentation, process execution, use of IT tools to support implementation and control of processes, performance measurement, process ownership, and taking into account customer requirements [33]. The level of organization involvement in BPM initiatives and programs determines process maturity, i.e., the higher the process maturity of an organization, the higher the level of BPM adoption. However, we can examine closer process maturity only when the organization decides to adopt BPM and starts the first associated initiatives and programs. In the proposed study, we identify technological, organizational and environmental conditions that are conducive to the successful adoption of BPM before the decision on this adoption is taken. Therefore, the BPM adoption model refers to factors that are predictors of the successful use and acceptance of BPM and are conducive to the development of BPM maturity.

BPM adoption factors, to the best of our knowledge, have not been identified thus far. Therefore, they need initially to be understood and they merit in the first step systematic literature review and qualitative approach [43], [20]. Thus, this research aims at the identification of the initial list of BPM adoption factors, based on a systematic literature review on BPM Critical Success Factors and of a methodology of qualitative research for constructing conceptual frameworks by Jabareen [20]. To group BPM adoption factors the TOE (Technology-Organization-Environment) framework was used as one of the most recognized concepts for innovative solutions adoption on the organizational level [41], [4]. The proposed study is an indispensable basis for the initial formulation of research hypotheses for the BPM adoption model.

The BPM adoption model should elucidate what factors predict successful BPM use in an organization, and, as a model to develop a theory, it should provide predictability based on clearly defined assumptions and be precise and falsifiable [37]. These reasons led to the proposed use of the TOE. The TOE framework was introduced by Tornatzky and Fleischer in 1990 [41] to indicate widespread theoretical

This work was supported by the Polish National Science Centre, Poland, Grant No. 2017/27/B/HS4/01734

perspective on factors influencing business innovation in organizations. This framework explains how a technological, organizational, and environmental context influences the adoption and implementation of innovations [41], [4].

BPM as a driver of organizational innovation [36, pp. 3-15] enables the development and implementation of process innovations [32], [40]. Previous research on BPM indicates that not only are resources within an organization important for BPM adoption, but also in the broadly understood context and external environment [44]. There is also strong evidence-based research showing a relationship between Information Systems (and general Information Technology) and BPM in organizations [1], [22], [27], [13]. These arguments led to the use of the TOE framework within a technological, organizational, and environmental context as the theoretical lens of BPM adoption's determinants.

Thus, to realize the main aim of the proposed research three research questions are formulated in this study:

RQ1: Which technological factors are conducive to successful BPM adoption?

RQ2: Which organizational factors are conducive to successful BPM adoption?

RQ3: Which environmental factors are conducive to successful BPM adoption?

This paper will be organized as follows: firstly, the literature underlying the BPM adoption will be presented, followed by a literature review on BPM Critical Success Factors (CSFs) and the factors identified in the research using the TOE framework. Next, the research methodology and obtained results will be presented. Finally, a discussion, contribution, and direction of future research will be also proposed.

II. BACKGROUND

Adoption of BPM

The topic of BPM is widely explored in empirical research [17] and used in organizations primarily to increase organizational effectiveness, prepare organizations to implement IT systems and increase customer satisfaction [29], [42]. One common definition of BPM is proposed in [39, p. 87]: "Business Process Management (BPM) is a discipline involving any combination of modeling, automation, execution, control, measurement, and optimization of business activity flows in applicable combination to support enterprise goals, spanning organizational and system boundaries, and involving employees, customers, and partners within and beyond the enterprise boundaries". This definition was created as a result of a broad discussion of the researchers of the phenomenon and practitioners of BPM implementation, and because it covers both, the BPM technological context (BPM as a technology) and the business context (BPM as a management discipline), it is considered the most comprehensive [39, pp. 87-88].

Despite the great popularity of process-based management concepts and the benefits they bring, BPM is still not adopted as a practice in many organizations, particularly in those of the public sector. It is unclear what causes this lack of acceptance [42]. Moreover the term "adoption" in the context of BPM is seldom used in literature, although it seems to be analogous to the area of Enterprise Resource Planning (ERP)

TABLE I.
BPM CRITICAL SUCCESS FACTORS

BPM CSFs	Source
Top management support, Management involvement, Leadership.	[3] Bai and Sarkis, 2013; [5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015
Information technology, Development of service-oriented business applications and adapting the IT infrastructure, IS support.	[3] Bai and Sarkis, 2013; [5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015; [31] Ravesteyn and Batenburg, 2010; [11] De Bruin and Rosemann, 2005; [38] Skrinjar and Trkman 2013; [42] Trkman, 2010
Strategic alignment, Alignment of processes to organizational goals.	[3] Bai and Sarkis, 2013; [5] Bandara et al, 2009; [7] Buh, Kovacic and Stemberger, 2015; [11] De Bruin and Rosemann, 2005; [38] Skrinjar and Trkman 2013; [42] Trkman, 2010
Governance, Clearly defined process owners, Appointment of process owners.	[7] Buh, Kovacic and Stemberger, 2015; [11] De Bruin and Rosemann, 2005; [42] Trkman, 2010
Methods, Methodology.	[5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015; [31] Ravesteyn and Batenburg, 2010; [11] De Bruin and Rosemann, 2005
Project management, Change Management, Ability to implement the proposed changes.	[3] Bai and Sarkis, 2013; [5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015; [31] Ravesteyn and Batenburg, 2010; [38] Skrinjar and Trkman 2013; [42] Trkman, 2010
Performance measurement, Measurement and control.	[3] Bai and Sarkis, 2013; [5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015; [31] Ravesteyn and Batenburg, 2010; [38] Skrinjar and Trkman 2013; [42] Trkman, 2010
People, Level of employee's specialization, Training and empowerment of employees, Motivated employees.	[5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh, Kovacic and Stemberger, 2015; [11] De Bruin and Rosemann, 2005; [38] Skrinjar and Trkman 2013; [42] Trkman, 2010
Culture, Communication, Teamwork, Social Networks.	[3] Bai and Sarkis, 2013; [5] Bandara, Alibabaei and Aghdasi, 2009; [7] Buh et al., 2015; [11] De Bruin and Rosemann, 2005

systems research [13]. This is particularly notable if we wish to explore factors that can explain the use and acceptance of BPM assumptions in an organization [12], [16].

Previous BPM adoption studies that have used the "adoption" term have focused on success factors or single, selected aspects of BPM implementations and applications. Hribar and Mendling [16] in quantitative research analyzed the role of organizational culture for the successful adoption of BPM. According to this study organizational culture influences the success of BPM adoption. Moreover, this research includes an analysis of the types of organizational culture and indicates which culture types to a greater extent contribute to the success of BPM adoption and the resulting increased performance.

Malinova and Mendling [23] in their qualitative research proposed a conceptual framework for the adoption of BPM indicating possible causes of adoption, action and implementation strategies, and the anticipated effects of this adoption. However, there is no measuring instrument in the study, so it cannot serve as a model for adoption in accordance with the principles of modeling in organizational science [37]. Eikebrokk, Iden, Olsen and Opdahl [12] applied a similar approach to the analysis of factors that influence the acceptance and use of process modeling in organizations: a process modeling acceptance model was developed and tested empirically using survey data from companies.

In general, research on BPM devotes much more attention to BPM after the adoption decision. Studies on the entire BPM life cycle and studies examining different aspects of BPM maturity are definitely dominant. In addition to the afore mentioned studies of the type of culture as a predictor of successful BPM adoption [16] and research on the acceptance of BPM tools [12], literature sources on BPM prediction are virtually non-existent.

BPM Critical Success Factors

In the context of the successful adoption of BPM there have been several studies on critical success factors (CSFs) [3], [5], [7], [11], [31], [38], [42]. Researchers used different terms for identical factors frequently critical to the success of BPM. These terms include: matching processes to organizational strategies, selecting appropriate project management and management methods, supporting top management, using the right information technology, and building a BPM-driven organizational culture.

All these factors are important for understanding the factors behind a successful BPM adoption. However, the adoption model should provide predictability based on clearly defined assumptions and be precise and falsifiable [37]. A BPM adoption model should explain which factors are positive or negative predictors of BPM adoption. For this reason, a review of critical success factors diagnosed after BPM (ex-post) adoption may be a starting point to consider which of these factors can be predicted earlier. For example, if top management support is a BPM critical success factor, then this factor probably also occurred in other organizational change projects implemented in the organization. If top management did not support other projects, it can be assumed that it will not support BPM initiatives and programs. Thus, the review of BPM CSFs studies is the first step to diagnosing prediction factors and for the development of a BPM adoption model, which is missing in BPM research. Table I. presents a summary of a literature review on BPM CSFs.

TOE as a Conceptual Framework for BPM Adoption

To identify and group BPM adoption factors in an organization the TOE framework is applied. The technological context of the TOE concerns the availability of internal and external technologies and new technologies relevant to the organization; the organizational context describes the characteristics of the organization such as communication processes and internal resources; the environmental context refers to the environmental conditions in which the organization operates, e.g. nature and/or strength of competitors and government regulations [25], [4].

The TOE framework takes into account the three aforementioned perspectives important for the adoption of new solutions, and have therefore been chosen as the theoretical basis in various areas of IS research such as cloud computing adoption [2], [6], e-business adoption [25], enterprise resource planning systems (ERP) adoption [26] and e-government assimilation [28]. In BPM research, the TOE framework was used to study BPM software adoption [15]. In order to identify contextual factors a literature review on TOE

TABLE II.
FACTORS AFFECTING INNOVATION'S ADOPTION BASED ON TOE FRAMEWORK

	Factors in research using the TOE framework	Source
Technological Context	Complexity, Compatibility, Satisfaction with existing systems, Technology Competence, Technology readiness, Technology integration.	[2] Alshamaila, Papagiannidis and Li, 2013; [6] Borgman, Bahli, Heier and Schewski, 2013; [9] Chau and Tam, 1997; [15] He and Wang, 2014; [19] Ismail and Ali, 2013; [30] Ramdani, Kawalek and Lorenzo, 2009; [45] Zhu, Kraemer and Dedrick, 2004; [46] Zhu and Kraemer, 2005
Organizational Context	Perceived benefits, Perceived costs, Perceived barriers, Top management support, Organizational readiness, Extent of coordination, Employees knowledge, Financial commitment.	[2] Alshamaila, Papagiannidis and Li, 2013; [6] Borgman, Bahli, Heier and Schewski, 2013; [9] Chau and Tam, 1997; [19] Ismail and Ali, 2013; [21] Kuan and Chau, 2001; [25] Oliveira and Martins, 2010; [28] Pudjianto, Zo, Ciganek and Rho, 2011; [30] Ramdani, Kawalek and Lorenzo, 2009; [46] Zhu and Kraemer, 2005
Environmental Context	Perceived environmental pressure, Market uncertainty, Regulatory policy and support.	[9] Chau and Tam, 1997; [21] Kuan and Chau, 2001; [25] Oliveira and Martins, 2010; [26] Pan and Jang, 2008; [30] Ramdani, Kawalek and Lorenzo, 2009; [47] Zhu, Dong, Xu and Kraemer, 2006

framework applications and on success factors of BPM implementations was conducted. Table II. provides an overview of the factors using in the research with TOE framework. The factors are grouped according to the technological, organizational and environmental context. Factors are listed according to the name under which they occur in the research. However, some names have the same meaning.

III. Research process and Methodology

To identify the TOE factors that allow us to predict successful BPM adoption we used the methodology for constructing conceptual framework by Jabareen [20]. There are two main reasons why we have chosen this methodology. Firstly, this methodology defines framework as an integrated set of factors that enable the theoretical explanation of the studied phenomenon. Thus, according to this definition, the chosen methodology can be used for the preliminary identification of factors. Secondly, it is a methodology of qualitative research that can be based on literature review research, this being the initial step of every researcher as the necessary basis for developing new knowledge and systematizing the existing one [43], [20]. Moreover, the literature review creates a foundation for the development of new models and theories [43], and this is the main aim of proposed research: to create a foundation for the BPM adoption model.

Jabareen's methodology for building conceptual frameworks from existent multidisciplinary literature is a process of theorization [20]. According to the chosen research procedure, we have used in this study the following research steps presented on Figure 1. In the first step of our research process we collected, mapped, and read the literature sources. Our data sources search included literature on BPM in general, especially BPM critical success factors. The literature regarding the current applications of the TOE framework in the research on the adoption of new innovative solutions was also crucial.

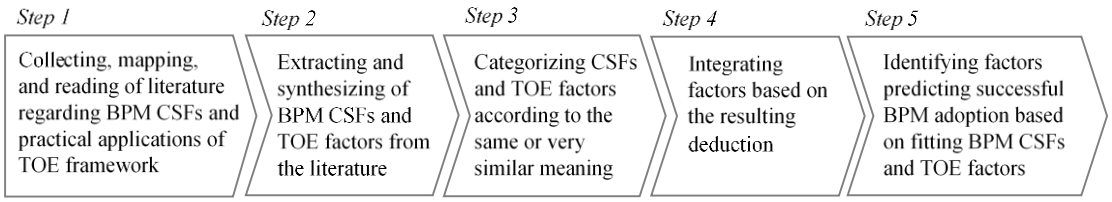


Fig. 1 Research process to identify initial list of factors predicting successful BPM adoption

In the second step of our research process (Fig. 1), we reviewed and synthesized the critical success factors for BPM, as presented in Table I. Also, the factors used in the TOE studies were collected and grouped by technological, organizational, and, environmental context. Findings of this research step are included in Table II.

In the next, third step, we categorized the factors found in the literature, both BPM CSFs and those used in the TOE applications according to the same or very similar meaning. Results of this investigation present first and second column in Table III. In the next, key stage of the study (step 4 on Fig. 1), as a result of the subsequent deduction, we combined factors of the same meaning, while reducing their number. For example, the key success factor of strategic alignment shows that the organization can indicate the impact of processes on the implementation of strategic objectives. Thus, this factor allows an assessment of the expected costs and benefits of BPM implementation. Factors matching the concept in the framework of the TOE are: perceived benefits, perceived costs, perceived barriers. These factors can all be measured with the help of a factor "perceived strategic benefits of using BPM". This approach allows us to identify in the fifth step an initial list of factors predicting successful BPM adoption. This approach resulted from the most common objective for literature review being a combination of past literature aiming at "formulating general statements that characterize multiple specific instances of research, methods, theories, or practices" [10, p. 4].

IV. Results and Discussion

Technological Context of BPM Adoption

Technological factors of BPM adoption can refer to the information technologies that are dedicated to modeling, analysis, simulation, automation, and process management in general. "Complexity", defined as "the degree to which an innovation is perceived as relatively difficult to understand and use" [34, p. 257] is considered as one of the fundamental factors that adversely affect adoption in many past IT adoption studies [30], [19]. Based on results of the BPM software adoption study [15], we suggest the same relationship for BPM in general, i.e. that the complexity of a BPM system and notation has a negative effect on the BPM adoption.

The adoption of BPM can also be affected by other relationships between processes and information technology (IT), such as Business-IT alignment [22], [42] or the adoption of ERP systems, where a process-driven approach is commonly used in system implementations [27], [13]. Reference [9] note that the satisfaction level with existing IT systems plays a significant role in the shaping of motivations to change. Satisfaction with existing processes that are automated in ERP systems, or, satisfaction with service delivery processes through IT, may discourage organizational

TABLE III.
FITTING OF BPM CSFS AND TOE FACTORS

BPM CSFs	TOE factors	TOE factors for BPM
Top management support, Management involvement, Leadership	Top management support	Top management support for previous projects of organizational change
Information technology, Development of service-oriented business applications and adapting the IT infrastructure, IS support	Complexity / Compatibility Satisfaction with existing systems Technology competence / readiness / integration	Complexity of BPM system and notation Satisfaction with existing systems Business-IT alignment level
Strategic alignment, Alignment of processes to organizational goals	Perceived benefits / costs / barriers / Financial commitment	Perceived strategic benefits of using BPM
Governance, Clearly defined process owners, Appointment of process owners	Extent of coordination	Extent of coordination
Methods, Methodology	Organizational readiness	Organizational readiness
Project management, Change Management, Ability to implement the proposed changes	Organizational readiness	Organizational readiness
Performance measurement, Measurement and control		Performance measurement
People, Level of employee's specialization, Training and empowerment of employees, Motivated employees	Employees knowledge	Culture conducive to organizational changes
Culture, Communication, Teamwork, Social Networks		Culture conducive to organizational changes
	Perceived environmental pressure / market uncertainty / Regulatory policy and support	Perceived environmental pressure

changes. Thus, the next technological factor concerns the Business-IT alignment level: the higher the technological factor the more positive the impact on BPM adoption. However, satisfaction with existing systems has a negative effect on the BPM adoption. The more an organization is satisfied with its systems, the less it wants to change the processes to which it is accustomed.

Organizational Context of BPM Adoption

The TOE organizational context refers to the characteristics, structures, processes, and resources of an organization that may constrain or facilitate the adoption of innovation [9], [19]. The first factor takes into consideration the perceived strategic benefits of BPM. Awareness of benefits such as efficiency, effectiveness, and agility [35] can be a basic driver of a decision to adopt. Therefore, the hypothesis is formulated as follows: perceived benefits of using BPM have a positive effect on the BPM adoption.

Top management support is one of the most commonly mentioned CSFs, not only for BPM, but also for all organizational change projects. Effective decisions, monitoring and promoting acceptance of the project, and general change of management from the top, are crucial for a successful BPM adoption [1], [3]. The supporting role of top management in the previous change projects can be also a positive predictor for BPM adoption. Therefore, top management support has a positive effect on BPM adoption.

Organizational readiness is defined in research on innovation adoption as "the availability of the needed organizational resources for adoption" [18, p. 467], [2], [19]. Of particular importance is the perceived assessment by managers on the financial resources held by the organization and the organizational competence to undertake the adoption. This viewpoint confirms studies on BPM CSFs [3], [31]. The following statement is therefore suggested: organizational readiness has a positive effect on BPM adoption.

Coordination mechanisms can take the form of "processes, roles, or structural arrangements […] as teams, informal linking roles, like those of change agents" [8], [28]. This factor seems highly important in the context of the development of BPM governance that establishes appropriate and transparent roles and responsibilities for BPM [11]. The resulting factor is the extent of coordination. The use of a coordination mechanism can have a positive effect on BPM adoption.

Performance measurement does not exist in previous studies using the TOE framework. However, the need to measure the effectiveness of the organization and its processes, and the inclusion of process performance measurement for continuous process improvement, are essential to the high level of BPM adoption and organizational maturity [31], [42]. The need for performance measurement results from strategic considerations and fosters the adoption of BPM. This result in the identification of the factor investigating the use of performance measurement, what can have a positive effect on BPM adoption.

Factors regarding organizational culture have not been mentioned so far in research using the TOE framework.

Probably because it was difficult to measure the impact of this factor in the context of adopting new technologies and innovations. However, in the area of BPM research there are strong arguments in the work of Hribar and Mendling [16] indicating the type of organizational culture that creates an environment for BPM. The results of these tests will allow to build an appropriate measurement instrument and thus to include the cultural factor in the BPM adoption model. BPM's adoption is very strongly linked to cultural and human aspects [11], [16] and so we believe that this factor must be included in the model.

Environmental Context of BPM Adoption

BPM adoption can be the result of pressure exerted on an enterprise by its environment or external circumstances. Pressure can be exerted by business partners, competitors or government policies [21], [14]. A study of McCormack and Johnson [24] indicates that BPM maturity in a least advanced organization determines the level of cooperation and adaptation of inter-organizational processes. External conditions can also force the adoption of BPM in an organization, when, for example, the improvement and development of internal processes is forced upon it as the result of feedback from customers and suppliers [14], [38], [7]. In summary, perceived environmental pressure can have a positive effect on BPM adoption.

V. Conclusion

In this research we explored the factors that create environments receptive or unreceptive to BPM adoption and use. Factors have been identified within the TOE framework and grouped by the technological, organizational and environmental context.

The method of literature review and its qualitative analysis was used to identify factors. Table IV. presents the results of the study. As indicated in the title, this research paper presents initial steps for BPM model development. A review of prior literature and identifying of preliminary list of factors that can predict successful BPM adoption creates a foundation for future research and facilitates theory and model development.

The identified list of factors is the foundation for the BPM adoption model which has not been proposed in studies so far.

VI. Future Research

The initial identification of BPM adoption factors presented in this research is the basis for the development of the adoption model. However, as such factors have not yet been investigated, further studies are needed to verify the initial proposed list of factors. Subsequent research should be of qualitative and quantitative research. In order to verify factors derived from the literature review and perhaps adding new ones, a multiple case study method could be applied. This multiple case study analysis is considered the most suitable in examining organizations that adopted BPM within the real-life context. We plan to use a qualitative study with the aim of developing relevant hypotheses for future quantitative research on the phenomena. Thus, to develop BPM adoption model we plan further research aims: to examine identified factors in a qualitative study, and then formulate and test research hypothesis based on qualitative research. To develop a model to explain BPM adoption is our target aim.

Table IV.
BPM Adoption Factors Based on TOE Framework

	TOE factors for BPM adoption
Technological Context	Complexity of BPM system and notation Satisfaction with existing systems Business-IT alignment level
Organizational Context	Top management support for previous projects of organizational change Perceived strategic benefits of using BPM Extent of coordination Organizational readiness Performance measurement Culture conducive to organizational changes
Environmental Context	Perceived environmental pressure

VII. Contribution

The results of the proposed research could contribute to the development of a consistent theoretical model that would include the various factors influencing the successful adoption of BPM and thus contribute to the theory development. The methodological approach utilizing the TOE framework as the basis of the adoption model is novel in BPM research.

The proposed initial list of BPM adoption factors may provide the foundation for further research. This list can be developed and modified using other data sources and types of research mentioned in the section about future research.

The exploration of BPM adoption factors can contribute to the development of both individuals and entire organizations, and in both the public and private sectors. For individual employees and managers, it will be possible to raise awareness of BPM and identify gaps in competency delaying the adoption of BPM. At an organizational level, the model will help streamline organizational planning and resource development in all areas of the TOE framework. Knowledge about factors influencing successful BPM adoption can help predict the effects of BPM application in organizations that are less mature. A high level of BPM adoption allows an increase in the efficiency of processes carried out for citizens in the public and customers in the private sector, thus benefiting a country as a whole.

References

[1] Al-Mudimigh, A. S. (2007). "The role and impact of business process management in enterprise systems implementation." Business Process Management Journal 13 (6), 866-874.

[2] Alshamaila, Y., S. Papagiannidis, and F. Li (2013). "Cloud computing adoption by SMEs in the north east of England: A multi-perspective framework." Journal of Enterprise Information Management 26 (3), 250-275.

[3] Bai, C. and J. Sarkis (2013). "A Grey-Based DEMATEL Model for Evaluating Business Process Management Critical Success Factors." International Journal of Production Economics 146 (1), 281-292.

[4] Baker, J. (2012). "The technology-organization-environment framework." In: Dwivedi, Y. K., M. R. Wade, and S. L. Schneberger (eds.) Information Systems Theory. New York: Springer, pp. 231-245.

[5] Bandara, W., A. Alibabaei and M. Aghdasi (2009). "Means of achieving Business Process Management success factors." In: Proceedings of the 4th Mediterranean Conference on Information Systems, 25-27 September 2009 , Athens University of Economics and Business, Athens.

[6] Borgman, H. P., B. Bahli, H. Heier and F. Schewski (2013). "Cloudrise: exploring cloud computing adoption and governance with the TOE framework." In: Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS). IEEE, pp. 4425-4435.

[7] Buh, B., A. Kovacic and M. I. Stemberger (2015). "Critical Success Sactors for Different Stages of Business Process Management Adoption – a Case Study." Economic Research-Ekonomska Istraživanja 28 (1), 243-258.

[8] Chatterjee, D., R. Grewal and V. Sambamurthy (2002). "Shaping up for e-commerce: institutional enablers of the organizational assimilation of web technologies." MIS Quarterly 26 (2), 65-89.

[9] Chau, P. Y. K. and K. Y. Tam (1997). "Factors affecting the adoption of open systems: an exploratory study." MIS Quarterly, 1997, 21 (1), 1-24.

[10] Cooper, H. and L. V. Hedges (2009). "Research synthesis as a scientific process" In: Cooper, H., L. V., Hedges and J. C. Valentine (Eds.). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

[11] De Bruin, T. and M. Rosemann (2005). "Towards a Business Process Management Maturity Model." In: *Proceedings of the Thirteenth European Conference on Information Systems, ECIS*, 26-28 May 2005, Germany, Regensburg.

[12] Eikebrokk, T. R., J. Iden, D. H. Olsen and A. L. Opdahl (2011). "Understanding the determinants of business process modelling in organisations." Business Process Management Journal 17 (4), 639-662.

[13] Gabryelczyk, R. and N. Roztocki (2017). "Effects of BPM on ERP Adoption in the Public Sector." In: Proceedings of the 23rd Americas Conference on Information Systems (AMCIS). Boston, USA: AIS Library. URL: http://aisel.aisnet.org/amcis2017/AdvancesIS/Presentations/14/

[14] Gabryelczyk, R. and N. Roztocki (2018). "Business Process Management Success Framework for Transition Economies." Information Systems Management 35(3), 234-253.

[15] He, Y. and W. Wang (2014). "BPM software adoption in enterprises based on TOE framework and IS success model." Computer Modelling & New Technologies 18 (12C), 195-200.

[16] Hribar, B. and J. Mendling (2014). "The Correlation of Organizational Culture and Success of BPM Adoption." In: Proceedings of the European Conference on Information Systems (ECIS). Tel Aviv, Israel: AIS Library. URL: http://aisel.aisnet.org/ecis2014/proceedings/track06/2

[17] Houy, C., P. Fettke and P. Loos (2010). "Empirical Research in Business Process Management – Analysis of an Emerging Field of Research." Business Process Management Journal 16 (4), 619-661.

[18] Iacovou, C. L., I. Benbasat, and A. S. Dexter (1995). "Electronic data interchange and small organizations: Adoption and impact of technology." MIS Quarterly 19 (4), 465-485, p. 467.

[19] Ismail, W. and A. Ali (2013). "Conceptual Model for Examining the Factors that Influence the Likelihood of Computerised Accounting Information System (CAIS) Adoption Among Malaysian SME." International Journal of Information Technology and Business Management 15 (1), 122-151.

[20] Jabareen, Y. (2009). "Building a Conceptual Framework: Philosophy, Definitions, and Procedure." International Journal of Qualitative Methods, 8(4), 49-62.

[21] Kuan, K. K., and P. Y. Chau (2001). "A perception-based model for EDI adoption in small businesses using a technology–organization–environment framework." Information & management 38 (8), 507-521.

[22] Luftman, J. (2000). "Assessing Business-IT Alignment Maturity". Communications of the Association for Information Systems 4 (14). AIS Library. URL: http://aisel.aisnet.org/cais/vol4/iss1/14

[23] Malinova, M. and J. Mendling (2013). "A Qualitative Research Perspective on BPM Adoption and the Pitfalls of Business Process Modeling." Lecture Notes in Business Information Processing, LNBIP 132, 77–88. Berlin Heidelberg: Springer.

[24] McCormack, K. P. and W. C. Johnson (2001). Business process orientation: Gaining the e-business competitive advantage. CRC Press.

[25] Oliveira, T. and M. F. Martins (2010). "Understanding e-business adoption across industries in European countries." Industrial Management & Data Systems 110 (9), 1337-1354.

[26] Pan, M. J., and W. Y. Jang (2008). "Determinants of the adoption of enterprise resource planning within the technology-organization-environment framework: Taiwan's communications industry." Journal of Computer Information Systems 48 (3), 94-102.

[27] Panayiotou, N.A., S. P. Gayialis, N. P. Evangelopoulos and P. K. Katimertzoglou (2015). "A business process modeling-enabled requirements engineering framework for ERP implementation." Business Process Management Journal 21 (3), 628-664.

[28] Pudjianto, B., H. Zo, A. P. Ciganek and J. J. Rho (2011). "Determinants of e-government assimilation in Indonesia: An empirical investigation using a TOE framework." Asia Pacific Journal of Information Systems 21 (1), 49-80.

[29] Ram, J. and D. Corkindale (2014). "How 'critical' are the critical success factors (CSFs)?: Examining the role of CSFs for ERP." Business Process Management Journal 20 (1), 151-174.

[30] Ramdani, B., P. Kawalek and O. Lorenzo (2009). "Predicting SMEs' adoption of enterprise systems." Journal of Enterprise Information Management 22 (1/2), 10-24.

[31] Ravesteyn, P. and R. Batenburg (2010). "Surveying the Critical Success Factors of BPM-Systems Implementation." Business Process Management Journal 16 (3), 492-507.

[32] Recker J. (2015). "Evidence-Based Business Process Management: Using Digital Opportunities to Drive Organizational Innovation." In: vom Brocke J. and T. Schmiedel (eds). BPM - Driving Innovation in a Digital World. Management for Professionals. Cham: Springer, pp. 129-143.

[33] Reijers, H.A. (2006). "Implementing BPM systems: the role of process orientation." Business Process Management Journal 12 (4), 389-409.

[34] Rogers, E. M. (2003). Diffusion of innovations. Fourth Edition. New York: Free Press. p. 257.

[35] Rudden, J. (2007). "Making the Case for BPM-A Benefits Checklist." BPTrends January 2007.

[36] Schmiedel, T. and J. vom Brocke (2015). "Business Process Management: Potentials and Challenges of Driving Innovation." In: vom Brocke J. and T. Schmiedel (eds). BPM - Driving Innovation in a Digital World. Management for Professionals. Cham: Springer, pp. 3-15.

[37] Shapira, Z. (2011). "I've Got a Theory Paper—Do You?: Conceptual, Empirical, and Theoretical Contributions to Knowledge in the Organizational Sciences." Organization Science 22 (5), 1312–1321.

[38] Skrinjar, R. and P. Trkman (2013). "Increasing Process Orientation with Business Process Management: Critical Practices." International Journal of Information Management 33 (1), 48-60.

[39] Swenson, K. D. and M. von Rosing (2015). "Phase 4: What Is Business Process Management?" In: von Rosing M., H. von Scheel and A.-W. Scheer (eds.). The Complete Business Process Handbook. Body of Knowledge from Process Modeling to BPM. 1st Edition. Morgan Kaufmann, p. 87.

[40] Tarafdar, M. and S. R. Gordon (2007). "Understanding the influence of information systems competencies on process innovation: A resource-based view." The Journal of Strategic Information Systems 16 (4), 353-392.

[41] Tornatzky, L.G. and M. Fleischer (1990). The Processes of Technological Innovation. Lexington Books.

[42] Trkman, P. (2010). "The Critical Success Factors of Business Process Management." International Journal of Information Management 30 (2), 125-134.

[43] Webster, J. and R. T. Watson (2002). "Analyzing the past to prepare for the future: Writing a literature review." *MIS Quarterly*, xiii-xxiii.

[44] vom Brocke, J., S. Zelt and T. Schmiedel (2016). "On the Role of Context in Business Process Management." International Journal of Information Management 36 (3), 486-495.

[45] Zhu, K., K. L. Kraemer and J. Dedrick, (2004). "Information technology payoff in e-business environments: An international perspective on value creation of e-business in the financial services industry." Journal of Management Information Systems, 21(1), 17-54.

[46] Zhu, K. and K. L. Kraemer (2005). "Post-adoption variations in usage and value of e-business by organizations: cross-country evidence from the retail industry." Information Systems Research, 16(1), 61-84.

[47] Zhu, K., S. Dong, S. X. Xu, and K. L. Kraemer (2006). "Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies." European Journal of Information Systems 15 (6), 601-616.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 769–778
DOI: 10.15439/2018F336
ISSN 2300-5963 ACSIS, Vol. 15

MCDA-based Approach to Sustainable Supplier Selection

Artur Karczmarczyk*, Jarosław Wątróbski†, Grzegorz Ladorucki† and Jarosław Jankowski*
*Faculty of Computer Science and Information Technology
West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
Email: {artur.karczmarczyk,jaroslaw.jankowski}@zut.edu.pl
†Faculty of Economics and Management
University of Szczecin
Mickewicza 64, 71-101 Szczecin, Poland
Email: {jwatrobski, gladorucki}@wneiz.pl

***Abstract*—The process of sustainable supplier selection is crucial to a company's business continuity. Distortions in poorly chosen suppliers can lead to an impediment or even complete downtime of the company's operations. The paper proposes a new unique approach in which classical MCDA paradigm is extended with aspects of temporal evaluation and various temporal aggregation strategies are provided. The partial MCDA evaluations are performed with three MCDA methods – AHP, TOPSIS and COMET – to allow for hierarchical structuring of the decision problem, creation of a reference model and avoiding rank reversal. The proposed approach is verified on a case study with an actual company and its supplier selection from a group of 30 potential suppliers.**

I. Introduction

SINCE the very beginning of the supply chain thinking, it has been understood that the selection of a proper supplier is the pillar of developing a competitive supply chain [1]. It is a strategic decision, which can considerably affect the company's competitive advantage [2]. The consequences of such decision can be intensified even further if a company plans expansion to new markets. A successful selection of decent suppliers can positively affect the company productivity and effectiveness, as well as decrease the operational costs [3]. Moreover, an apt arrangement of a supply base is crucial for effective and efficient materials and products logistics. Therefore, such selection directly affects the company's business continuity [2].

Integrated relationships between supply chain partners drive the supply chain efficiency, however, if over-dependence occurs, it might lead to propagation and amplification of various disruptions [4]. There are multiple works studying the negative effects of incompleteness or negligence of the supplier evaluation processes. In the early studies, Meade pointed out that wrong selection of suppliers can have negative effects on multiple processes within an organization [5], whereas the subsequent works expanded the negative consequences over the full integrated supply chain (Supply Chain Management - SCM) [6]. Moreover, Chan points out that the negligence of the process of suppliers evaluation can lead not only to disruptions in the supply chain process, but also to ceasing the primary operations of a company at all [6].

The development of the Internet and ICT (Information and Communication Technologies) lead to a considerable increase of the data processing efficiency. As a result, an on-going monitoring and evaluation of suppliers was made possible. Moreover, nowadays, such evaluation can easily be performed repeatedly over a span of time. The development of modern smart management information systems allowed the evaluation of suppliers on a temporal level, thus opening new research areas. From the methodological and practical points of view, an on-going monitoring – as opposed to a one-time evaluation – of the quality of services and products supplied by the company's key supply partners becomes an interesting research problem. Such evaluation takes into account an important, yet often overlooked, aspect of the appraisal changeability in time. For example, the businesses such as e-commerce websites can have as much as 30% of raise or loss in transactions count between the busy December and slow summer holiday months [7]. On the other hand, for the farming industry the summer months are the peak of the fruit picking season and the companies are looking for additional outsourcing suppliers to cope with the increased amount of work [8].

Over the span of the last two decades, the studies about environment protection have been increasingly gaining importance in the world [9]. The increased environmental awareness has lead to pressures from various stakeholders [10] for the companies to realize the significance of incorporating the green practices into their daily operations [11]. Therefore, the evaluation criteria used in the previous decades [2] needed to be expanded to support the evaluation of the green supply chain management (GSCM) practices [10], [12].

Multi-criteria decision analysis (MCDA) methods have been successfully applied in such evaluation problems to find "good" (but not optimal) solutions. However, the exclusive use of an MCDA method provides the "here and now" evaluation of the suppliers, yet it does not take into account the temporal validity of the aggregated data and the partial evaluations. The authors attempted to address this problem in their prior work (see [13]). However, the evaluations obtained with the most popular MCDA methods may be not fully reliable, because many of them are prone to the rank

reversal phenomenon, which means that introduction of a new supplier to the evaluation process can reverse the ranks of the other, unrelated, suppliers. Therefore, the authors expand their approach presented in [13] and their contribution in this paper is to provide a framework for dynamic MCDA-based sustainable supplier selection, which takes into account the temporal aspects, the hierarchical structure of the decision problem, as well as it studies the full space of the decision problem, thus preventing the rank-reversal phenomenon. In practical terms, the introduction of the temporal aspects to the MCDA-based evaluation means introduction of a set of time-anchored MCDA-based models and providing the mechanisms of their aggregation according to the DM's needs.

The rest of this paper is organized as follows. In section 2 a literature review regarding the current state of art is provided. Section three presents the methodological framework. An empirical research and its results are presented in section four. The conclusions and future work directions are presented in section 5.

II. Literature Review

The supplier selection is a process of a significant strategic importance for all the parties involved. The literature analysis provides a wide spectre of theoretical solutions and practical studies where the authors evaluate and select the suppliers in the supply chain. A numerous set of studies has been performed in the areas of electronics industry [10], [11], automotive [9], [15], [16], manufacturing [18], or food supply chain [17], [20] to name just a few.

Nowadays, due to the increased awareness of the environmental issues, the sustainable supplier development has become a necessity, as companies increasingly compete on the ground of having green supply chain capabilities [9]. Many companies struggle with the eco-friendly supplier selection, yet the advancements in the green supply chain management practice strategies can help in this selection [11]. The literature review of the research methodologies shows that the evaluation process is performed with the use of numerous analytical methods. A profound discussion of the approaches can be found in [21]. It can be also noted, that MCDA methods are becoming increasingly popular in such type of applications. Moreover, in some studies fuzzy variants of the crisp MCDA methods, as well as hybrid solutions are used.

Table I presents some of the recent applications of the MCDA methods in sustainable supplier selection. Kannan et al. [10] used the fuzzy TOPSIS method on a set of criteria based on green supply chain management (GSCM) practices to select green suppliers for a Brazilian electronics company. Similarly, Uppala et al. [11] used a hybrid approach of fuzzy AHP and fuzzy TOPSIS to select green suppliers for an Indian electronics company. For the same kind of industry in Taiwan, Chateterjee et al. [12] used a hybrid set of DEMATEL, ANP and MAIRCA methods with 15 criteria in 5 dimensions. Razaei et al. [2] used BWM (the Best-Worst Method) to evaluate 34 suppliers in edible oils industry for a company seeking to expand to a new country, whereas Banaeian et al. [17] used the fuzzy variants of TOPSIS, VIKOR and GRA for a green supplier selection for an actual company from agri-food sector in Iran. Akman [9] used VIKOR and fuzzy c-means clustering to evaluate 198 automotive industry suppliers in Turkey, based on 4 performance and 9 green criteria. Govindan et al. [20] used a mixture of Fuzzy TOPSIS, Fuzzy AHP and Fuzzy SAW for green supplier selection and order allocation in a low-carbon paper industry in India. A more comprehensive literature review of MCDA methods usage for the green supplier evaluation and selection can be found in [22], [23].

It is important to note, however, that the aforementioned MCDA-based approaches produce an assessment based on criteria measurements collected for a single moment in time. In case of the supplier selection problem, it is often required to consider the variability of each suppliers' evaluations in time. There have been some efforts to extend the MCDA methods to provide the ability to aggregate measurements and evaluations collected over a period of time. Banamar and De Smet [24] extended the PROMETHEE II method to allow temporal evaluations. Sahin and Mohamed [25] introduced a Spatial Temporal Decision framework, based upon a combination of System Dynamics modelling, Geographical Information Systems modeling and multi-criteria analyses of stakeholders' views with the use of the AHP method. Zhu and Hipel [26] used multiple stages grey target decision making method for vendor evaluation of a commercial airplane in China. Arasteh et al. [27] used the Goal Programming MCDA method to consider a 6-project portfolio over five investment periods and compared the use of their model in fuzzy and crisp scenarios. Last, but not least, a framework extending the TOPSIS method capabilities to evaluate and select green suppliers based on temporal analysis has been constructed [28], [13]. However, the latter approach still did not take into account the hierarchical structure of the decision problem, nor the rank reversal problem. Thus, the performed literature review allows to observe an interesting research gap of the sustainable suppliers selection problem which would simultaneously take into account the decision problem hierarchical structure, temporal aspects of the evaluation as well as protect the produced outcome from the rank reversal phenomenon.

III. Methodological Framework

The selection of a sustainable supplier in the Green Supplier Chain Management is a complex problem that requires a proper approach. The popular MCDA-based approaches are not without shortcomings. They are based on the classic MCDA paradigm, where constancy of all the elements of the decision support process is assumed. It should be noted, however, that the process of sustainable supplier evaluation requires taking into account its characteristics - its hierarchical structure as well as changeability of the appraisal elements in time. Based on the above, the authors propose using a complementary approach based on precise mapping of the structure of the decision problem (derived from AHP), building a supplier reference model (TOPSIS), as well as minimizing

TABLE I
MCDA METHODS APPLICATION IN THE SUSTAINABLE SUPPLIER SELECTION PROBLEM

Ref	MCDA Methods	Hybrid	Sens. Anal.	Application	Country	Criteria	Suppliers
[14]	ANP, GRA	yes	no	automotive	Iran	6	5
[15]	no	no	no	automotive	Malaysia		153
[16]	AHP	no	yes	automotive	Pakistan	4	3
[9]	VIKOR, fuzzy c-means clustering	yes	no	automotive	Turkey	13	198
[17]	Fuzzy TOPSIS, VIKOR, GRA	yes	no	edible oils	Iran	4	10
[2]	BWM (best worst method)	no	no	edible oils	new country	37	34
[10]	Fuzzy TOPSIS	no	yes	electronics	Brasil	GSCM	12
[11]	Fuzzy AHP and Fuzzy Topsis	yes	no	electronics	India	GSCM	10
[12]	DEMATEL, ANP, MAIRCA	yes	yes	electronics	Taiwan	15	
	Fuzzy TOPSIS, Fuzzy AHP, Fuzzy SAW	yes	no	low-carbon paper	India	5	4
[18]	Fuzzy AHP	no	yes	manufacturing	global	25	2
[19]	AHP, Fuzzy AHP, TOPSIS, Fuzzy TOPSIS, IRP and weighted IRP	yes	no	various	India, Germany, Switzerland	24	41

the shortcomings of the two methods by incorporating the COMET method. Moreover, the characteristics of the sustainable supplier evaluation process requires taking into account the variable effect of each supplier appraisal in a period of time. Therefore, the authors propose using time-conditioned evaluation aggregation strategies. The framework is visually presented on Fig. 1 and is described in detail in the following subsections.

A. MCDA Foundations of the Proposed Framework

The problem of sustainable supplier selection is a multi-criteria problem, since it requires to take into consideration multiple, not only performance but also environmental, criteria. For example, Rezaei et al. [2] provided a list of 23 supplier selection criteria most utilised in the periods 1966-1990 and 1990-2001, and combined them with 15 modern environmental criteria. However, such a vast set of criteria makes the evaluation difficult to perform. Therefore, in the proposed approach we utilise the AHP method to organize the evaluation criteria in a hierarchy (see subsection III-B).

The AHP method produces a ranking of suppliers with the percentage score of the DM's preference of each supplier over the others. However, in the problem of sustainable supplier evaluation it would be beneficial not only to know how much one supplier is preferred over its competition, but also to compare such supplier with a potential ideal supplier. Therefore, the proposed framework utilizes the TOPSIS method to compute a potential ideal and anti-ideal supplier (see subsection III-C).

Unfortunately, neither the AHP nor the TOPSIS method are resistant to the rank reversal phenomenon. Therefore, in the last step of the MCDA analysis of the suppliers, the proposed approach explores the complete space of the decision problem criteria values, thus providing a universal solution immune to rank reversal (see subsection III-D).

Last, but not least, the outputs of the aforementioned three MCDA components of the framework constitute the input to the temporal aggregation (see subsection III-E).

B. Hierarchical Structure of the Sustainable Supplier Selection Problem

The Analytical Hierarchy Process (AHP) by Saaty [30] is one of the best known and most widely used MCDA approaches. It is built on three main principles [31]: construction of a hierarchy, setting priorities and logical consistency. The decision problem is decomposed and structured into a hierarchy of sub-objectives, attributes, criteria and alternatives. In case of the proposed approach, the hierarchy of criteria is presented in Table II. Subsequently, the decision maker (DM) uses a pairwise comparison mechanism to determine the relative priority of each element at each level of the hierarchy. When comparing the elements of the hierarchy, a scale of $1-9$ is used to indicate the degree of preference of one element over the other. In case an element is less preferred, a reciprocal value is used, i.e. $\frac{1}{9}-1$.

The comparison results are stored in the pairwise comparison matrix, and the weights of individual elements are obtained. Each element of the matrix represents the dominance of an element in the column on the left over an element in the row on top. If the element on the left column is less important than the element in the row on top, a reciprocal value is inserted. The elements on the diagonal of the matrix are always

TABLE II
EVALUATION CRITERIA GROUPED INTO CRITERIA, SUBCRITERIA AND SECOND LEVEL SUBCRITERIA

Criteria	Subcriteria	2 Level Subcriteria	Ref
Cost		Product costs, Total supply cost which impact on final product, Financial cost, Operating expenditure, After-sales costs, Sunk/loss cost/customer dissatisfaction, Suppliers production pauses	[6], [9], [10], [14], [28], [13]
Quality		Product quality, Quality of Service, Warranty, Quality system certificate of the supplier, Quality assurance, Conformance quality, Quality image, Vendor specific, Quality manual, Documentation control, Archive of quality records, Receiving Inspection, Calibration control, Non-conforming material control system, Corrective and preventive action system, Audit mechanism	[9], [11], [16], [13]
Logistics		Choice of transportation, Reliability of quality, Delivery flexibility, Serious delivery delay rate, Compliance delivery with quantity, Supplier Stock Management, Technology Level, Capability of R & D, Order fulfill rate, Capability of Product Development, Procurement, Return forecast for each client, Warehouse management, IT management, Confirmed fill rate, Total order cycle time, System flexibility index, Integration technologies level, Increment in market share	[6], [29], [21], [28]
Social		The interests and rights of employee, The interests and the right of shareholders, Information disclosure, Expose nonfinancial information, Respect for the policy, Discrimination in employment, Child labor	[9], [11], [18], [19], [28], [13]
Profile		Customer base, Performance history, Production facility and capacity, Facility location, The number of working years in this sector, References, Communication capability, The number of personnel, Education status of the personnel, Machine capacity and capability, Manufacturing technology, Facilities manufacturing capacity, Technical capability, Manufacturing planning capability, Handling and packaging capability	[6], [17], [18], [13]
Green	Innovation	Green Technology Capabilities, Green Process/Production Planning, Recycling Product Design, Renewable Product Design, Green R & D Project, Redesign of Product	[9], [29], [14], [18], [28]
	Environment protection	Environment Efficiency, Eco-design, Environment Protection System Certification, Environmental Protection policies/plans	[29], [11], [12], [19], [28], [13]
	Environment Management	Production of material ecologically efficient, Eco-design requirements for energy using products, Level of restriction of hazardous substance in the production process, Compliance with the local regulation and policies	[29], [14], [15], [28]
	Pollution control	Air Emissions, Waste water, Pollution Control Capability, Pollution Reduction Capability	[9], [12], [19], [28]
	Hazardous Substance Management	Management of hazardous substances in the production procedure, Prevention of mixed material, Process Auditing, Warehouse Management, Inventory of Hazardous Substance	[29], [15], [22], [13]
	Image	Ratio of green customers to total customers, Green customers market share, Stakeholders relationship, Green materials coding and recording	[9], [11], [18], [28], [13]
	Product	Recycle, Green Packaging, Cost of Component Disposal, Green Production, Reuse, Re-Manufacture, Disposal	[6], [12], [20], [22], [28]
	Materials	Materials used in the supplied components that reduce the impact on natural resources, Ability to alter process and product for reducing the impact on natural resources	[29], [15], [19], [28], [13]

equal to 1. Therefore, a total of $n(n-1)/2$ comparisons needs to be performed. The procedure is repeated on all subsystems of the hierarchy. Sometimes, the DM's judgments can be inconsistent. However, in the AHP method, the inconsistency can be considered a tolerable error in measurement, as long as it does not exceed 10%.

C. Positive-Ideal and Negative-Ideal Supplier

The TOPSIS method (Technique for Order Performance by Similarity to Ideal Solution) utilized in the proposed approach, is a popular MCDA decision-making technique, originally developed by Hwang and Yoon [32], based on the idea to compare relative the distances between the alternatives and the ideal (PIS, positive ideal solution) and anti-ideal solutions (NIS, negative ideal solution). The best alternative should be as close as possible to the PIS, and, at the same time, as far as possible from the NIS.

The algorithm of the TOPSIS method comprises of six stages. In the first of them, the decision maker (DM) is required to choose m alternatives and n criteria for use in solving the problem, which are used to build the decision matrix $D[x_{ij}]$. The rows of the matrix represent alternatives and the columns represent criteria. The x_{ij} element is a representation of the decision attribute of the ith alternative regarding the jth criterion:

$$D[x_{ij}] = \begin{pmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1n} \\ x_{21} & x_{22} & x_{23} & ... & x_{2n} \\ x_{31} & x_{32} & x_{33} & ... & x_{3n} \\ ... & ... & ... & ... & ... \\ x_{m1} & x_{m2} & x_{m3} & ... & x_{mn} \end{pmatrix} \quad (1)$$

The second step of the procedure is the decision matrix normalization. Each decision attribute is normalized separately for each criterion. The following formulae are used to normalize benefit and cost criteria respectively:

$$r_{ij} = \frac{x_{ij} - min_i(x_{ij})}{max_i(x_{ij}) - min_i(x_{ij})} \quad (2)$$

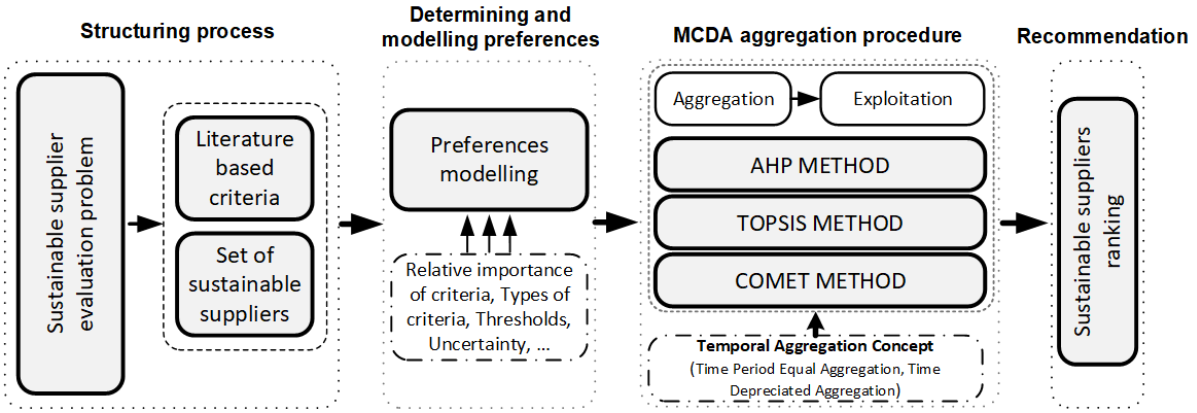


Fig. 1. Visual illustration of the proposed approach

$$r_{ij} = \frac{max_i(x_{ij}) - x_{ij}}{max_i(x_{ij}) - min_i(x_{ij})} \tag{3}$$

In the third step of the procedure, a weighted normalized decision matrix is created with the following formula:

$$v_{ij} = w_j \cdot r_{ij} \tag{4}$$

The PIS (V_J^+) and NIS (V_J^-) are obtained in the fourth step:

$$V_j^+ = \{v_1^+, v_2^+, v_3^+, \ldots, v_n^+\} \tag{5}$$

$$V_j^- = \{v_1^-, v_2^-, v_3^-, \ldots, v_n^-\} \tag{6}$$

In the fifth step, the Euclidean distances between the alternatives and PIS and NIS are computed:

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^+)^2} \tag{7}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2} \tag{8}$$

In the last step of the algorithm, the relative closeness of the alternative to the ideal solution is calculated:

$$CC_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{9}$$

Based on the CCi values, the final ranking of alternatives is created. In case of the proposed approach, the CCi value of each supplier allows to understand how far it is from a potential ideal supplier.

D. Prevention of the Rank Reversal Phenomenon

The proposed approach utilizes the Characteristic Objects METhod (COMET) [33] for exploring the complete space of possible solutions. The COMET method is based on the fuzzy sets theory and instead of comparing alternatives, as it is in AHP and TOPSIS, characteristic objects are created and compared in it, in order to create a linguistic rule base. Each evaluated alternative is subsequently scored in a defuzzification process. It is important to note, that due to the fact that a complete space of possible solutions of the decision problem has been explored by the comparisons of the problem's characteristic objects, introduction of a new supplier to the analysis will not change the evaluations of the remaining suppliers.

In the first step of the COMET procedure, the expert determines the dimensionality of the problem by selecting r criteria, $C_1, C_2, ..., C_r$. Then, a set of fuzzy numbers is selected for each criterion C_i, e.g. $\{\tilde{C}_{i1}, \tilde{C}_{i2}, ..., \tilde{C}_{ic_i}\}$ (10):

$$\begin{aligned} C_1 &= \{\tilde{C}_{11}, \tilde{C}_{12}, \ldots, \tilde{C}_{1c_1}\} \\ C_1 &= \{\tilde{C}_{21}, \tilde{C}_{22}, \ldots, \tilde{C}_{2c_2}\} \\ &\ldots \\ C_r &= \{\tilde{C}_{r1}, \tilde{C}_{r2}, \ldots, \tilde{C}_{rc_r}\} \end{aligned} \tag{10}$$

where $c_1, c_2, \ldots, c_r$ are the ordinals of the fuzzy numbers for all criteria.

In the second step, characteristic objects (CO) are obtained as a Cartesian product of the fuzzy numbers' cores of all the criteria (11):

$$CO = C(C_1) \times C(C_2) \times \ldots \times C(C_r) \tag{11}$$

As a result, an ordered set of all CO is obtained (12):

$$\begin{aligned} CO_1 &= C(\tilde{C}_{11}), C(\tilde{C}_{21}), ..., C(\tilde{C}_{r1}) \\ CO_2 &= C(\tilde{C}_{11}), C(\tilde{C}_{21}), ..., C(\tilde{C}_{r2}) \\ &\ldots \\ CO_t &= C(\tilde{C}_{1c_1}), C(\tilde{C}_{2c_2}), ..., C(\tilde{C}_{rc_r}) \end{aligned} \tag{12}$$

where t is the count of COs and is equal to (13):

$$t = \prod_{i=1}^{r} c_i \tag{13}$$

In the third step of the procedure, the expert determines the Matrix of Expert Judgment (MEJ) by comparing the COs pairwise. The matrix is presented below:

$$MEJ = \begin{pmatrix} \alpha_{11} & \alpha_{12} & .... & \alpha_{1t} \\ \alpha_{21} & \alpha_{22} & ... & \alpha_{2t} \\ ... & ... & ... & ... \\ \alpha_{t1} & \alpha_{t2} & ... & \alpha_{tt} \end{pmatrix} \tag{14}$$

where α_{ij} is the result of comparing CO_i and CO_j by the expert. The function f_{exp} denotes the mental judgment function of the expert. It depends solely on the knowledge of the expert. The expert's preferences can be presented as (15):

$$\alpha_{ij} = \begin{cases} 0.0, & f_{exp}(CO_i) < f_{exp}(CO_j) \\ 0.5, & f_{exp}(CO_i) = f_{exp}(CO_j) \\ 1.0, & f_{exp}(CO_i) > f_{exp}(CO_j) \end{cases} \tag{15}$$

After the MEJ matrix is prepared, a vertical vector of the Summed Judgments (SJ) is obtained as follows (16).

$$SJ_i = \sum_{j=1}^{t} \alpha_{ij} \tag{16}$$

Finally, the values of preference are approximated for each characteristic object. Correspondingly, a vertical vector P is obtained, where the $i-th$ row contains the approximate value of preference for CO_i.

Then, in the fourth step, each characteristic object and its value of preference is converted to a fuzzy rule as follows (17):

$$IF \quad C(\tilde{C}_{1i}) \quad AND \quad C(\tilde{C}_{2i}) \quad AND \quad ... \quad THEN \quad P_i \tag{17}$$

Thus, a complete fuzzy rule base is obtained.

Eventually, in the final step, each alternative is presented as a set of crisp numbers, e.g., $A_i = \{a_{1i}, a_{2i}, ..., a_{ri}\}$. This set corresponds to the criteria $C_1, C_2, ..., C_r$. Mamdani's fuzzy inference method is used to compute the preference of the $i-th$ alternative.

E. Temporal Aggregation of the Supplier Evaluation Results

The classic MCDA procedure requires both the alternatives and criteria to be constant [34], [35]. However, if criteria measurements are collected over a span of time, the ones closest to the time of the evaluation are intuitively the most valid. By all means, the criteria measurements from all periods can be aggregated using for example fuzzy sets theory and fuzzy numbers. However, it would affect the accuracy of the evaluation method input data and, consequently, could lead to oversimplifying the model and reducing the quality of the decision support. Therefore, in the proposed approach, instead of aggregating the input data, the DM should perform a temporal aggregation of the outcomes produced by evaluations produced in each period.

The temporal aggregation concept is based on the construction of a general utility function with an additional attribute called *forgetting*. Two possible types of forgetting strategies can be used:

TPEA
: Time Period Equal Aggregation – the impact of individual ratings on the final outcome of the assessment is balanced;

TDA
: Time Depreciated Aggregation – along with increasing distance of the historical data to the current period, its significance is being diminished.

Regardless of the forgetting strategies chosen, the general utility of a supplier can be obtained with the formula:

$$V(a^i) = \sum_{k=1}^{n} S_{ik} \cdot p(t_k) \tag{18}$$

where $V(a^i)$ denotes the general utility for the ith supplier on the basis of all n periods taken into consideration, S_{ik} means the utility of the ith supplier in period k and $p(t_k)$ means the significance of data for the k period in time t, based on the selected forgetting strategy. S_{ik} is determined in the previous step by the AHP, TOPSIS and COMET methods.

IV. Empirical Research

The proposed approach was empirically verified on a real company. A set of thirty suppliers of the company were selected for the research. The suppliers for the research were selected based on the yearly and monthly turnover. The criteria for the study were chosen as a result of a thorough literature review and are presented in Table II. The companies' performances in some of the criteria, such as time to confirm delivery, delivery time, delivery on-time, complaints, turnover, cost of transport, terms of payment and currency were fetched automatically from ERP systems, whereas for criteria where automation was not possible, surveys and expert judgment were utilized.

The obtained measurements of each criterion were normalized and mapped according to the Likert scale. In the next step, the AHP, TOPSIS and COMET methods were used to obtain the utility values of each supplier for each period. Eventually, temporal aggregation was performed based on five strategies TPEA, TDA1, TDA2, TDA3 and TDA4, which are illustrated in Table III and Fig 2.

In case of the TPEA strategy, the $p(t_k)$ value is always equal 1. In case of TDA1, TDA2 and TDA3 the forgetting

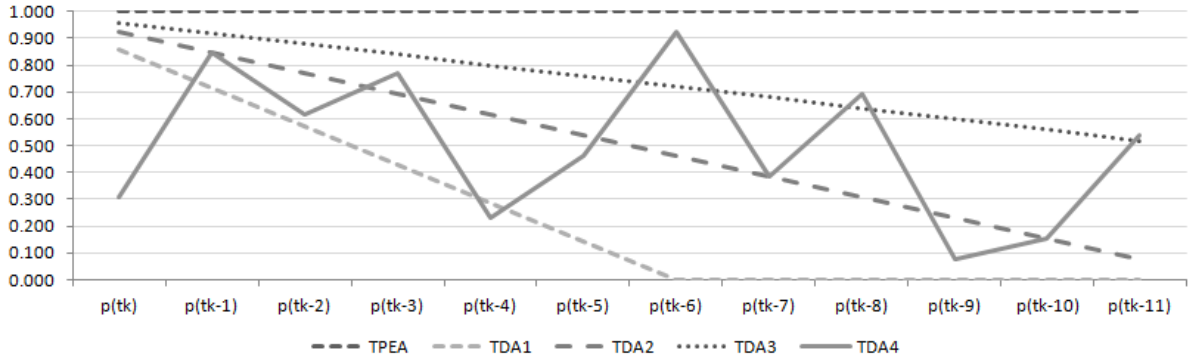


Fig. 2. Forgetting functions for TPEA, TDA1, TDA2, TDA3 and TDA4 strategies

TABLE III
FORGETTING FUNCTION PARAMETERS FOR TPEA, TDA1, TDA2, TDA3 AND TDA4 STRATEGIES

Aggregation	p(tk)	p(tk-1)	p(tk-2)	p(tk-3)	p(tk-4)	p(tk-5)	p(tk-6)	p(tk-7)	p(tk-8)	p(tk-9)	p(tk-10)	p(tk-11)
TPEA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
TDA1	0.857	0.714	0.571	0.429	0.286	0.143	0.000	0.000	0.000	0.000	0.000	0.000
TDA2	0.923	0.846	0.769	0.692	0.615	0.538	0.462	0.385	0.308	0.231	0.154	0.077
TDA3	0.960	0.920	0.880	0.840	0.800	0.760	0.720	0.680	0.640	0.600	0.560	0.520
TDA4	0.308	0.846	0.615	0.769	0.231	0.462	0.923	0.385	0.692	0.077	0.154	0.538

function is spread over 6, 12 and 24 months respectively. In case of TDA4, the value of the forgetting function depends on the supplies turnover.

The results of the temporal evaluation of the suppliers based on the AHP, TOPSIS and COMET methods partial evaluations are presented in Tables IV, V and VI respectively. For the reasons of brevity, the number of suppliers presented in the paper was limited to twelve.

As it can be noticed from the analysis of Fig. 3, all obtained rankings are highly correlated. A higher correlation can be observed between the rankings produced by the same method. However, the rankings obtained based on the AHP or TOPSIS methods are more correlated than any of these methods with the COMET method. This is caused by the fact that the COMET method explores the complete space of the decision problem, whereas the AHP and TOPSIS methods operate locally on the provided alternatives (suppliers).

The analysis of the temporal evaluation of the suppliers based on the three MCDA methods allowed to observe that depending on the aggregation strategy and the MCDA method used, the ranks of the suppliers vary slightly. However, it was noticed that the supplier A12 appeared on the majority of the rankings within the group of the best 5 suppliers.

V. CONCLUSIONS

The process of sustainable supplier selection is crucial to the companies' business continuity. Distortions in invalidly chosen suppliers can lead to a considerable impediment or even to a complete cease of company's operations. The current research focus is double-track. The first track focuses on the evaluation methods development. The second one focuses on the sustainability factors of the green cities, taking into account not only greening, but also human well being. While many prior studies focused on evaluation of suppliers based on performance and environmental criteria, the approach presented in this paper extended them with the following authors' contributions:

- possibility to organize the evaluation criteria into a multi-level hierarchy for better manageability of the decision problem;
- possibility to obtain a potential positive-ideal and negative-ideal supplier description and evaluation of the actual suppliers in relation to those two test cases;
- possibility to comprehensively explore the complete space of solutions of the sustainable supplier selection decision problem, thus preventing reversals in the produced rankings;
- possibility of temporal aggregation of the rankings obtained over a span of time with various forgetting strategies.

Compared to study [13], the performed research clearly shows that usage of a single hierarchical MCDA method in the process of sustainable supplier selection might not always be sufficient if the exploration of the complete space of the selection problem solutions or potential positive-ideal and negative-ideal supplier descriptions are needed.

TABLE IV
TEMPORAL EVALUATION OF 12 OF THE SUPPLIERS BASED ON THE AHP METHOD OUTPUT

Supplier	**Supplier Temporal Evaluation ($V(a_i)$)**					**Supplier Ranking in Temporal Evaluation**				
Ai	TPEA	TDA1	TDA2	TDA3	TDA4	TPEA	TDA1	TDA2	TDA3	TDA4
A1	0.3277	0.0825	0.1634	0.2422	0.1646	21	23	24	23	25
A2	0.3544	0.0891	0.1782	0.2628	0.177	15	17	16	15	16
A3	0.4262	0.1067	0.2123	0.315	0.2103	12	12	12	12	12
A4	0.3971	0.1001	0.1978	0.2934	0.1982	14	14	14	14	14
A5	0.4847	0.1181	0.2399	0.3574	0.2418	4	5	5	4	4
A6	0.4709	0.1179	0.2361	0.3488	0.234	6	6	6	6	7
A7	0.2488	0.0701	0.1347	0.1894	0.1331	29	27	29	29	29
A8	0.454	0.1149	0.2302	0.3376	0.2298	8	9	8	8	9
A9	0.5059	0.1323	0.2585	0.3773	0.2531	3	1	2	3	3
A10	0.4325	0.1086	0.2174	0.3207	0.2172	10	11	11	11	11
A11	0.3308	0.0843	0.1662	0.2452	0.1664	19	22	22	22	23
A12	0.5136	0.1257	0.2552	0.3792	0.2571	2	3	3	2	2

TABLE V
TEMPORAL EVALUATION OF 12 OF THE SUPPLIERS BASED ON THE TOPSIS METHOD OUTPUT

Supplier	**Supplier Temporal Evaluation ($V(a_i)$)**					**Supplier Ranking in Temporal Evaluation**				
Ai	TPEA	TDA1	TDA2	TDA3	TDA4	TPEA	TDA1	TDA2	TDA3	TDA4
A1	7.4096	1.892	3.7341	5.4983	3.7436	15	17	17	16	16
A2	7.5508	1.9118	3.8349	5.6185	3.7822	13	16	13	13	14
A3	7.6704	1.9496	3.8679	5.6931	3.8418	11	14	12	12	13
A4	7.1932	1.802	3.5955	5.3224	3.5904	17	20	18	17	18
A5	8.6405	2.1266	4.3261	6.397	4.378	7	7	7	7	6
A6	8.9969	2.2919	4.5817	6.701	4.4994	2	3	3	2	4
A7	5.5666	1.5739	3.014	4.2392	2.9679	29	25	27	29	28
A8	7.4461	1.931	3.8171	5.559	3.7492	14	15	15	14	15
A9	8.9034	2.3761	4.6095	6.6706	4.5037	3	1	2	3	3
A10	8.2315	2.0806	4.1645	6.1167	4.1403	8	8	8	8	8
A11	6.9282	1.8095	3.5188	5.1553	3.475	19	19	20	20	21
A12	9.0961	2.3101	4.6211	6.7691	4.5967	1	2	1	1	1

The research has identified possible areas of improvement and future work directions. The presented approach is only a framework which requires further verification in a complete data space, expanding the presented case study.

REFERENCES

[1] R. R. Colton, *Industrial Purchasing: Principles and Practices*. CE Merrill, 1962.

[2] J. Rezaei, T. Nispeling, J. Sarkis, and L. Tavasszy, "A supplier selection life cycle approach integrating traditional and environmental criteria using the best worst method," *Journal of Cleaner Production*, vol. 135, pp. 577–588, Nov. 2016. doi: 10.1016/j.jclepro.2016.06.125. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652616308022

[3] R. Khodaverdi and L. Olfat, "A fuzzy mcdm approach for supplier selection and evaluation: a case study in an automobile manufacturing company," in *2011 IEEE International Conference on Industrial Engineering and Engineering Management*, 2011, pp. 25–27.

[4] A. Świerczek, "The impact of supply chain integration on the "snowball effect" in the transmission of disruptions: An empirical evaluation of the model," *International Journal of Production Economics*, vol. 157, pp. 89–104, Nov. 2014. doi: 10.1016/j.ijpe.2013.08.010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925527313003654

[5] L. Meade and J. Sarkis, "Strategic analysis of logistics and supply chain management systems using the analytical network process1this work was partially supported by NSF Grants 9320949 and 9505967, and Texas Higher Education Coordinating Board ATP Grant Number 003656-036.1," *Transportation Research Part E: Logistics and Transportation Review*, vol. 34, no. 3, pp. 201–215, Sep. 1998. doi: 10.1016/S1366-5545(98)00012-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S136655459800012X

[6] F. T. S. Chan and N. Kumar, "Global supplier development considering risk factors using fuzzy extended AHP-based approach," *Omega*, vol. 35, no. 4, pp. 417–431, Aug. 2007. doi: 10.1016/j.omega.2005.08.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030504830500112X

[7] T. Grant, "10 Ways to Drive E-Commerce Sales During Slow Online Shopping Months." [Online].

TABLE VI
TEMPORAL EVALUATION OF 10 OF THE SUPPLIERS BASED ON THE COMET METHOD OUTPUT

Supplier	**Supplier Temporal Evaluation ($V(a_i)$)**					**Supplier Ranking in Temporal Evaluation**				
Ai	TPEA	TDA1	TDA2	TDA3	TDA4	TPEA	TDA1	TDA2	TDA3	TDA4
A1	9.0881	2.329	4.5962	6.7523	4.6717	5	2	2	4	2
A2	8.9049	2.2777	4.5325	6.6312	4.5816	8	7	6	7	5
A3	8.6651	2.2267	4.3968	6.4456	4.4126	11	11	11	12	10
A4	8.9418	2.2138	4.4665	6.6146	4.5174	7	13	9	8	8
A5	9.1468	2.2858	4.5789	6.7715	4.5951	3	5	4	3	3
A6	9.4306	2.3692	4.7489	6.9961	4.7525	1	1	1	1	1
A7	5.6706	1.6237	3.0464	4.306	3.0743	29	25	29	29	28
A8	8.6468	2.2358	4.4163	6.4469	4.3546	12	9	10	11	13
A9	8.6993	2.215	4.389	6.4579	4.37	10	12	12	10	12
A10	8.513	2.1355	4.2691	6.3062	4.2833	15	16	14	15	14
A11	7.1745	1.8728	3.6351	5.334	3.6209	21	24	24	23	25
A12	9.1024	2.2848	4.5677	6.7444	4.5695	4	6	5	5	6

	TPEA - AHP	**TDA1 - AHP**	**TDA2 - AHP**	**TDA3 - AHP**	**TDA4 - AHP**	**TPEA - TOPSIS**	**TDA1 - TOPSIS**	**TDA2 - TOPSIS**	**TDA3 - TOPSIS**	**TDA4 - TOPSIS**	**TPEA - COMET**	**TDA1 - COMET**	**TDA2 - COMET**	**TDA3 - COMET**	**TDA4 - COMET**
TPEA - AHP	1.000	0.937	0.959	0.980	0.959	0.926	0.847	0.909	0.921	0.926	0.828	0.762	0.795	0.810	0.783
TDA1 - AHP	0.937	1.000	0.986	0.975	0.947	0.875	0.902	0.913	0.899	0.895	0.762	0.788	0.772	0.774	0.737
TDA2 - AHP	0.959	0.986	1.000	0.990	0.977	0.887	0.895	0.919	0.911	0.919	0.760	0.757	0.767	0.766	0.737
TDA3 - AHP	0.980	0.975	0.990	1.000	0.977	0.904	0.877	0.920	0.919	0.927	0.794	0.771	0.792	0.794	0.769
TDA4 - AHP	0.959	0.947	0.977	0.977	1.000	0.851	0.828	0.874	0.868	0.917	0.732	0.706	0.730	0.726	0.735
TPEA - TOPSIS	0.926	0.875	0.887	0.904	0.851	1.000	0.932	0.979	0.993	0.957	0.881	0.799	0.846	0.869	0.812
TDA1 - TOPSIS	0.847	0.902	0.895	0.877	0.828	0.932	1.000	0.976	0.957	0.936	0.761	0.767	0.773	0.778	0.715
TDA2 - TOPSIS	0.909	0.913	0.919	0.920	0.874	0.979	0.976	1.000	0.994	0.971	0.828	0.792	0.822	0.834	0.783
TDA3 - TOPSIS	0.921	0.899	0.911	0.919	0.868	0.993	0.957	0.994	1.000	0.967	0.857	0.798	0.840	0.855	0.800
TDA4 - TOPSIS	0.926	0.895	0.919	0.927	0.917	0.957	0.936	0.971	0.967	1.000	0.816	0.768	0.810	0.812	0.806
TPEA - COMET	0.828	0.762	0.760	0.794	0.732	0.881	0.761	0.828	0.857	0.816	1.000	0.930	0.975	0.993	0.960
TDA1 - COMET	0.762	0.788	0.757	0.771	0.706	0.799	0.767	0.792	0.798	0.768	0.930	1.000	0.972	0.953	0.947
TDA2 - COMET	0.795	0.772	0.767	0.792	0.730	0.846	0.773	0.822	0.840	0.810	0.975	0.972	1.000	0.989	0.978
TDA3 - COMET	0.810	0.774	0.766	0.794	0.726	0.869	0.778	0.834	0.855	0.812	0.993	0.953	0.989	1.000	0.965
TDA4 - COMET	0.783	0.737	0.737	0.769	0.735	0.812	0.715	0.783	0.800	0.806	0.960	0.947	0.978	0.965	1.000

Fig. 3. Correlation matrix between AHP, TOPSIS and COMET evaluations rankings based on TPEA, TDA1, TDA2, TDA3 and TDA4 forgetting strategies

Available: https://www.infusionsoft.com/business-success-blog/sales/e-commerce/10-ways-to-drive-e-commerce-sales-during-slow-months

[8] "Broadwater Farm - Summer Fruit Picking - The Job and Pay." [Online]. Available: https://www.broadwaterfarm.biz/summer-fruit-picking-uk/

[9] G. Akman, "Evaluating suppliers to include green supplier development programs via fuzzy c-means and VIKOR methods," *Computers & Industrial Engineering*, vol. 86, pp. 69–82, Aug. 2015. doi: 10.1016/j.cie.2014.10.013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835214003441

[10] D. Kannan, A. B. L. d. S. Jabbour, and C. J. C. Jabbour, "Selecting green suppliers based on GSCM practices: Using fuzzy TOPSIS applied to a Brazilian electronics company," *European Journal of Operational Research*, vol. 233, no. 2, pp. 432–447, Mar. 2014. doi: 10.1016/j.ejor.2013.07.023. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221713006048

[11] A. K. Uppala, R. Ranka, J. J. Thakkar, M. V. Kumar, and S. Agrawal, "Selection of Green Suppliers Based on GSCM Practices: Using Fuzzy MCDM Approach in an Electronics Company," *Handbook of Research on Fuzzy and Rough Set Theory in Organizational Decision Making*, pp. 355–375, 2017. doi: 10.4018/978-1-5225-1008-6.ch016. [Online]. Available: https://www.igi-global.com/chapter/selection-of-green-suppliers-based-on-gscm-practices/169495

[12] K. Chatterjee, D. Pamucar, and E. K. Zavadskas, "Evaluating the performance of suppliers based on using the R'AMATEL-MAIRCA method for green supply chain implementation in electronics industry," *Journal of Cleaner Production*, vol. 184, pp. 101–129, May 2018. doi: 10.1016/j.jclepro.2018.02.186. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652618305055

[13] J. Wątróbski, W. Sałabun, and G. Ladorucki, "The Temporal Supplier Evaluation Model Based on Multicriteria Decision Analysis Methods," in *Intelligent Information and Database Systems*, N. T. Nguyen, S. Tojo, L. M. Nguyen, and B. Trawiński, Eds. Cham: Springer International Publishing, 2017, vol. 10191, pp. 432–442. ISBN 978-3-319-54471-7 978-3-319-54472-4. [Online]. Available: http://link.springer.com/10.1007/978-3-319-54472-4_41

[14] S. H. Hashemi, A. Karimi, and M. Tavana, "An integrated green supplier selection approach with analytic network process and improved Grey relational analysis," *International Journal of Production Economics*, vol. 159, pp. 178–191, Jan. 2015. doi: 10.1016/j.ijpe.2014.09.027. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925527314003053

[15] S. Zailani, K. Govindan, M. Iranmanesh, M. R. Shaharudin, and Y. Sia Chong, "Green innovation adoption in automotive supply chain: the Malaysian case," *Journal of Cleaner Production*, vol. 108, pp. 1115–1122, Dec. 2015. doi: 10.1016/j.jclepro.2015.06.039. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959652615007684

[16] F. Dweiri, S. Kumar, S. A. Khan, and V. Jain, "Designing an integrated AHP based decision support system for supplier selection in automotive industry," *Expert Systems with Applications*, vol. 62, pp. 273–283, Nov. 2016. doi: 10.1016/j.eswa.2016.06.030. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416303104

[17] N. Banaeian, H. Mobli, B. Fahimnia, I. E. Nielsen, and M. Omid, "Green supplier selection using fuzzy group decision making methods: A case study from the agri-food industry," *Computers & Operations Research*, vol. 89, pp. 337–347, Jan. 2018. doi: 10.1016/j.cor.2016.02.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0305054816300399

[18] S. Gold and A. Awasthi, "Sustainable global supplier selection extended towards sustainability risks from (1+n)th tier suppliers using fuzzy AHP based approach," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 966–971, Jan. 2015. doi: 10.1016/j.ifacol.2015.06.208. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896315004474

[19] H. Kaur, S. P. Singh, and R. Glardon, "An Integer Linear Program for Integrated Supplier Selection: A Sustainable Flexible Framework," *Global Journal of Flexible Systems Management*, vol. 17, no. 2, pp. 113–134, Jun. 2016. doi: 10.1007/s40171-015-0105-1. [Online]. Available: https://link.springer.com/article/10.1007/s40171-015-0105-1

[20] K. Govindan and R. Sivakumar, "Green supplier selection and order allocation in a low-carbon paper industry: integrated multi-criteria heterogeneous decision-making and multi-objective linear programming approaches," *Annals of Operations Research*, vol. 238, no. 1-2, pp. 243–276, Mar. 2016. doi: 10.1007/s10479-015-2004-4. [Online]. Available: https://link.springer.com/article/10.1007/s10479-015-2004-4

[21] C.-H. Wang, "Using quality function deployment to conduct vendor assessment and supplier recommendation for business-intelligence systems," *Computers & Industrial Engineering*, vol. 84, pp. 24–31, Jun. 2015. doi: 10.1016/j.cie.2014.10.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360835214003362

[22] K. Govindan, S. Rajendran, J. Sarkis, and P. Murugesan, "Multi criteria decision making approaches for green supplier evaluation and selection: a literature review," *Journal of Cleaner Production*, vol. 98, pp. 66–83, Jul. 2015. doi: 10.1016/j.jclepro.2013.06.046. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095965261300437X

[23] J. Wątróbski, "Outline of Multicriteria Decision-making in Green Logistics," *Transportation Research Procedia*, vol. 16, pp. 537–552, Jan. 2016. doi: 10.1016/j.trpro.2016.11.051. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146516306652

[24] I. Banamar and Y. De Smet, "An extension of promethee ii to temporal evaluations code-smg–technical report series," 2016.

[25] O. Sahin and S. Mohamed, "A spatial temporal decision framework for adaptation to sea level rise," *Environmental Modelling & Software*, vol. 46, pp. 129–141, Aug. 2013. doi: 10.1016/j.envsoft.2013.03.004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1364815213000558

[26] J. Zhu and K. W. Hipel, "Multiple stages grey target decision making method with incomplete weight based on multi-granularity linguistic label," *Information Sciences*, vol. 212, pp. 15–32, Dec. 2012. doi: 10.1016/j.ins.2012.05.011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0020025512003556

[27] A. Arasteh, A. Aliahmadi, and M. M. Omran, "A Multi-stage Multi Criteria Model for Portfolio Management," *Arabian Journal for Science and Engineering*, vol. 39, no. 5, pp. 4269–4283, May 2014. doi: 10.1007/s13369-014-0987-9. [Online]. Available: http://link.springer.com/10.1007/s13369-014-0987-9

[28] J. Wątróbski and W. Sałabun, "Green Supplier Selection Framework Based on Multi-Criteria Decision-Analysis Approach," in *Sustainable Design and Manufacturing 2016*, R. Setchi, R. J. Howlett, Y. Liu, and P. Theobald, Eds. Cham: Springer International Publishing, 2016, vol. 52, pp. 361–371. ISBN 978-3-319-32096-0 978-3-319-32098-4. [Online]. Available: http://link.springer.com/10.1007/978-3-319-32098-4_31

[29] D. Kannan, K. Govindan, and S. Rajendran, "Fuzzy Axiomatic Design approach based green supplier selection: acase study from Singapore," *Journal of Cleaner Production*, vol. 96, pp. 194–208, Jun. 2015. doi: 10.1016/j.jclepro.2013.12.076. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095965261300930X

[30] T. L. Saaty, "Decision making with the analytic hierarchy process," *International Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, Jan. 2008. doi: 10.1504/IJSSci.2008.01759. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJSSci.2008.01759

[31] C. Macharis, J. Springael, K. De Brucker, and A. Verbeke, "PROMETHEE and AHP: The design of operational synergies in multicriteria analysis.: Strengthening PROMETHEE with ideas of AHP," *European Journal of Operational Research*, vol. 153, no. 2, pp. 307–317, Mar. 2004. doi: 10.1016/S0377-2217(03)00153-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037722170300153X

[32] C.-L. Hwang, Y.-J. Lai, and T.-Y. Liu, "A new approach for multiple objective decision making," *Computers & Operations Research*, vol. 20, no. 8, pp. 889–899, Oct. 1993. doi: 10.1016/0305-0548(93)90109-V. [Online]. Available: http://www.sciencedirect.com/science/article/pii/030505489390109V

[33] W. Sałabun, "The characteristic objects method: A new distance-based approach to multicriteria decision-making problems," *Journal of Multi-Criteria Decision Analysis*, vol. 22, no. 1-2, pp. 37–50, 2015.

[34] A. Guitouni and J.-M. Martel, "Tentative guidelines to help choosing an appropriate MCDA method," *European Journal of Operational Research*, vol. 109, no. 2, pp. 501–521, Sep. 1998. doi: 10.1016/S0377-2217(98)00073-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221798000733

[35] B. Roy and D. Vanderpooten, "The European school of MCDA: Emergence, basic features and current works," *Journal of Multi-Criteria Decision Analysis*, vol. 5, no. 1, pp. 22–38, Mar. 1996. doi: 10.1002/(SICI)1099-1360(199603)5:1<22::AID-MCDA93>3.0.CO;2-F. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-1360%28199603%295%3A1%3C22%3A%3AAID-MCDA93%3E3.0.CO%3B2-F

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 779–788
DOI: 10.15439/2018F333
ISSN 2300-5963 ACSIS, Vol. 15

Information System Backsourcing: A Systematic Literature Analysis

Christian Leyh
Technische Universität Dresden
Chair of Information Systems, esp. IS in Manufacturing and Commerce, Helmholtzstr. 10, 01069 Dresden, Germany
Email: Christian.Leyh@tu-dresden.de

Thomas Schäffer
University of Applied Sciences Heilbronn, Faculty of Business Administration, Max-Planck-Str. 39, 74081 Heilbronn, Germany
Email: Thomas.Schaeffer@hs-heilbronn.de

Trung Duc Nguyen
Technische Universität Dresden
Chair of Information Systems, esp. IS in Manufacturing and Commerce, Helmholtzstr. 10, 01069 Dresden, Germany

Abstract*—As information technology (IT) in private and public organizations continues to gain importance, information system outsourcing (ISO) has become a critical component of corporate strategy for many institutions. Consequently, a substantial amount of research has investigated topics around ISO decisions and outcomes. However, despite decades of ISO research, analyses that focus on information system (IS) backsourcing remain scarce. Therefore, in this paper, we share the results of a systematic literature analysis of papers that consider IS backsourcing. Within our paper, IS backsourcing is integrated into the wider research landscape of ISO. Finally, our study clearly identifies the need for further and more extensive research on IS backsourcing. The high dissatisfaction and failure of outsourcing arrangements should not be ignored. Both sourcing decisions and existing outsourcing arrangements must be analysed carefully and in the long term to ensure the success of the company.

I. Introduction

Information system outsourcing (ISO) has become a common alternative to running in-house information technology (IT) operations and development activities across company and national borders. Looking at the ever-growing range of IT services that are available in the global outsourcing market, companies can purchase not only small development and hosting services but also entire business process and infrastructure solutions. Thus, information system outsourcing (ISO) / information technology outsourcing (ITO) has become a well-established field of research. Typical research topics evolved from ISO motivations and success factors in the 1990s to the relationship between vendor and client in the 2000s to modern forms of sourcing in approximately 2008 [1].

ISO/ITO are backed by many drivers: optimized cost situation (mainly through high labour cost differences), access to highly qualified staff, access to new markets, high flexibility and technical feasibility. These factors weigh even more heavily if the company-owned IT department experiences a lack of competence, high costs or a lack of attention because it does not belong to the core business [2].

What seems like an ideal solution comes with a variety of risks and problems. These range from global client risks, which include the state of the labour market and the infrastructure in the provider's country, to local client risks, such as cultural differences, different time zones, language problems, knowledge transfer and technical skills, which hamper the quality of the cooperation [3]. Further risk factors are, for example, interest conflicts, low product or service quality, high turnover rates on the provider-side, additional work and extra costs for the client and technology-related risks [4]. Referring to practice-oriented studies, many of the stated risks have become problems [5]. This argument is also supported by surveys that report that 50% of analysed offshore outsourcing contracts that are signed by North American companies failed to meet their expectations [5], 30 to 50% of the companies that are involved in offshore outsourcing had cancelled their contracts [6], and 20% of those outsourcing contracts are cancelled in the first year [7].

One of several alternatives for a company that is facing those problems is to terminate its relationship with the vendor and relocate its IT services. Therefore, this paper will focus on backsourcing as one type of relocation. IS backsourcing in the IS literature is generally defined as the process of recalling operations back in-house after they have been outsourced [8]. The practice of IS backsourcing has been only briefly discussed in the scientific literature, compared to the growing literature on ISO/ITO [9]–[11]. This is surprising when examining some of the numerous prominent cases, where banks such as *JP Morgan Chase* and *Bank One* prematurely terminated their multi-billion-dollar ITO contracts to pull those services back in-house [12]. Such cases highlight the practical relevance of IS backsourcing. To address the knowledge void that surrounds this phenomenon, it is necessary to analyse the state of the academic literature on IS backsourcing, to answer the most important question, namely, "*Why backsource IT services?*", and determine what is known about the transition process to answer the question "*How can IT services be backsourced?*".

Therefore, the main objective of our paper is to identify drivers for IS backsourcing and factors that influence the transition phase in the existing academic literature. To fulfil this objective, we conducted a systematic literature analysis.

This work was not supported by any organization

With this analysis, we aim at answering two research questions:

- *What are drivers for companies to backsource their IT services?*
- *How do companies backsource their IT services?*

To obtain an initial answer to these questions, the paper is structured as follows: First, we present a short overview of our methodology. Next, as the main focus of our paper, we describe in detail the findings of our literature analysis. Then, we conclude with a discussion and summary of our results and identify aspects for future research.

II. Research Approach

A. Literature Review

To obtain a general overview of the body of IS backsourcing research, we followed the guidelines of [13], in which a five-step approach for conducting IS literature reviews is provided. In the first step, we defined the review scope, backed by a taxonomy of literature reviews that were developed by [14]. In steps 2 and 3, relevant working definitions (for a common understanding of the used search terms) and the search process (sources and the selection criteria for the literature) are described. The definition of analysis and synthesis is assigned to the fourth step, including a categorization referring to [12], [15]. The final step is composed of summarization of key findings and specification of a research agenda.

As an initial step, we examined IS journals and IS conference proceedings using the databases *AIS Electronic Library*, *EBSCOhost Business Source Complete*, *Emerald Insight*, *IEEE Xplore* and *ScienceDirect*. We conducted electronic searches of *titles*, *keywords* and *abstracts* for the following search term*: ["backsourc*" OR "backshor*" OR "reshor*" OR "insourc*"]*. Furthermore, we performed a second search on the following search term: *[("offshor*" OR "outsourc*") AND "fail*"]*. With these searches, we identified 290 publications. After analysing each article's abstract and keywords, and/or the full article when necessary, we excluded 220 articles that were duplicates or did not appear to be concerned with or relevant to our research focus. As a third analysing step, by reading the abstracts of the remaining 70 papers, we selected only 15 papers for deeper review. In the last step, five additional publications were identified through backward and forward search (as suggested by [16]). Table 2 in the Appendix gives an overview of the 20 publications that elaborate a consolidated view of the current field of IS backsourcing. In the remainder of this article, we focus on these papers.

B. Literature Analysis and Synthesis Framework

For the analysis and synthesis of the relevant literature, the analytical framework of [4] is partially considered, which refers to the perspectives *research focus* and *research approach*.

Research Focus: Dibbern et al. [2] developed a five-stage model of ISO. These five stages are divided into two main phases: *the decision process* and *the implementation*. The decision process contains the following questions and stages: ***Stage Decision*****:** (1) Why does an organization consider outsourcing? (e.g., drivers, antecedents), (2) What is outsourced? (e.g., functions, organizations) and (3) Which choices are made? (e.g., with a decision model or guideline); ***Stage Implementation*****:** (4) How is outsourcing carried out? (e.g., selecting vendors, transition of knowledge) and (5) What are the outcomes of outsourcing? (e.g., experience, lessons learned). Since backsourcing is a major decision in an organization, the adoption of this stage model appears appropriate. Thereby, our paper focusses on the two main stages by reducing the decision process to the "**WHY**" stage and the implementation to the "**HOW**" stage. This limitation is backed not only by the foci of most of the IS backsourcing literature, which instead examines the antecedents of backsourcing, but also by the similarity of the transition process in outsourcing and backsourcing in terms of influencing factors (see [17]).

Research Approach: Referring to the view on research approaches of [2], the identified literature was analysed and we differentiated between empirical and non-empirical approaches. Thereafter, the epistemology within the approaches was determined by following [4]. **Empirical approaches** contain the following types of epistemology: *interpretivism*, *positivism* and *descriptivism*, whereas **non-empirical approaches** can be distinguished between *conceptual* and *mathematical methods*.

III. Findings

As a result of the literature analysis and synthesis, this section outlines the determinants for backsourcing IT services and the factors that influence the transition phase. Working definitions are specified in the first part, followed by a descriptive analysis of the findings and their methodologies. Subsequently, the drivers and transition process influencers for IS backsourcing are analysed and synthesized.

A. Conceptual Background

Prior to identifying the drivers for outsourcing failure and relocating IT services from an offshore location to the home country, various working definitions must be clarified. As backsourcing is a type of general sourcing in the academic literature, both IS and manufacturing definitions could be applied. Especially in manufacturing literature, multiple synonyms are used, such as *back-shoring*, *reshoring*, *back-sourcing* and *de-internationalization* (see [18]). In IS research, mainly the terms *backsourcing* and *insourcing* are being used. Table 1 presents an overview of existing definitions.

A comparative analysis of the definitions that are found in both the IS and manufacturing literature (see Table 1)

reveals several characteristics of the relocation of manufacturing or IT services back to the home country of the company: **(a)** the relocation is the reverse decision with respect to a previous offshoring process, **(b)** the relocation does not necessarily involve repatriation or the closure of all of a company's offshore activities or plants **and (c)** a difference between the ownership modes can be identified (backshoring and backsourcing).

Referring to [4], in which a framework was specified for classifying the characteristics of the term offshoring, almost all characteristics can be assumed to be parallel to backsourcing. Only the category *Distance* may be left out, since the destination for the backshoring action is usually the domestic country of the company by definition. Thus, the focus lies on three characteristics (see [2], [4], [19]):

- **Ownership** (What property model shall be used?),
- **Function** (What IS services shall be backshored?) and
- **Degree** (To what extent shall IS services be backshored?).

These characteristics can be further divided into internal, external, partial, selective and total.

(1) Ownership: As opposed to offshoring, where the action begins in the home country of the company, backsourcing starts in the country to where the services have already been located. Hence, several reshoring alternatives are possible. Company **internal** IS services can generally be reshored to captive organizational units that are still located in foreign countries (in-house reshoring following [20]). They can also be reshored to **partially** owned companies (e.g., joint ventures or strategic alliances, following [10], [21]) or **externally** owned companies (near-/ offshore outsourcing, following [22]).

(2) Function: Strasser and Westner [23] determined that most studies do not specify which IS services are transferred, as often only general terms such as *information system development activities* and *IS functions* are stated. By clustering these terms, three groups of activities are defined: *infrastructure services*, *application development services* and *business process services* ([4], [21]).

(3) Degree: Here, distinctions between **total** and **selective** reshoring can be made ([21], [24], [25]). Contrary to the case of total offshoring, total reshoring of previously outsourced or offshored services is realistic.

This paper will focus on backsourcing as one case of relocating IS services. As the drivers for relocating IS services to the home country of a company will be investigated, backshoring, which is used primarily in the manufacturing reshoring literature, will be considered as well. This paper follows the initial definition of backsourcing of [26]: "[…] backsourcing, where companies who initially outsourced their IT decide to bring it back in-house." For the remainder of our review, further specifications were set to avoid ambiguities in the stated characteristics of reshoring: The backsourced IS services are considered to be integrated into the company-internal organization, which is located in the home country of the company. Only internal functions are backsourced entirely.

B. Applied Research Methodologies

The selected conference and journal publications regarding IS backsourcing can be divided into *non-empirical* (conceptual or mathematical) and *empirical* (interpretive, descriptive or positivist) papers. Mathematical modelling papers were not found. An overview on who used which approach can be found in Table 2 in the Appendix.

The **non-empirical conceptual** research papers include issues and perspectives that are related to the phenomenon of IS backshoring. Akoka and Comyn-Wattiau [27] designed a framework for understanding "Why to backsource IT" by defining rational and irrational factors. Another framework for understanding the decision to backsource IT was created by [28], who regarded the reasons for IT backsourcing not only as a problem-solving strategy but also as an internally or externally motivated opportunity for stability and growth. Another backsourcing decision model was constructed by [29], in which a decision process is designed by integrating knowledge from the research literature and expert interviews.

TABLE I.
OVERVIEW OF TYPES OF RESHORING

Concept	Definition	References
Backshoring	"Re-concentration of parts of production from own foreign locations as well as from foreign suppliers to the domestic production site of the company" [30]	[30]–[32]
Backsourcing	"[...] where companies who initially outsourced their IT decide to bring it back in-house." [26] "Production return relocation from an [...] external entity" [31]	[26], [31], [33]
De-internatio-nalisation	"Any voluntary or forced action that reduces a company's engagement in or exposure to current cross border activities" [34]	[34]–[36]
Insourcing	"Insourcing is the practice of evaluating the outsourcing option, but confirming the continued use of internal IT resources to achieve the same objectives of outsourcing" [11]	[11]
Reshoring	"Moving manufacturing back to the country of its parent company" [37]	[20], [37]

Moreover, McLaughlin and Peppard [9] integrated backsourcing into an end-to-end sourcing model, which means that backsourcing is one of several alternatives of sourcing options. Finally, Beardsell [38] tried to determine whether backsourcing improves the firm's innovative capability by integrating a broad range of theories.

Many **empirical interpretivist** research papers were identified, which were mostly based on real case studies. Butler et al. [17] conducted six semi-structured interviews in a company that recently backsourced its whole IT department to identify parallels between backsourcing and outsourcing. A similar approach was taken by several authors ([12], [33], [39], [40]), who examined antecedents for IS backsourcing, not only based on interviews but also by analysing external media publications. This method was extended by [41]–[44], in which internal company data, such as presentations and e-mails, were considered in the analysis of determinants that have led to the failure of ISO arrangements.

Several **empirical positivist** research papers were identified [10], [28], [45]. They interviewed over 250 employees from various companies in field studies on the success and failure factors for ITO and identified reasons why companies backsource.

Bhagwatwar et al. [46] used an **empirical descriptive** approach to analyse two case studies and developed best-practices for the success of an IS backsourcing arrangement in terms of knowledge re-integration based on their observations.

C. Why-Stage: Backsourcing drivers

Comparing the foci of the relevant IS backsourcing literature in this paper with those of selected papers in IS offshoring literature reviews ([1], [4], [23]), striking differences are observed. While the IT offshoring literature focusses on multiple dimensions, such as *distance* (onshore, nearshore, offshore), *function* (infrastructure, application, etc.), *degree* (selective, total), *perspective* (vendor, customer, consultant) and the *stages* of why, how, what and which to offshore, the IS backsourcing literature is very limited in its research perspectives. This underlines the weak pervasion of the subject matter from a research point of view.

Regarding the analysed papers, multiple types of classifications are proposed. Veltri [47] classified backsourcing drivers into *costs*, *uncertainty and risk*, *goal conflict* and *opportunism*. Wong [44] categorized his findings into *strategic factors*, *power and politics*, *outsourcing expectation gaps* and *changes in vendor organization*. A very general classification was made by Wong in 2008 [40], who categorized his findings into **(1)** *outsourcing expectation gaps*, **(2)** *internal organizational changes* and **(3)** *external environmental changes*. Driven by the content analysis of the selected publications, the categorization and sub-categorization by [40] are most suitable due to the broad variety of the obtained results.

In the following, expectation gaps are stated, followed by internal organizational changes and external environmental changes. A comparison of the results with **(a)** *the results to the ISO* and **(b)** *manufacturing backshoring* literature is performed afterwards.

(1) Backsourcing drivers through expectation gaps: The most commonly mentioned factor for moving IS services back among the selected research papers is *unsatisfying service quality* (mentioned in eleven of the papers). In particular, concrete factors are *low product and service quality*, *poor communication*, *lower productivity*, *poor commitment of the vendor* and a *lack of transparency*. An additional striking factor is *cultural differences*, which is reflected into *different understandings of hierarchy*, *punctuality, acknowledging mistakes*, and *accuracy and responsibility over tasks*. In addition, *knowledge mismatch*, which describes a lack of business and technology knowledge, and *process comprehension and experience* might lead to inefficiencies in collaboration. Independent from possible mismatches through differences of any kind, *opportunistic behaviour* must be considered as well, since the vendor has latitude, for example, in appointing key personnel to specific positions, which he might use to further his interests.

Furthermore, *cost aspects* are found to be of higher relevance, specifically unrealized cost savings through agency costs, transaction costs, hiring and retaining costs, lost performance and uncertainty costs. Through these categories, it becomes clear that cost and service quality are both highly considered in decisions on outsourcing and backsourcing IS services [26].

Losing control over the vendor's activities is also considered a central driver for backsourcing regarding possible principal-agent problems, such as inefficiencies through incorrect working directions and, especially, insecurity issues for sensitive information. From a strategic point of view, *failing to achieve defined outsourcing goals* is one of the most striking arguments for backsourcing. Additional drivers are *uncertainties regarding objectives*, *performance measurements* and *missing measures for low performance or failure*.

Lastly, an important factor for IS backsourcing is *missing access to latest technologies*, which refers not only to state-of-the-art soft- and hardware technologies but also to highly educated human and knowledge capital, which might lead to deficiencies in communication and cooperation and, finally, lower competitiveness.

(2) Backsourcing drivers through internal or external changes: From a company-internal organizational point of view, backsourcing can result from trivial causes, such as *changes of the (top) management*, *changes of the role of IT in the company* or *general changes in the strategic direction*. The last two factors go hand in hand due to the rising relevance of IT in the operations and strategies of

companies, especially in times of digitalization of assets and products. Deduced from this, outsourcing can also occur for reasons of business model changes, organizational changes in structures (e.g., through acquisitions, mergers, or divestments) or simply through a shift or lack of top executive support.

(3) External environmental changes: Comparatively few factors were identified regarding IT backsourcing. Those factors mostly refer to *changes in the vendor's strategy or organization* or, from an economic perspective, *uncertainties in demand and supply through economic ups and down*s. A link to the *technology factor* can also be made, meaning that disruptive technologies might lead to new markets, which again might lead to new business models and environments that require new sourcing considerations.

(a) Comparison to ISO literature: The most frequently identified drivers for ISO are of financial and strategic nature ([2], [4], [23]). These primarily include cost reduction, wealth maximization by leveraging cost savings, access and proximity to highly skilled employees and markets, focus on core competencies and higher flexibility and technical feasibility [4]. Comparing the most striking motivations of both fields, except for the cost factor, a relatively small number of matchings occur. Considering the reviewed determinants for consideration of ISO as a sourcing option by [23], who specify factors such as advanced technology access, chance for organizational changes and higher innovativeness, similar results dominate.

However, both phenomena contribute to the success of an organization via reconsideration of the business strategy and adaption to the business environment. Additional matchings can be identified when the stated risks for ISO are compared to the drivers for IS backsourcing. Gonzalez et al. [3] addressed risks for the client from different aspects, such as economic (e.g., unemployment rates, poor infrastructure), local (e.g., differences in culture, mentality, language and knowledge transfer or legal problems) and managerial risks (e.g., low quality, additional effort, hidden costs). A striking matching regarding the results of the review is the risk of impacting (internal and external) customer relationships, which is rarely stated in the ISO literature. In their analysis, Butler et al. [17] stated that backsourcing cannot be viewed as "outsourcing in reverse", which can be confirmed in this part of the analysis.

(b) Comparison to manufacturing backshoring literature: For this comparison, two existing systematic literature reviews were used ([15], [48]), which resulted in 22 and 20 selected publications for reshoring manufacturing. Although those reviews are similar in their analysed literature and period, different methods of categorizing the drivers for reshoring manufacturing were used.

Stentoft et al. [15] synthesized and summarized their findings from the reshoring-company perspective. The following aspects were considered: *cost*, *quality*, *time and flexibility*, *access to skills and knowledge*, *risks*, *market* and *other factors* (e.g., incentives from governments and change of a company's strategy).

On the other side, Wiesmann et al. [48] considered an economic perspective by selecting the following driver categories: *global competitive dynamics*, *host and home country*, *supply chain and firm specifics*.

However, the disparity between the manufacturing reviews makes a direct comparison with the findings of our analysis difficult. Therefore, a differentiated comparison appears appropriate. On an enterprise level, the major difference between the IS service and manufacturing business seems to be the subject matter of the backsourcing arrangement: intangible vs. tangible assets, whereas in IS backsourcing, the aspects of cost and quality are considered factors that influence the collaboration between client and vendor; those factors refer more to asset and logistical costs and product quality on the manufacturing-side. Linked to the relevance of collaboration in IS, related factors, such as cultural differences and communication as well as project management, are of high importance. In terms of control, only a few factors are specified in the manufacturing literature. One reason might be the deeper integration of IS services in the company, since IS services are being used by employees abroad whereas manufacturing functions as its own entity for the most part. This stresses the relevance of IS in terms of operations and strategy (see [2]). This goes hand in hand with the high relevance of designing well-conceived contracts for facing all types of contingencies. Lastly, the IS backsourcing literature concentrates on the company layer and considers changes in strategy, management and structure as possible drivers.

In contrast, the characteristic of tangibility influences most of the arguments that are stated in the review of [15], such as production and delivery reliability, supply chain risks and the value of "Made in X"-brandings. These points emphasize that operational artefacts, especially employees, products and the production process, are of interest in the analysis of manufacturing backshoring factors. A variety of parallels and similarities can be detected. As an example, delivery reliability can be found in IS services as well in terms of system and service availability. Both fields face unplanned efforts in terms of transaction costs, miscalculations and high employee turnover rates. In addition, the access to state-of-the-art technologies, the lack of trust and commitment and the risk of theft of intellectual property are factors that are considered as drivers for backsourcing in both IS and manufacturing.

From an economic point of view, Wiesmann et al. [48] conducted a more differentiated review than we did in our analysis by including the categories that are mentioned above. While our paper identifies backsourcing drivers that come from external sources, Wiesmann et al. [48] amplified the influence of political, economic and structural

circumstances, specifying, for example, changes in the international and national economy, political risks, access to qualified personnel, the increasing degree of automation and international differences between productivity rates and work morale among staff. Due to their business character, these arguments can also be considered for the IS field. Competition for resources, sustainability and environmental aspects and difficulties in estimating supply and demand volumes appear to fit into the manufacturing area at first glance but touch the IS area indirectly as well (see [28], [47]).

D. How-Stage: Re-transition process

Comparing our findings in the IS backsourcing literature with those in the ISO literature, the infancy of IS backsourcing becomes visible only in the "how"-stage. Whereas backsourcing results in three publications (two in IS backsourcing), Dibbern et al. [2] identified 36 papers, Wiener et al. [4] considered six papers and Strasser and Westner [23] 13 papers. One reason for this difference is the maturity of ISO research. Furthermore, the limitations, which were mentioned at the beginning of this section, have to be considered, since factors such as supplier selection play an important role in the "how"-stage and might lead to the higher number of findings. Research in the "how"-stage, if narrowed down to IS backshoring literature, is comprised of four areas:

(1) transfer and management of knowledge,
(2) project management needs and challenges,
(3) the relevance of relationship management and
(4) hiring or re-hiring strategies.

(1) Transfer and management of knowledge: As IS services are more integrated into a company's infrastructure than isolated manufacturing activities, a delimitation is difficult to make ([2]). Thus, the transfer of knowledge in an IS backsourcing arrangement must be structured and accurate, due to multiple barriers, such as business requirements, geography or distance, limitations of information and communication technologies, language and problems with sharing beliefs and cultural norms ([46]). Adapting and modifying the approach in Strasser and Westner's [23] systematic literature review on ISO, this section can be divided into *knowledge transfer factors* and *knowledge processes and roles*.

Knowledge transfer factors: Most IS backsourcing and a wide range of ISO studies examine the knowledge transfer process between the client and the vendor and identify central factors that influence this process positively or negatively. As an example, in reviewing two case studies of IS backsourcing, Bhagwatwar et al. [46] argued that high transparency and the willingness to cooperate lead to positive impacts on the transfer, while neglecting the communication and the integration of the employees into the transition process lead to negative impacts. The scope of knowledge, in terms of product specifications and processes, and an environment of clear instructions play a crucial role in transferring concrete knowledge from one entity to another. Indirect influencers are formal factors, such as the level of knowledge on each side, organizational characteristics, and additional efforts for privacy preservation of company-internal data.

Knowledge processes and roles: The knowledge transfer process can be divided into different types and can therefore be explained in different process models (see [49]–[51]). It becomes clear that various types of knowledge exist; hence, different transfer methods should be applied. A prior step to the transfer is to enable the process by sensitizing affected employees on the client and vendor sides to prepare the cooperation and communication on an organizational level. Wang et al. [52] developed a process of boundary formation and spanning activities and defined the role of a boundary spanner, who navigates and negotiates existing boundaries. A second role, namely, the bridge system engineer, is defined. This role is to help minimize all types of issues regarding knowledge gaps and make the client staff aware of cultural differences between the client and the vendor (see [53], [54]).

In their literature review, Strasser and Westner [23] extended Wiener et al.'s [4] findings by identifying additional organizational practices that influence the knowledge transfer. In particular, the relevance of intermediaries and learning activities for successful knowledge transfer was determined. Comparing these findings with the existing IS backsourcing literature, most of the factors that are specified in the ITO literature are also identified but only briefly analysed.

(2) Project management challenges: Adapting and modifying Wiener et al.'s [4] results, the project management challenges for ISO can be divided into three categories: *cultural differences*, *distances* and *psychological contract*. Referring to the definition and characteristics of IS backsourcing, these categories can be applied in this research area as well. In the lessons that they learned from two case studies, Bhagwatwar et al. [46] emphasize the relevance of a guided re-integration process, backed by a backsourcing project team and plan. This team ideally consists of not only executives, managers and technical staff but also the mentioned bridge system engineers. The most obvious tasks of the team are to relay decisions of the vendor to all relevant parties, pay attention to existing and defined security policies and perform the business continuity planning [46]. In addition, it is an unobvious but crucial challenge to lay the groundwork for working and collaborating by defining milestones and responsibilities and overseeing deadlines and costs [4]. On an unconscious level, the project team is responsible for handling upcoming challenges in terms of providing platforms and methods for overcoming any mentality, language or communication barriers that might

hamper the collaboration between the client and the vendor. Coordinating cultural groups over a geographical distance in different time zones while integrating all relevant stakeholders increases the difficulty of the task of project management in IS backsourcing.

(3) Relationship management: As knowledge transfer can be considered the main task and project management the main tool in the "how"-stage, effective relationship management involves an enabler and a facilitator for both aspects. According to [23], relationship management can be divided into *relationship management factors*, *relationship management practices and strategies* and *client and supplier middle-management capabilities and roles*.

Relationship management factors: Since a relationship exists between the client and the vendor, the similarity of factors between IS backsourcing and ISO seems obvious. Primarily, the interests of both client and vendor must be considered since different and possibly hidden motivations drive the engagement on either side [55]. On a more operational and interpersonal level, various aspects have an impact on the relationship management. While trust and the motivation for collaboration lead to a successful relation [56], missing commitment of senior management and weak employee identification influence the relationship negatively [46]. In addition, the various aspects of distance play a role in managing relationships, parallel to the challenges in project management. Since any type of backsourcing has the characteristic of finality, short-term activities seem appropriate, whereas ISO also focusses on establishing long-term strategic partnerships.

Relationship management practices and strategies: Since two organizationally and culturally different groups are in contact, specific practices and strategies appear to be necessary for a collaboration. Abbott and Jones [57] developed a framework for obtaining a better understanding of complex cross-cultural practices and processes. Based on their interviews, Mehta and Mehta [58] emphasized the need for investments in the vendor relationship to minimize the client's risk of relationship breakdown. Such investments may be face-to-face contacts or interactions and the motivation of both the vendor's and the client's employees [59]. In case of a deterioration of the relationship, for example, due a lack of team identity or blockages of communication, Mathew [60], Zimmermann [61] and Butler et al. [17] suggested contingency plans and risk mitigation strategies, such as accelerating the transition.

Client and supplier middle management capabilities and roles: Surprisingly, scant research has been published on middle managers, who execute the outsourcing on an operational level and report to the top management [2]. Willcocks and Griffiths [62] identified the capabilities and roles of middle management for both client and vendor that ensure the effectiveness of an outsourcing arrangement. To clarify the difference from the project management approach that was mentioned earlier, middle managers are domain experts, behaviour managers or governance specialists who are directly confronted with upcoming problems from the operational side. In contrast, project managers are responsible for general organizational issues regarding the project. However, overlaps in roles and tasks exist.

(4) Hiring and re-hiring strategies: Parallel to ISO and manufacturing backshoring research, information on handling human resource capacities is lacking. This phenomenon might occur in ISO, since hiring new staff is an issue of the vendor. Bhagwatwar et al. [46] stressed the relevance of having a strategy for re-transferring existing employees and hiring new employees. Backsourcing without the needed manpower is impossible, which makes it necessary to consider the availability and the need to transfer or hire staff in advance. This need is emphasized by the fact that running in-house IT functions requires people with expertise. The hiring and training of highly skilled staff and service quality assurance are time and cost issues that also must be considered in the backsourcing decision [17].

Comparison to ISO and manufacturing literature: While [2] focused more on conceptualizing and building a relationship between client and vendor, Wiener et al.'s [4] review examined the challenges of offshore relationships, including risk mitigation techniques and success factors. Strasser and Westner [23] extended these findings by specifying a range of factors that emphasize the relevance of communication and commitment of all stakeholders. In addition, they identified additional research fields regarding the role and capabilities of middle management, cross-cultural and organizational learning processes and offshoring attitudes and resulting behaviours that influence relationship management of offshoring initiatives. In their study, Butler et al. [17] stressed the importance of relationship management in terms of investing in the relationship to enable a smooth knowledge transfer and avoid a relationship breakdown during the transition.

IV. Discussion and Conclusions

This literature review presents a consolidated view of the current IS backsourcing field of study and is the first of its kind. Twenty publications critically reflect the state of research of the period between 2003 and 2016. In this article, the current state of the IS backsourcing research stream was reviewed and analysed. By partially referring to the analytical framework of [4], the perspectives of research focus and research approach were adapted. With the help of this modified framework, a common understanding of basic terms and, thus, the basis for the analysis of prior academic IS backsourcing literature was enabled. According to an analysis of the findings, the chosen framework appears to be appropriate and encourages further research in the field along the framework perspectives.

A. Current State of Research

With its first mention in the 2000s, IS backsourcing became a field of interest in the upcoming years. Most of the papers that address the IS backsourcing phenomenon were published between 2003 and 2010, whereas only a few publications that investigated failed outsourcing arrangements appeared from 2014 to 2016. According to the main path analysis by [1], who analysed ISO research from 1992 to 2013, IS backsourcing is only mentioned as an alternative among IS sourcing possibilities.

In total, 20 papers were identified in our literature analysis, which were published between 2003 and 2016 and consisted of eleven conference papers from nine conferences and nine journal papers in nine journals (see Table 2 in the Appendix). With the literature review at hand, one overarching finding becomes immediately apparent. Backsourcing research is in a stage of infancy. This finding is based solely on the number of publications and the foci of the papers compared to the ISO research field; such as in [23], in which the authors were able to identify and analyse 95 articles that were published from 2009 to 2013 for their literature review on ISO. However, with other reviews emerging (e.g., see [63]) the topic of backsourcing seems to gain momentum.

B. Research Focus

The findings demonstrate that the focus of research is the decision process, especially the drivers for enterprises to backsource their IT services (16 papers). Most of those findings address the concrete IS backsourcing subject, whereas three articles instead investigate failures of outsourcing arrangements. Thus, currently, it seems to be the most mature branch of the IS backsourcing research stream. One reason for this domination might be related to how backsourcing is viewed. Initially viewed as a solution for poor service quality and unmet expectations, backsourcing has become a strategy for change and innovation over time. In analysing the first large wave of publications on ISO, Dibbern et al. [2] encountered a similar domination.

Unlike the literature reviews on ISO, few articles focus on the "how" question, which refers to the implementation of the IT services back to the home country of the company. In drawing parallels to other research fields, influencing factors could be found indirectly and partially matched to findings in the IS backsourcing field due to the similarity of various characteristics of transition processes. Accordingly, future research should further address the implementation aspect of the re-implementation stage of IS backsourcing.

According to the search results, a stronger focus should be laid on the implementation phase, to determine what influences the transition phase and what outcomes can be expected. Distinctions between IS backsourcing and backshoring could be examined to a similar extent as in ISO research. Switching the point of view may lead to additional insights. Integrating various stakeholder perspectives might enhance the robustness of IS backsourcing research results. Furthermore, research on hiring and re-hiring strategies should be conducted, both in IS backsourcing and ISO. Having this in mind, more research should be conducted on comparing the phenomena of backsourcing and outsourcing.

C. Research Approach

Most of the reviewed publications make use of empirical research methods (13 papers), which are dominated by interpretive (nine papers) and followed by positivist research (three papers). Interpretive research is conducted across both stages whereas positivist research only considers the "why"-stage. Descriptive research is used only once. Among the empirical research methods, case studies are by far the most popular. Non-empirical research was conducted in seven articles, in each case in a conceptual manner. The allocation of the empirical papers corresponds to the findings of [4] and [23], except that the share of the conceptual papers is higher.

Considering the current predominance from an interpretive epistemological view, a more balanced application of interpretive and positivist methods seems appropriate. As the research field is emergent, descriptive studies should be conducted as well. The obvious dominance of case study research should be complemented by a wider use of other methods (e.g., field study research and action research) and the design of research approaches.

D. Future Research

Apart from the small number of search results for IS backsourcing, future research should primarily consider all perspectives along the multi-perspective framework, following [2], [4]. Thus, one goal might be a higher pervasion of IS backsourcing research to be able to subdivide the two main stages that are specified in this paper into sub-stages according to the five-stage model. To complete the analytical framework of [4], a third perspective should be considered in future, namely, *reference theory*. Matching various approaches and their conclusions with existing theories might lead to additional insights and research questions and could function as an extension of the review at hand. Future research should be aimed at building a fundamental understanding of the phenomenon of backsourcing by varying the points of view, investigating various cases and scenarios and applying various research approaches to verify and extend previous findings.

V. References

[1] H. Liang, J.-J. Wang, Y. Xue, and X. Cui, "IT outsourcing research from 1992 to 2013," *Inf. Manage.*, vol. 53, no. 2, pp. 227–251, 2016. doi: 10.1016/j.im.2015.10.001.

[2] J. Dibbern, T. Goles, R. Hirschheim, and B. Jayatilaka, "Information systems outsourcing," *ACM SIGMIS Database*, vol. 35, no. 4, pp. 6–102, 2004. doi: 10.1145/1035233.1035236.

[3] R. Gonzalez, J. Gasco, and J. Llopis, "Information Systems Offshore Outsourcing," *Inf. Syst. Manag.*, vol. 27, no. 4, pp. 340–355, 2010. doi: 10.1080/10580530903455205.

[4] M. Wiener, B. Vogel, and M. Amberg, "Information Systems Offshoring - A Literature Review and Analysis," *Commun. Assoc.*

Inf. Syst., vol. 27, no. 1, pp. 455–492, 2010.
[5] R. Aron and J. V. Singh, "Getting Offshoring Right," *Harv. Bus. Rev.*, vol. 83, no. 12, pp. 135–143, 2005.
[6] H. T. Barney, G. C. Low, and A. Aurum, "The Morning After: What Happens When Outsourcing Relationships End?," in *Information Systems Development*, G. A. Papadopoulos, W. Wojtkowski, G. Wojtkowski, S. Wrycza, and J. Zupancic, Eds. Boston, MA: Springer, 2009, pp. 637–644. doi: 10.1007/b137171_66.
[7] C. Ebert, "Optimizing Supplier Management in Global Software Engineering," in *Proceedings of ICGSE 2007*, 2007.
[8] T. Kern and L. Willcocks, *The relationship advantage: information technologies, sourcing, and management*. Oxford , New York: Oxford University Press, 2001.
[9] D. Mclaughlin and J. Peppard, "IT backsourcing: from 'make or buy' to 'bringing it back in-house,'" in *Proceedings of ECIS 2006*.
[10] D. Whitten and D. Leidner, "Bringing IT Back: An Analysis of the Decision to Backsource or Switch Vendors," *Decis. Sci.*, vol. 37, no. 4, pp. 605–621, 2006. doi: 10.1111/j.1540-5414.2006.00140.x.
[11] R. Hirschheim and M. Lacity, "The myths and realities of information technology insourcing," *Commun. ACM*, vol. 43, no. 2, pp. 99–107, 2000. doi: 10.1145/328236.328112.
[12] S. F. Wong, "Understanding IT Backsourcing Decision," in *Proceedings of PACIS 2008*, 2008.
[13] J. vom Brocke, A. Simons, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven, "RECONSTRUCTING THE GIANT," in *Proceedings ECIS 2009*, 2009.
[14] H. M. Cooper, L. V. Hedges, and J. C. Valentine, Eds., *The handbook of research synthesis and meta-analysis*, 2nd ed. New York: Russell Sage Foundation, 2009.
[15] J. Stentoft, J. Olhager, J. Heikkilä, and L. Thoms, "Manufacturing backshoring: a systematic literature review," *Oper. Manag. Res.*, vol. 9, no. 3–4, pp. 53–61, 2016. doi: 10.1007/s12063-016-0111-2.
[16] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future," *MIS Q.*, vol. 26, no. 2, pp. 13–23, 2002.
[17] N. Butler, F. Slack, and J. Walton, "IS/IT Backsourcing – A Case of Outsourcing in Reverse?," in *Proceedings of HICSS 2011,* 2011.
[18] L. Fratocchi, C. Di Mauro, P. Barbieri, G. Nassimbeni, and A. Zanoni, "When manufacturing moves back," *J. Purch. Supply Manag.*, vol. 20, no. 1, pp. 54–59, 2014. doi: 10.1016/j.pursup.2014.01.004.
[19] M. Amberg and M. Wiener, *IT-Offshoring: Management internationaler IT-Outsourcing-Projekte*. Heidelberg: Physica-Verlag, 2006.
[20] J. V. Gray, K. Skowronski, G. Esenduran, and M. Johnny Rungtusanatham, "The Reshoring Phenomenon: What Supply Chain Academics Ought to know and Should Do," *J. Supply Chain Manag.*, vol. 49, no. 2, pp. 27–33, 2013. doi: 10.1111/jscm.12012.
[21] M. Westner and S. Strahringer, "Current state of IS offshoring research. A descriptive meta-analysis," in *Proceedings of the First Workshop on Offshoring of Software Development – Methods and Tools for Risk Management*, 2008.
[22] E. Carmel and R. Agarwal, "Tactical approaches for alleviating distance in global software development," *IEEE Softw.*, vol. 18, no. 2, pp. 22–29, 2001. doi: 10.1109/52.914734.
[23] A. Strasser and M. Westner, "Information Systems Offshoring: Results of a Systematic Literature Review," *J. Inf. Technol. Manag.*, vol. 26, no. 2, pp. 70–142, 2015.
[24] J. Dibbern and A. Heinzl, "Outsourcing of Information Systems Functions in Small and Medium Sized Enterprises: A Test of a Multi-Theoretical Model," *Bus. Inf. Syst. Eng.*, vol. 1, no. 1, pp. 101–110, 2009. doi: 10.1007/s12599-008-0008-1.
[25] R. Hirschheim, C. Loebbecke, M. Newman, and J. Valor, "Offshoring and its Implications for the Information Systems Discipline," in *Proceedings of ICIS 2005*, 2005.
[26] R. Hirschheim, G. Beena, and S. F. Wong, "Information technology outsourcing: The move towards offshoring," *Indian J. Econ. Bus.*, vol. 3.2004, pp. 103–123, 2004.
[27] J. Akoka and I. Wattiau, "Developing a Framework for Analyzing IS/IT Backsourcing," in *Proceedings of the 11th International Conference of the Association Information and Management (AIM 2006)*, 2006.
[28] N. F. Veltri, C. S. Saunders, and C. B. Kavan, "Information Systems Backsourcing: Correcting Problems and Responding to Opportunities," *Calif. Manage. Rev.*, vol. 51, no. 1, pp. 50–76, 2008. doi: 10.2307/41166468.
[29] B. Martens and F. Teuteberg, "Bewertung von Backsourcing-Entscheidungen im Umfeld des Cloud Computing," in *Proceedings MKWI 2010*, 2010.
[30] S. Kinkel and S. Maloca, "Drivers and antecedents of manufacturing offshoring and backshoring," *J. Purch. Supply Manag.*, vol. 15, no. 3, pp. 154–165, 2009. doi: 10.1016/j.pursup.2009.05.007.
[31] R. Holz, *An investigation into offshoring and backshoring in the German automotive industry*. Doctoral Thesis, Swansea University, 2011.
[32] S. Kinkel, "Trends in production relocation and backshoring activities," *Int. J. Oper. Prod. Manag.*, vol. 32, no. 6, pp. 696–720, 2012. doi: 10.1108/01443571211230934.
[33] J. Kotlarsky and L. Bognar, "Understanding the process of backsourcing," *J. Inf. Technol. Teach. Cases*, vol. 2, no. 2, pp. 79–86, 2012. doi: 10.1057/jittc.2012.7.
[34] G. R. G. Benito, B. Petersen, and L. S. Welch, "Mode Combinations and International Operations," *Manag. Int. Rev.*, vol. 51, no. 6, pp. 803–820, 2011. doi: 10.1007/s11575-011-0101-4.
[35] J. L. Calof and P. W. Beamish, "Adapting to foreign markets: Explaining internationalization," *Int. Bus. Rev.*, vol. 4, no. 2, pp. 115–131, 1995. doi: 10.1016/0969-5931(95)00001-G.
[36] R. V. Ţurcan, M. M. Mäkelä, O. J. Sørensen, and M. Rönkkö, "Mitigating theoretical and coverage biases in the design of theory-building research," *Int. Entrep. Manag. J.*, vol. 6, no. 4, pp. 399–417, 2010. doi: 10.1007/s11365-009-0122-7.
[37] L. M. Ellram, "Offshoring, Reshoring and the Manufacturing Location Decision," *J. Supply Chain Manag.*, vol. 49, no. 2, pp. 3–5, 2013. doi: 10.1111/jscm.12023.
[38] J. Beardsell, "IT Backsourcing: Is it the Solution to Innovation?," SMC Working Papers Series, No. 02/2010, Swiss Management Center, 2010.
[39] B. B. Nujen, L. L. Halse, and H. Solli-Sæther, "Backsourcing and Knowledge Re-integration: A Case Study," in *Advances in Production Management Systems*, S. Umeda, M. Nakano, H. Mizuyama, H. Hibino, D. Kiritsis, and G. von Cieminski, Eds. Cham: Springer, 2015, pp. 191–198. doi: 10.1007/978-3-319-22759-7_22.
[40] S. F. Wong, "Drivers of IT Backsourcing Decision," *Commun. IBIMA*, vol. 2, no. 14, pp. 102–108, 2008.
[41] R. Chandrasekaran, A. Tayeh, and V. Nagoore, "Understanding Information System Outsourcing Failure: Lessons from a Case Study," in *Proceedings of AMCIS 2007*, 2007.
[42] N. B. Moe, D. Šmite, G. K. Hanssen, and H. Barney, "From offshore outsourcing to insourcing and partnerships," *Empir. Softw. Eng.*, vol. 19, no. 5, pp. 1225–1258, 2014. doi: 10.1007/s10664-013-9272-x.
[43] T. Philip, G. Schwabe, and K. Ewusi-Mensah, "Critical Issues of Offshore Software Development Project Failures," in *Proceedings of ICIS 2009*, 2009.
[44] S. F. Wong, "Bringing IT Back Home: Developing Capacity for Change," in *Proceedings of ICIS 2006*, 2006.
[45] G. P. A. J. Delen, R. J. Peters, C. Verhoef, and S. F. M. van Vlijmen, "Lessons from Dutch IT-outsourcing success and failure," *Sci. Comput. Program.*, vol. 130, pp. 37–68, 2016. doi: 10.1016/j.scico.2016.04.001.
[46] A. Bhagwatwar, R. Hackney, and K. C. Desouza, "Considerations for Information Systems 'Backsourcing'," *Inf. Syst. Manag.*, vol. 28, no. 2, pp. 165–173, 2011. doi: 10.1080/10580530.2011.562132.
[47] N. F. Veltri, "Antecedents of IS Backsourcing," in *Proceedings of AMCIS 2003*, 2003.
[48] B. Wiesmann, J. R. Snoei, P. Hilletofth, and D. Eriksson, "Drivers and barriers to reshoring," *Eur. Bus. Rev.*, vol. 29, no. 1, pp. 15–42, 2017. doi: 10.1108/EBR-03-2016-0050.
[49] J. Chen and R. J. McQueen, "Knowledge transfer processes for different experience levels of knowledge recipients at an offshore technical support center," *Inf. Technol. People*, vol. 23, no. 1, pp. 54–79, 2010. doi: 10.1108/09593841011022546.
[50] Y. Feng, H. Ye, and S. L. Pan, "Delivering Knowledge across Boundaries," in *Proceedings of PACIS 2010*, 2010.

[51] K. Schott, "Vendor-Vendor Knowledge Transfer In Global ISD Outsourcing Projects," in *Proceedings of PACIS 2011*, 2011.
[52] Z. Wang, E. J. Chen, S.-L. Pan, and Y. Wu, "Bridging Boundaries in Offshore Outsourcing Organizations," in *Proceedings HICSS 2011,* 2011.
[53] N. T. Huong, U. Katsuhiro, and D. H. Chi, "Knowledge Transfer in Offshore Outsourcing," *J. Glob. Inf. Manag.*, vol. 19, no. 2, pp. 27–44, 2011. doi: 10.4018/jgim.2011040102.
[54] V. Mahnke, J. Wareham, and N. Bjorn-Andersen, "Offshore middlemen: transnational intermediation in technology sourcing," *J. Inf. Technol.*, vol. 23, no. 1, pp. 18–30, 2008. doi: 10.1057/palgrave.jit.2000124.
[55] S. S. Bharadwaj, K. B. C. Saxena, and M. D. Halemane, "Building a successful relationship in business process outsourcing," *Eur. J. Inf. Syst.*, vol. 19, no. 2, pp. 168–180, 2010. doi: 10.1057/ejis.2010.8.
[56] H. Kefi, A. Mlaiki, and R. L. Peterson, "IT Offshoring: Trust Views from Client and Vendor Perspectives," *Int. J. Inf. Technol. Proj. Manag.*, vol. 2, no. 2, pp. 14–31, 2011. doi: 10.4018/IJITPM.201104012011040102.
[57] P. Y. Abbott and M. R. Jones, "Everywhere and nowhere: nearshore software development in the context of globalisation," *Eur. J. Inf. Syst.*, vol. 21, no. 5, pp. 529–551, 2012. doi: 10.1057/ejis.2012.7.
[58] N. Mehta and A. Metha, "Reducing Client Risks from Offshore IT Vendors' HR Challenges," *MIS Q. Exec.*, vol. 8, no. 4, pp. 191–201, 2009.
[59] A. Boden, B. Nett, and V. Wulf, "Operational and Strategic Learning in Global Software Development," *IEEE Softw.*, vol. 27, no. 6, pp. 58–65, 2010. doi: 10.1109/MS.2009.113.
[60] S. K. Mathew, "Mitigation of risks due to service provider behavior in offshore software development: A relationship approach," *Strateg. Outsourcing Int. J.*, vol. 4, no. 2, pp. 179–200, 2011. doi: 10.1108/17538291111148008.
[61] A. Zimmermann, "Offshoring attitudes, relational behaviours, and departmental culture," in *Proceedings of ECIS 2011,* 2011.
[62] L. Willcocks and C. Griffiths, "The Crucial Role of Middle Management in Outsourcing," *MIS Q. Exec.*, vol. 9, no. 3, pp. 177–193, 2010.
[63] von Bary, Benedikt and Westner, Markus, "Information Systems Backsourcing: A Literature Review," *J. Inf. Technol. Manag.*, vol. 29, no. 1, pp. 62–78, 2018.
[64] L. C. e Silva, A. P. H. de Gusmao, M. M. Silva, T. Poleto, and A. P. C. S. Costa, "Analysis of IT Outsourcing Services Failures Based on an Existing Risk Model," in *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015)*, 2015.
[65] H. Solli-Sæther and P. Gottschalk, "Stages-of-Growth in Outsourcing, Offshoring and Backsourcing: Back to the Future?," *J. Comput. Inf. Syst.*, vol. 55, no. 2, pp. 88–94, 2015. doi: 10.1080/08874417.2015.11645760.

VI. Appendix

TABLE II.
OVERVIEW OF THE RESEARCH PAPERS FROM THE LITERATURE REVIEW

	Applied Research Methodologies					**Drivers for IS Backsourcing**						
	Empirical			**Non-Empirical**		**Why-Stage**			**How-Stage**			
Reference	Interpretivist	Positivist	Descriptivism	Conceptual	Mathematical Methods	Expectation Gaps	Internal or External Changes	Environmental Changes	Transfer and Management of Knowledge	Project Management Challenges	Relationship Management	Hiring and Re-hiring Strategies
Akoka and Wattiau 2006 [27]				x		x	x					
Beardsell 2010 [38]				x								
Bhagwatwar et al. 2011 [46]			x									
Butler et al. 2011 [17]	x											
Chandrasekaran et al. 2007 [41]	x					x						
Delen et al. 2016 [45]		x										
Kotlarsky and Bognar 2012 [33]	x					x	x					
Martens and Teuteberg 2010 [29]				x		x						
McLaughlin and Peppard 2006 [9]				x		x	x	x				
Moe et al. 2014 [42]	x					x						
Nujen et al. 2015 [39]	x											
Philip et al. 2009 [43]	x					x						
Silva et al. 2015 [64]				x		x	x	x				
Solli-Sæther and Gottschalk 2015 [65]				x								
Veltri et al. 2008 [28]				x		x	x	x				
Veltri 2003 [47]		x				x		x				
Willcocks and Griffiths [62]	x					x						
Whitten and Leidner 2006 [10]		x				x						
Wong 2008 [40]	x					x	x					
Wong 2006 [44]	x					x	x	x				
Amount (Σ 20)	**9**	**3**	**1**	**7**	**0**	**14**	**7**	**5**				

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 789–793
DOI: 10.15439/2018F161
ISSN 2300-5963 ACSIS, Vol. 15

Prospective Financial Assessment Based on Real Options in Small and Medium-Sized Company

Bartłomiej Nita*, Piotr Oleksyk*, Jerzy Korczak**, Helena Dudycz*
*Wrocław University of Economics, Wroclaw, Poland
**International University of Logistics and Transport, Wrocław, Poland

Email: {bartlomiej.nita, piotr.oleksyk, jerzy.korczak, helena.dudycz}@ue.wroc.pl

Abstract—**The article presents a prototype of an intelligent Early Warning System based on real option approach to prospective financial assessment of Small and Medium-Sized Enterprises (SME). The described approach constitutes a continuation of development of the intelligent cockpit for managers (InKoM project), the main objective of which was to facilitate financial analysis and evaluation of economic status of a company. The current project is related to the design of smart evaluation of critical financial situations of SME using real options, domain ontology, and AI methods. The content of the knowledge is focused on essential financial concepts and relationships connected with risk assessment, taking into consideration internal and external economic and financial information. A case study based on the real option has been carried out on financial data extracted from financial information system.**

I. Introduction

TODAY, innovative methods combined with advanced financial analysis tools are required to correctly assess critical economic standings of Small and Medium-Sized Enterprises. The main stumbling block and difficulty is that managers of SMEs often do not possess solid background knowledge in financial analysis and new available IT solutions, in particular Decision Support Systems. The problem is often caused by lack of the knowledge required to correctly interpret economic indicators. This knowledge may be improved using real option approach to investment appraisal in SME.

In general, an enterprise works better on the competitive space if it tries to identify development opportunities and threats of disruption a company's leading activity. This requires implementation of prospective financial assessment in SME. Most of SME managers are not skilled enough to understand and respond to threats coming from the business environment. Real options have been applied in practice to solve such issues as just described. In addition, they are treated as the risk management instruments used to assess financial risk of high-risk development projects as well as to influence the company's ability to continue as a going concern in the future [1-2].

Taking into consideration all managerial requests and the complexity of business problems, solutions are needed to integrate managerial knowledge and computational methods so as to support intelligent analysis and decision making [3].

The aim of the paper is to present the of a prototype based on the real option approach that integrates financial knowledge, predictive models, and business reasoning to support financial assessment in Early Warning Systems. The term "real option" can be defined using the analogy to the financial option. Real option therefore means the right of its holder to buy or sell some underlying assets (basic instrument, which is usually an investment project) in specified sizes, at a fixed price and at a given time [4, p. 172]. Generally, it can be said that the real option is the right to modify an investment project in an enterprise [5, p. 269].

Thus, we demonstrate how to employ real option approach in Early Warning Systems to support financial assessment in SME with the aim of avoiding bankruptcy. In our project, it is assumed that financial knowledge is formally defined by the domain ontology. The essential part of the work is to develop a smart solution facilitating automated analysis of information available in financial databases and external data.

The paper has been structured as follows. The first part in a critical way introduces the application of Early Warning Systems in the context of financial assessment. In the succeeding section, the concept of application of real option for the purposes of investment appraisal is discussed. In section three, knowledge conceptualization and reasoning are elaborated, with proposal of an appropriate ontology. Next we present a case study explaining the prototype that refers to prospective financial assessment based on real option approach. Finally, in the last section, some conclusions are drawn and the future directions of the project discussed.

II. Critical analysis of Early Warning Systems in the context of financial analysis

Early warning is a process which allows an organization to consistently anticipate and address competitive threats. As far back as early seventies, managers of firms had started thinking about methods that would allow for early identification of opportunities and threats present in their business environment. It led to the emergence of Early Warning Systems, which were to as early as possible forewarn of approaching threats and opportunities and explore their weak signals. Many methods have been developed to analyze SME performance aimed at creating the Early Warning

System [6]. Unfortunately, they are more often based on past data, and this at present is simply not enough. The essential requirement for SME to survive in a competitive market is development of mechanisms allowing for generation of revenues from core operations in the future. In planning future activities, company's managers emphasize the need to maintain existing customers. If this is not possible, attempts are made to search for new customers. It is also necessary to analyze competitive actions, which in the near future could lead to a significant decrease in market share.

One of the main weaknesses of existing Early Warning Systems is the lack of a formal representation of the knowledge and analytical models that take into consideration both internal and external information. Managers using simple Early Warning Systems receive various alerts, but they don't know which problems should be addressed first. Moreover, these systems do not indicate for managers which suggestions are to be implemented, hence managers have to rely solely on their managerial intuition. It is therefore necessary to extend the EWS functionality.

The proposed prototype is focused on prospective information as well as value embedded in real options. Financial forecasts serve as the basis for the remedial actions that take into account contingent factors. Such activities are focused on searching for value hidden in real options, so as to take advantage of opportunities that may emerge in the future. This is not possible without extending the EWS with regard to external information. This kind of external information is not formalized, thus this extension is another challenge. Moreover, external information may be supplemented with data processing algorithms based on management accounting and finance learning tools.

III. Real options in assessing investments

Standard approach to investment appraisal is based on discounted cash flow methodology, in particular NPV (Net Present Value) analysis. This approach is currently insufficient mainly due to the high volatility of external factors affecting a company [7-8]. The commonly used net present value criterion is currently considered as static mainly because it is calculated at a given moment and does not anticipate changes that may occur in the future. As a result, the NPV criterion does not take into account the opportunity to react to new circumstances, such as [9]:

- an unexpected collapse of the market, which leads to a reduction in the business size,
- significant changes in prices, which may have a significant impact on the profitability of the project,
- an exceptionally favorable situation that allows for expanding the scope of activities.

Taking into account of the limitations of NPV criterion, address suggest to apply the concept of real options[1]. Techniques based on the net present value are still necessary and valuable, hence they should not be underestimated in any case. However, real options allow for a deeper analysis of the investment appraisal issue and somehow expand the traditional methods due to the identification of various investment possibilities embedded in the investment projects. Jahanshahi et al. [12] argues the role that real options can play in SME to increase market orientation and organizational learning, consequently providing a firm with the ability to both attain and sustain competitive advantage, particularly in a volatile environment.

The value of this flexibility is reflected in the option price (option premium); it increases if the probability of receiving new information increases and ability to risk bearing increases. The value of this flexibility is the difference between the value of the investment project with the right of managers to modify the project embedded and the value of the project in the absence of managerial discretionary to modify project. This relationship can be described as follows [see: 13]:

$$S\text{-}NPV = NPV + OV \tag{1}$$

where *S-NPV* – a strategic net present value,
NPV – a standard (static, passive, direct) net present value
OV – an option value.

The lack of flexibility is especially the main factor preventing managers from taking risk. Power and Reid [14] test empirically whether real options logic applies to small firms implementing significant changes (e.g. in technology). Their research findings imply that strategic flexibility in investment decisions is necessary for good long-run performance of small companies.

The valuation of real options is a difficult task and very often impossible to be carried out by manager of SME. It should be noted that value of real options is closely linked with high risk. A manager without advanced financial knowledge can increase the level of risk associated with running a business. Thus, it is necessary to build a prototype that will guide the manager through all the risks associated with the investment project taking into account contingency factors. The prototype also indicates additional opportunities which result from the company's environment. It is also required to create a smart analytical tool that processes signals coming from the environment and integrates them with the real option pricing module.

IV. knowledge Conceptualization and reasoning

Implementation of any development project should be preceded by a multi-faceted analysis confirming its profitability. The objective of analytical activities, mainly based on external sources, is focused on confirmation of the necessity of unconditional implementation of changes in the enterprise. After obtaining external information, it is necessary to integrate them with data describing an entity's potential to implement new projects and solutions.

The prototype includes extended analytical methods. This method is aimed at integrating data from internal reporting with external information. In the prototype, it is assumed that business knowledge is formally described using

1 The term „real option" was initially used in 1977 by S.C. Myers from *Massachusetts Institute of Technology* [10]. This concept was further developed by A.K. Dixit and R.S. Pindyck [11].

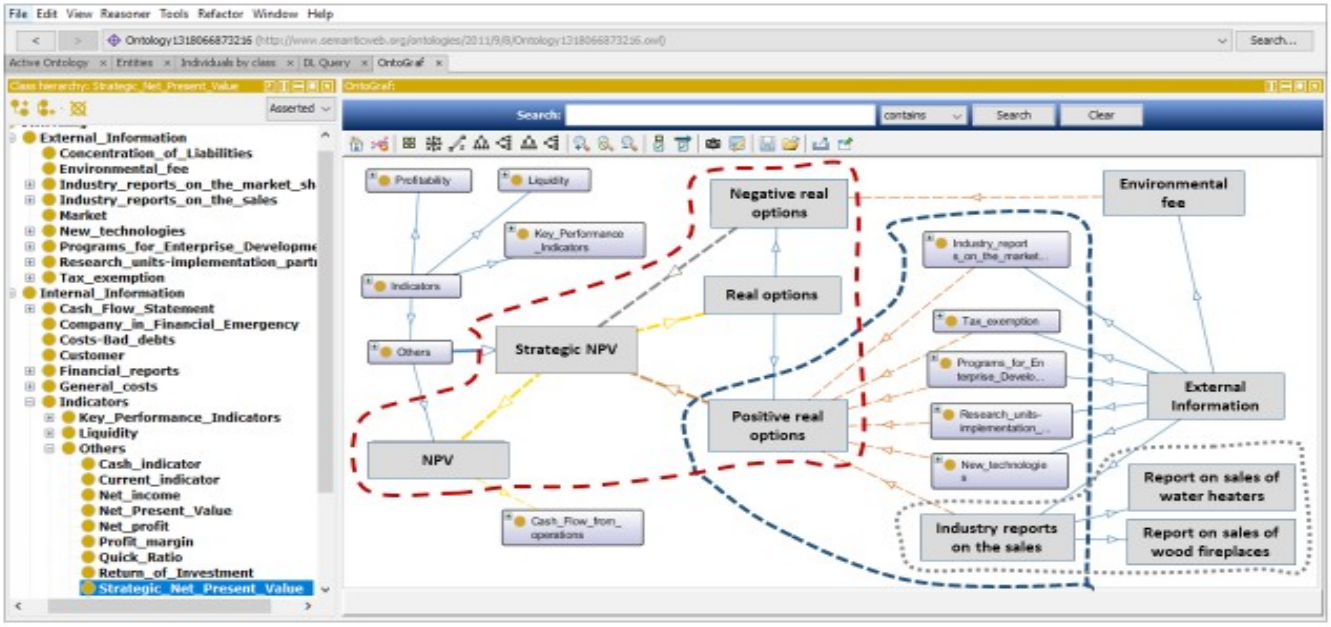


Fig. 1. An example of visualization of a semantic network of *Strategic NPV*
Source: own elaboration using Protégé editor.

ontology[2], which is one of commonly used method of representing knowledge in the information systems. The ontology includes internal information and external information essential for the functioning of the company. A manager can browse hierarchy of concepts, relationships, and annotations. In addition, post conditions such as positive and negative effects of decision process execution can be defined.

The ontology has been encoded using the Protégé platform (http://protege.stanford.edu/). It is important to note that the given ontology describes only static structures, namely the financial concepts and their relationships and external information having influence on the functioning of an enterprise (especially for SME). The ontology presented in Fig. 1 shows a few concepts related to the analysis of the strategic NPV.

Figure 1 illustrates a sample visualization of external and internal information focused on the issue of the strategic NPV. There are two panels on the screenshot. The panel to the left shows taxonomic relations, while the one to the right allows for visualization of taxonomic and semantic relations between defined topics (semantic network visualization). There are two types of lines between topics: (1) the solid line represents a relation subclass-of and (2) the dashed line represents the experts' defined relationships (for example: depends on) on the figure.

Figure 1 presents topics important topics in rectangles for the analysis of the strategic NPV. The presented part of the ontology shows that *Strategic NPV* depends on standard *NPV* and *Real options*, which contain *Positive real options* and *Negative real options*. Positive real options *increase* strategic NPV, while negative real options *decrease* Strategic NPV. Instances of positive and negative real options are *External information* (for example: *Industry reports on the sales*, *Environmental fee*). This part of ontology shows to the manager, that if he wants to calculate *Strategic NPV*, he should estimate *Real options*. The manager can see that he should analyze *External information* affecting the calculation of the positive and negative values of real options. The manager can add, modify as well as retrieve topics related to the problem at hand.

Our proposal to extend the functionality of the system is based on introduction of financial ontology, containing internal and external variables related to real options, as described in Fig. 2. The system processes selected information from financial reports and business environment, subsequently forecasting a company's economic and financial situation. In a situation of a negative forecast, in addition to warning messages, it indicates the possibility of using the real options that would allow a manager to exit a critical situation. The financial ontology not only helps identify the concepts and relationships between them, but it also helps in the interpretation of the current and future situations of the company.

In order to explain the operation of the system, let's describe the vector of the input information as follows:

$$[(f_1, \ldots, f_k), (e_1, \ldots, e_l), (r_1, \ldots r_m), (\delta_1, \ldots \delta_n)] \qquad (2)$$

where $f_1, \ldots, f_k$ denote information from financial reports,
$e_1, \ldots, e_l$ information from the business environment,
$r_1, \ldots r_m$ information about real options,
$\delta_1, \ldots \delta_n$ are estimates of future changes of variable values.

2 Ontology "is an explicit specification of a conceptualization" [15, p. 907].

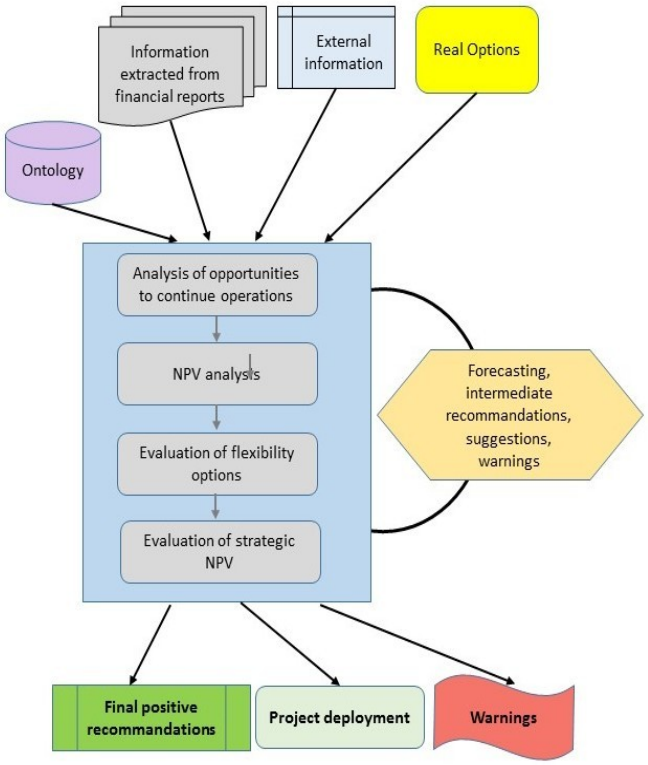


Fig. 2. Functional schema of smart EWS

This information is interpreted by the decision rules described in the ontology, for example:

if f_i ® *threshold f*
then message f-warning else message f-positive

where ® denotes a specific relationship between values of f_i and the threshold.

Depending on the values and number of thresholds, the messages can be more or less varied. If we left the output messages in this form, then the logic of our system would not be different from the classic EWS.

We have introduced several new solutions in the project. The first is the transformation of information qualification into the values of multivalued logic. An example of a transformation with regard to NPV is shown in Fig. 3.

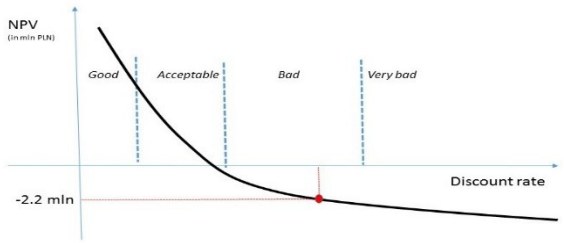


Fig. 3. Example of values transformation

In order to assess the overall situation of the company, these values (Very bad, Bad, Acceptable, and Good) can be transformed into numeric values, for example (*2,3,4,5*), and use in computational models. Additionally, a manager can give weights to variables in the range prototype *[0,1]*, where *0* is a negligible variable, and *1* is a very important variable. In the prototype, the weights of variables may be taken by default or defined by a manager.

With these assumptions, assessment of the company *S* situation can be computed in two ways: as an arithmetic average or as a geometric mean:

$$S = \Sigma (f_i * w_i + e_j, * w_j + \{r_k * w_k + \delta_l * w_l\})/m \quad (3)$$

or

$$S = \sqrt{} f_i * w_i * e_i, * w_j * \{r_k * w_k * \delta_l * w_l\}/m. \quad (4)$$

Note that the evaluation is performed twice, first without taking into account variables related to real options and estimating the increase in values resulting from accepting real options (expression $\{r_k * w_k + \delta_l * w_{lj}\}$, and the second time after accepting the real options. Interpretation of the assessment and its justification are prepared using financial ontology and data from an enterprise's information system.

The design of an inference process illustrating this concept will be discussed in the next section.

V. Case Study

To illustrate the need for real option valuation, we present the case of a project that would be rejected on the basis of traditional analytical methods. Based on valuation of flexible option to expand, we have shown that to avoid going bankrupt, the management should choose_to implement the project.

Assumptions of the case study:

- managers of a manufacturing company producing water heaters and wood fireplaces, while preparing sales forecasts, identify a significant problem with the company's ability to continue its operations,
- managers, based on their expertise and experience, foresee that if they decide to abandon development projects, the company will lose the ability to continue its operations within 5-7 years,
- when planning innovations in the enterprise, a new design of a fully ecological cogeneration fireplace meeting the most stringent ecological standards has been developed,
- the forecasted product cost suggests high selling price that does not allow for launching the project,
- it is necessary to implement changes in production technology, that would make it possible to reduce costs and offer a lower price of the new product, however, NPV analysis indicates that the project would still be unprofitable.

Due to the limited size of the article, it is not possible to indicate a detailed valuation of the real option, and thus strategic value of NPV. Such activities require a large number of calculations, calibration of input parameters, and adoption of discretionary assumptions for the valuation of future benefits. The prepared EWS prototype allows the manager to assign any rank to each source of information.

However, less experienced managers can use the hint embedded in the prototype, which suggests default solutions.

The prototype of the system presented in Fig. 2 contains next analytical steps. The ontology built into the prototype (Fig. 1) explains to the manager the basic concepts and problems associated with the sales profitability. The ontology also presents the knowledge that combines the profitability issue with the investment project appraisal. The manager receives information from the system that it is not possible to conduct development activities. The knowledge contained in ontology explains to the manager the essence of valuation of the flexibility option. The system presents a set of fundamental information needed to evaluate the flexibility option, but the information should be verified by the manager. The manager may or may not expand this information base based on his own expertise. In the analyzed company, the situation allows for initiating preparatory activities to launch the project.

Improvement of the financial situation should be the dominant objective of any manager. It is very difficult for the manager to determine the right moment to implement an investment. Contextual analysis of the impact of real options on a given project is very often beyond the scope of most SME managers. The SME manager is not in a position to take into consideration all the aspects of the development project on his own. It is often necessary to hire a consultant, which is an additional cost. Therefore, creation of a prototype that makes it possible to handle prospective analysis seems indispensable.

VI. Conclusion and Future Works

The main objective of the paper was to present foundations of a prototype based on real option approach that incorporates financial knowledge, predictive models, and business reasoning to support financial assessment in Early Warning Systems. The implemented prototype contains unique methods of prospective analysis used to assess profitability of an investment project. The novelty of the approach consists in applying real options embedded in the Early Warning System. The example is based on real data extracted from a small company. Risk of bankruptcy could be avoided by making decisions based on intelligent in-depth analysis of external information combined with the analysis of financial situation that allows for implementation of corrective solutions.

From a financial perspective, the presented case study supports the conclusion that the decision to undertake any investment cannot be based solely on estimation of standard NPV. It also requires analysis of various external factors determining decision making process. Managers of SMEs may take advantage of the proposed system that integrates financial knowledge and predictive models. Therefore, the system provides knowledge not only on the required internal information from various reports, but also from external information (which are weak signals). The proposed ontology seems to be a promising extension to Early Warning Systems. It not only improves the quality of analysis, but also enhances managerial ability to better understand relations between financial data (internal information) and various factors affecting development of the SMEs (external information).

Further work should be focused on a global process-oriented approach to financial assessment. This will not be possible without large databases of real case studies and use of knowledge possessed by experienced managers and financial analysts. For a company, the multidisciplinary approach to develop the prospective analysis in the Early Warning System could contribute to attainment of a competitive advantage and to increase its financial stability.

References

[1] M. Benaroch, Y. Lichtenstein and K. Robinson, Karl. 2006. "Real Options in Information Technology Risk Management: An Empirical Validation of Risk-Option Relationships," *MIS Quarterly*, 2006, vol. 30 (40), pp. 827-864

[2] G. Favato and R. Vecchiato, "Embedding real options in scenario planning: A new methodological approach, *Technological Forecasting and Social Change*, Elsevier, 2017, vol. 124(C), pp. 135-149, DOI: 10.1016/j.techfore.2016.05.016

[3] J. Korczak, H. Dudycz, B. Nita, P. Oleksyk and A. Kaźmierczak, "Attempt to extend knowledge of Decision Support Systems for small and medium-sized enterprises", in: *Proc. of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Annals of Computer Science and Information Systems, vol. 8, 2016, pp. 1263-1271, DOI:10.15439/2016F181

[4] B. Nita, „*Metody wyceny i kształtowania wartości przedsiębiorstwa*" [Methods of Corporate Valuation and Value-Based Management], PWE, Warszawa 2007.

[5] R. A Brealey and S. C. Myers, *Principles of Corporate Finance*, McGraw Hill, 2003.

[6] A.S. Koyuncugil and N. Ozgulbas (Eds.), *Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection*, IGI Global, 2009

[7] T. Adelaja, *Capital Budgeting: Capital Investment Decision Paperback*, CreateSpace Independent Publishing Platform, 2016

[8] U. Götze, D. Northcott and P. Schuster, *Investment Appraisal: Methods and Models*, Springer, 2015

[9] A. Damodaran, "The Promise of Real Option", in: J. M. Stern, D. H. Chew, Eds., *The Revolution in Corporate Finance*, Blackwell Publishing, 2003.

[10] S. C. Myers, "Determinants of Capital Borrowing", *Journal of Financial Economics*, 1977, vol. 5., pp. 147-175

[11] A. Dixit and R.S. Pindyck, *Investment under Uncertainty*, Princeton University Press. Princeton, New Jersey 1994

[12] A. Jahanshahi, K. Nawaser, N. Eizi and M. Etemadi, "The Role of Real Options Thinking in Achieving Sustainable Competitive Advantage for SMEs", *Global Business & Organizational Excellence*, November 2015, vol. 35, no. 1, pp. 35-44

[13] L. Trigeorgis, *Real Options. Managerial Flexibility and Strategy in Resource Allocation*, The MIT Press, Cambridge 1998

[14] B. Power and G. Reid, "Organisational change and performance in long-lived small firms: a real options approach", *European Journal Of Finance*, September 2013, vol. 19, no. 7/8, pp. 791-809

[15] T. R. Gruber, Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Technical Report KSL, Knowledge Systems Laboratory, Stanford University, 1993, http://tomgruber.org/writing/onto-design.pdf

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 795–799
DOI: 10.15439/2018F138
ISSN 2300-5963 ACSIS, Vol. 15

e-Assessment Management System for Comprehensive Assessment of Medical Students Knowledge

Jaroslav Majerník
Pavol Jozef Šafárik University in Košice, Faculty of Medicine, Trieda SNP 1, Košice, Slovakia
Email: jaroslav.majernik@upjs.sk

***Abstract*—The evaluation of students' knowledge, skills and performance is one of inseparable parts in education process. Except of various teaching approaches, the history brought also various, more or less effective assessment methods. Nowadays, thanks to the learning management systems, the e-assessment methods become more available and widespread across education institutions. The heterogeneity in usage of various assessment methods at our faculty, as well as involvement of natural subjective factors, forced us to search for the solution that can be applicable and acceptable in all study programs, courses and examinations. Furthermore, the system should be secure, intuitive and available for all teachers and their students. In this paper, we describe main principles and methodology of e-Assessment management system we implemented into the medical education for automatic assessment of students' knowledge and competences at the Faculty of Medicine in Košice, Slovakia.**

I. Introduction

PAPER based tests and oral examinations have been used as the main forms to evaluate students' knowledge for many years. Traditionally, teachers used the tests in exams to evaluate learners' knowledge obtained during educational process. Some forms of tests were also involved in course requirements to measure concept assimilation presented in lectures or classes [1]. Various clinical performance assessment tools are also used in practice education [2, 3]. However, there is still no general framework of competency assessment in medicine [4].

Advances in information and communication technologies (ICT) reduced the needs of papers and the time the teachers spent by evaluation of individual tests too. Many educational institutions already discovered advantages of modern innovative digital technologies and adopted some type of smart tools to facilitate assessment. Simulations and work-based assessments methods for specific purposes and clinical performance, including medical history taking, physical examination skills, procedural skills, clinical judgment etc. have also been used and involved in systems of medical education [5, 6]. The common electronic assessment tools are integrated either in Learning Management Systems (LMS) that offer complex modules for teaching, learning and assessment within education institution [7], or they are designed as individual or separate systems to fully manage all assessment needs, and are generally known as Assessment Management Systems (AMS). In both cases, assessment practices serve teachers and students as a part of continual teaching and learning.

In general, the e-Assessment management systems can be classified as systems based either on client-server architecture or as web-based services [8]. Nowadays, the majority of administrators prefer to adopt online available AMS, where everything can be organized through networks and without the needs to use any papers. Except of these environmental factors, there is also no need to install any clients on students' devices. Thus, the tests can be accessed anytime and anywhere, no matter which platform is used to manage assessment procedures.

The higher education benefits from the e-assessment as it assists learning and determines the effectiveness of the education system. e-Assessment systems have a great potential to improve or replace traditional paper-based assessment processes. It is because they allow users development and managing of various types of questions and tests; assigning of students to the tests/exams; setting of dates, times and places/rooms for the tests/exams; summarizing of tests results; analysing of questions' quality etc. In addition, a well implemented e-Assessment system and understood by the teachers can save the time needed to organize and evaluate exams. In this point of view, their performance is also positively affected as the marking load is reduced and the results are available immediately after the exam is completed [9]. However, the assessment should have clear purpose and has to match the educational programmes and learning outcomes. Thus, any assessment method must be reproducible to show similar results on different occasions and valid to reflect appropriate

This work was not supported by any organization

representation of educational content. Finally, the e-Assessment systems are considered comfortable in all assessment related tasks, including measurement and documentation of knowledge, skills, and attitudes of individual learner and/or learning community [10, 11, 12].

The e-Assessment technology principles should be based on methodology that, except of others, allow examiners to create a bank of questions, to generate different types of tests, to mix questions and/or answers in the tests, to specify exact dates and times when the learners must take the exams and to automatically score and share test results to learners. The capabilities of the e-Assessment systems should be also focused on the ways how the users interact with the systems and how it is adopted to their needs.

Aiming to solve the assessment issues in a complex and comprehensive way, we had to consider various factors and questions. Is there any system that will meet the requirements of our teachers and that can be integrated at institutional level? Do the systems allow specifying assessment plans in relation to the learning outcomes? How to grant the permissions of different groups of users to access the system? These and many other similar questions were solved and discussed during our initiative and resulted in a satisfying solution that was accepted very well by both the teachers as well as by the learners.

II. Materials and Methods

To find and/or design the AMS that will be well accepted by all of our teachers, we conducted a survey in which we wanted to discover what kind of assessment methods are currently used, what are the preferred forms to evaluate students' knowledge during diagnostic, formative and summative exams, and what are/should be the most preferred features of assessment system.

The survey was realized online using Google forms, and 65 teachers of our faculty participated on it. The findings illustrated wide usage of ICT in everyday praxis, however, the engagement with e-Assessment was only 12,3%, i.e. only 8 of 65 respondents actively utilize some electronic form to evaluate students' knowledge. The responses resulted in the list of features our teachers require from AMS. These features included: possibilities to test large number of students at the same time; place/room independence; protected access and high security of all exams related data; repository of questions and tests; multimedia support in tests; limited access to registered students only; easy to use interface in national language; reporting per examination; and not surprisingly for academic environment, low or no financial expenses.

Except of the above mentioned features, the technicians had to consider numerous technical and administrative related aspects too. Thus, the fully functional e−Assessment system required to solve the tasks related to the safe and reliable server(s), network infrastructure, computer classrooms, and professional administrative staff support.

Comparing the features, technical requirements and supporting documentation of various commercial (AEFIS, beSocratic, Blackboard Learn, Digication AMS, eLumen, LiveText, rGrade, Taskstream) and open-source (openIGOR, Rogō, Unicon, TAO) assessment systems, we decided to test the Rogō system. The Rogō AMS was developed at the University of Nottingham together with partner institutions, now involved in development community. The results of the tests and the functionalities offered in Rogō convinced us to integrate the system into the ICT infrastructure of the faculty including Slovak language pack developed during testing phase. Our decision was supported also by abilities of the system to integrate third party systems, including LDAP authentication and functionalities, which allow a VLE or other LMS to launch and single sign into Rogō.

III. Results

The integration of AMS was fully adopted to the faculty infrastructure and requirements. The hierarchic structure reflects the faculty units, study fields, courses with learning objectives, different assessment methods and the users with different roles in the system as it is shown on Figure 1.

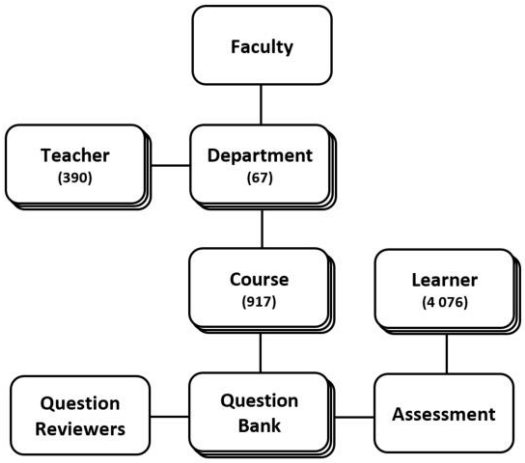


Fig. 1 The structure of e-assessment management system integrated at the Faculty of Medicine in Košice.

All authorised users of the system (teachers, question reviewers and learners) were connected to the accounts of faculty's LMS. Thus, 390 teachers and 4,076 learners were able to use the system without any registration procedures. Similarly, the full list of all courses (917) was imported to the system. Information related to the course registration lists ensured the teachers create questions and examination papers only within their courses and the learners do the exams only in courses they are enrolled in.

Depending on the course management, the Question Banks of particular courses were created by guarantors or by the team of teachers associated with courses. Naturally, the questions can be imported and/or added manually if there is no previously created electronic list of questions. Almost any type of questions is supported that makes the assessment easily adjustable to various types of courses as well as their

learning objectives. Except of commonly used Multiple Choice Questions (MCQ) the teachers can create questions like Area, Calculation, Dichotomous, Extended Matching, Fill-in-the-Blank, Image Hotspot, Labelling, Likert Scale, Matrix, Multiple Response, Random Question Block, Ranking, Script Concordance Test (SCT), Textbox or True-False as it is defined in the Table 1. The users can combine all these question types in Random Question Blocks if there is a requirement to organize exams with randomly generated questions. Once the questions are stored in the Question Bank then it is possible to export them to external QTI or Rogō files and use them in other systems or in other Rogō instances.

Considering various purposes for which the students are assessed and relations to in-course or end of course teaching activities, there was a need to organise different types of assessment. The most frequent types included summative, formative and diagnostic exams. In summative assessment, the learner performance against the standard knowledge is awarded by grades. Then, the grade can either be a part of in-course assessment, or assessment at the end of a course. Formative assessment is organized during the course, and provides feedback to learners. While the summative assessment is used for certification, the formative assessment helps students improve their learning as the failure rate can be reduced and the performance can be increased. Diagnostic assessment is used to evaluate the level of learning that has been achieved by learners. In general, it can be used at the beginning of the course to determine the level of knowledge, or at the end of lessons to know how the learners understood the topics. However, diagnostic assessment does not provide tools of feedback as it is in formative assessment. Individual types of supported assessment methods are show and described in Figure 2.

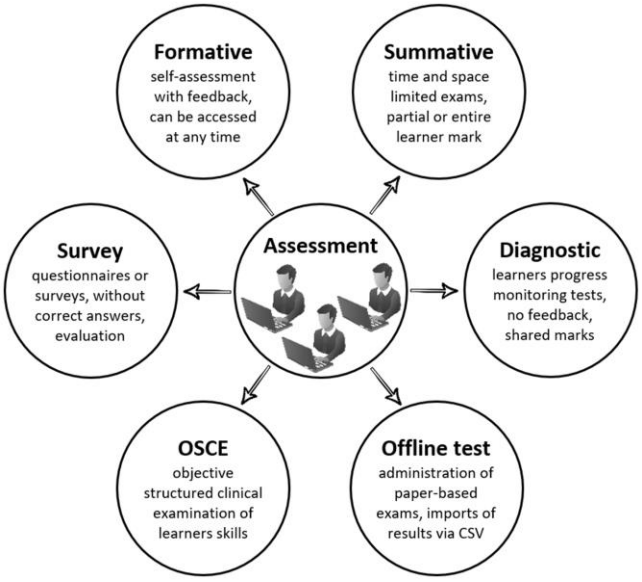


Fig. 2 Assessment methods integrated in e-assessment management system.

The questions for various assessments can be chosen from the same Question Bank of the course or group of courses.

TABLE I.
TYPES OF SUPPORTED QUESTIONS IN E-ASSESSMENT SYSTEM.

Question type	Purpose and description
Area	To draw a shape around a specified part of an image. The most commonly used formats including JPG, GIF and PNG are supported.
Calculation	To use one or more variables (random values) specified in the task by the teacher and to define formula for verification of calculated value given by the learner in the answer field.
Dichotomous	To present a lead-in question together with a number of stems displayed below. The learner must select either True or False for each stem.
Extended Matching	To present multiple scenarios based around a common theme. Each question has the same list of options from which the learner is asked to choose the answer.
Fill-in-the-Blank	An alternative to open ended question. Used to allow learners fill-in the blank textboxes or to select answer from dropdown lists.
Image Hotspot	To identify parts of the graphic, e.g. anatomical structures, body regions etc. One question may cover up to 10 different items to be identified by the learner.
Labelling	To place labels (one label can only be used once), using drag and drop method, to the spaces shown over the graphic.
Likert Scale	To list the answers with the support of both Likert Scales and Semantic Differential questions.
Matrix	To reduce test space via better visual appearance. The questions are presented in rows with possible answers as columns. Radio buttons are used as only one answer per row can be selected.
Multiple Choice	To choose correct answer from up to 20 options. Radio buttons are used in the interface so that only one option can be selected.
Multiple Response	To identify various number of correct options. Each option can be selected or unselected.
Ranking	To see if the learner can put various options in the correct order.
Script Concordance Test	To assess reasoning skills in clinical situations, specifically those with uncertain scenarios.
Textbox	To collect written answers, e.g. open-ended textual questions used for surveys.
TrueFalse	To confirm statement that is displayed with an option that is either True or False.

The system holds large amount of highly important data which must be kept safe at all times. Therefore, the users' data, question banks, exam tests, results as well as all the information stored in AMS are secured using several protection levels. From the security point of view, it is very important that the summative exams are not available anytime and anywhere. The students should not find the tests before the exam dates and the results must be delivered to them securely. On the other hand, the security issues are not necessary to be so strict in formative assessments.

Summative assessments can only be taken by learners assigned to the course during the time allocated to the exam in specific allocated room or place. Thus, the summative tests are not accessible to the students anywhere and at any other time. To increase security, the tests and all questions are locked and cannot be amended to ensure that the questions in the bank accurately match the results of the exam. The protection levels used in summative exams are shown in Figure 3.

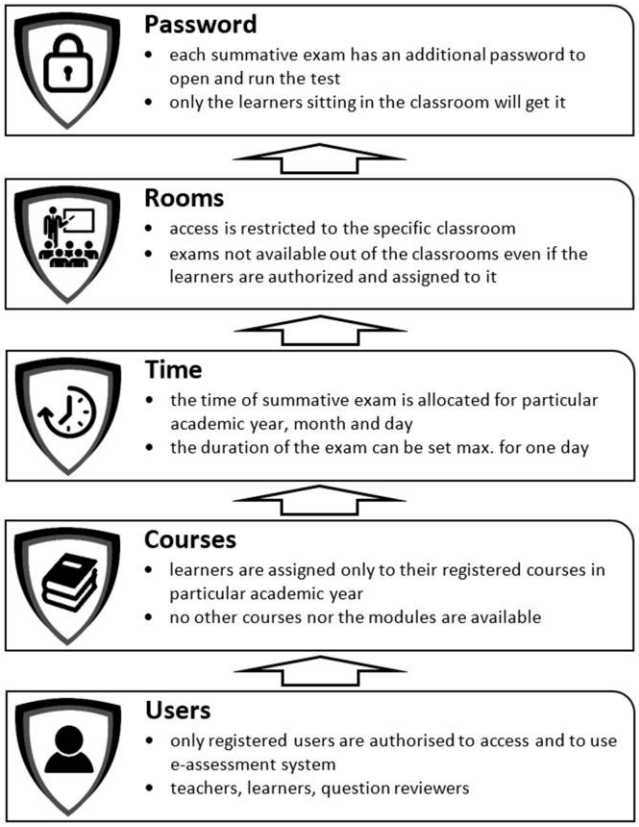


Fig. 3 Security levels applied in summative assessment.

Despite of combination of various security levels there were still some doubts of teachers related to the personal authentication of learners to be sure that the learner completing the assessment is learner that confirmed its identity. Regarding the importance of particular type of assessment, the summative types have to be delivered under invigilated conditions using secure systems. Other assessment forms, where no grading of the results is required, need not to be additionally secured. Thus, for example the formative assessments can be opened to be completed anytime, anywhere and even using learners own devices connected either to the faculty or commercial network. On the other hand, all summative exams at the faculty are organized using advanced mechanisms for personal as well as for equipment identification. Figure 4 shows main concept of additional summative assessment security mechanism we implemented to ensure the summative exams are performed personally by the learners.

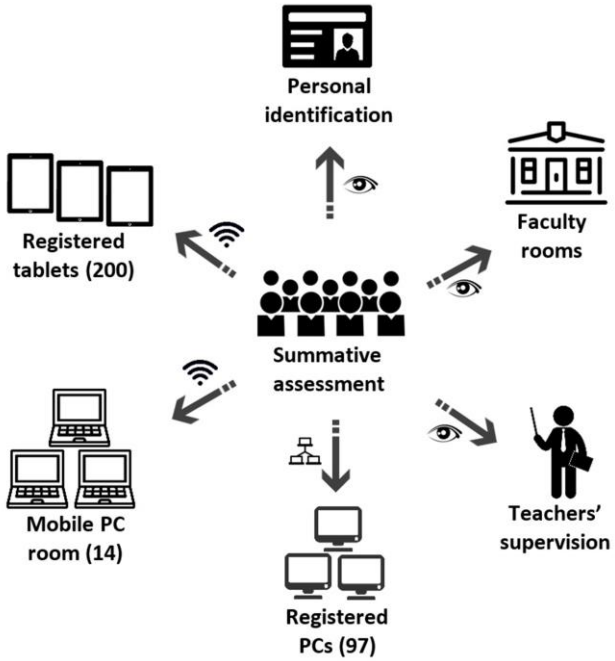


Fig. 4 Additional security mechanisms incorporated into the summative assessments.

All summative exams are organized only in designated faculty computer classrooms and/or lecture halls depending on the size of tested group. The learners are identified by the teacher(s) before they will enter the examination room. Learners' personal identification cards can be used in all lecture halls to register attendance. One or more teachers supervise the summative examination and offer the support to the learners if they have some technical problems during exam. Six computer classrooms with 97 PCs all together are used to test smaller groups of learners. All computers are protected and the internet connection is blocked. The e-Assessment is the only available service during the exams. If the test is restricted to the particular room, then it is not possible to see and open it in another room even if the learner is authorized to perform it. Small groups of learners can be also tested everywhere at the faculty using mobile computer classroom where the connection to the e-Assessment system is realized through protected and hidden WiFi network that is a part of that mobile classroom. However, the biggest challenge was to assess the mass of all learners in particular study field. Therefore, we built a network of secured wireless hotspots across main lecture halls (12 all together). This network has also restricted

access limited to the e-assessment system. In these lecture halls, the learners are doing the summative exams using tablets (200 learners can do the exam in one lecture hall). The tablets are set to access the exams only and everything is preloaded. So the learners are only asked to login to the system, to enter the exam's password that is announced by the teacher once the exam will start and to do the exam.

The AMS is also prepared to solve some problems in the case of unexpected events. The most frequently mentioned doubts of the teachers during their first exams related to the network failures or problems with computer hardware. The system registers each activity of the learners during exam. Therefore, if there is a network failure or computer related problem and AMS cannot be reached, then the assignment can be extended until the problem is solved. Then, the learner can be logged into the system again and continues to solve the exam with all previously marked answers. However, during almost two years' experience we noticed only one problem related to WiFi failure and no problems with PCs. So, the likelihood of such failures is very low.

The system was successfully implemented into the faculty infrastructure and the number of involved teachers is continuously growing. During the period of almost two academic years, the teachers generated almost 1,200 summative exams with more than 8,500 tests and more than 46,000 questions in their question banks.

IV. Conclusion

Implementation of AMS into the education at our faculty minimized the subjective assessment factors and saved the time of our teachers. Of course, many of them disagreed when they started to use it. Initially, they were loaded by the same tasks and problems as it was in paper-based forms. It was because they had to spend time by preparing questions and organizing of all assessment issues. Teachers mind was changed once they understood this is a long-term investment, in which the lifecycle of e-assessment material will save considerable development and supporting workload. Integration and adaptation of AMS brought also many other advantages, as reported by the teachers. These include possibilities to generate both the summative and formative exams with various types of questions; to follow progress in individual learners through stored results; to obtain course feedback or to identify problematic parts in taught topics via analysis of collected answers.

In the next stage of our work we plan to increase the awareness of formative assessment benefits among our teachers to be utilized more frequently in their curricula. The great potential of formative assessment is in instant feedback and continuous monitoring of learners' progress through which they can identify areas of their weakness and are motivated to study for better understanding of particular topics before final summative exams will take place.

The teachers have variety of reasons to use assessment tools, including to pass or fail students, to grade students, to select best ones for future courses, to prove what students have learnt, to reveal strengths and weaknesses of both students and courses, and many others. When implementing an AMS, it is necessary to clarify requirements and needs of particular educational institutions and staff working at these institutions. Only the well fitted system can be accepted across whole institution, can satisfy the need of users and may have positive effect on overall performance.

Acknowledgment

Results presented in this work were obtained with the support of the national agency's grant KEGA 017UPJS-4/2016 "Visualization of education in human anatomy using video records of dissections and multimedia teaching materials".

References

[1] Y. Lee, "Assessment Management System based on IMS QTI 2.1", International Journal of Software Engineering and Its Applications, Volume 8, Number 1, 2014, pp. 159-166, http://dx.doi.org/10.14257/ijseia.2014.8.1.14.

[2] A. O'Connor, O. McGarr, P. Cantillon, A. McCurtin, A. Clifford, "Clinical performance assessment tools in physiotherapy practice education: a systematic review", Physiotherapy 104, 2018, p. 46-53, https://doi.org/10.1016/j.physio.2017.01.005.

[3] J. Kubicek, T. Rehacek, M. Penhaker, M., I. Bryjova, "Software simulation of CT reconstructions and artifacts", Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, Volume 165, 2016, Pages 428-437, http://dx.doi.org/10.1007/978-3-319-29236-6_41.

[4] E. Sureda-Demeulemeester, C. Ramis-Palmer, A. Sesé-Abad, "The assessment of medical competencies", Rev Clin Esp. 2017, 217(9), pp. 534-542, http://dx.doi.org/10.1016/j.rceng.2017.05.004.

[5] J.J. Norcini, D.W. McKinley, "Assessment methods in medical education", Teaching and Teacher Education, 23, 2007, pp. 239-250, http://dx.doi.org/ 10.1016/j.tate.2006.12.021.

[6] S.C. Daly et al., "A Subjective Assessment of Medical Student Perceptions on Animal Models in Medical Education", Journal of Surgical Education, Volume 71, Number 1, 2014, pp. 61-64, http://dx.doi.org/10.1016/j.jsurg.2013.06.017.

[7] J.G. Moura, L.O. Brandão, A.A.F. Brandão, "A web-based learning management system with automatic assessment resources", Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual. IEEE, 2007.

[8] O.F. Bukie, "Understanding Technologies for E-Assessment: A Systematic Review Approach", Journal of Emerging Trends in Computing and Information Sciences, Vol. 5, No. 12, 2014, pp. 936 – 947.

[9] Y. Lee, "Assessment Management System based on IMS QTI 2.1", International Journal of Software Engineering and Its Applications Vol.8, No.1 (2014), pp.159-166, http://dx.doi.org/10.14257/ijseia.2014.8.1.14.

[10] M. Amelung, K. Krieger, and D. Rosner, "E-Assessment as a Service", Ieee Transactions on Learning Technologies, Vol. 4, No. 2, 2011, pp. 162 – 174, http://dx.doi.org/10.1109/TLT.2010.24.

[11] J. Živčák, R. Hudák, T. Tóth, "Rat skin wounds tensile strength measurements in the process of healing", IEEE 10th Jubilee International Symposium on Applied Machine Intelligence and Informatics, SAMI 2012 – Proceedings, 6208996, pp. 389-392.

[12] D. Schwarz, P. Štourač, M. Komenda, H. Harazim, M. Kosinová, J. Gregor, R. Hůlek, O. Smékalová, I. Křikava, R. Štoudek, L. Dušek, "Interactive algorithms for teaching and learning acute medicine in the Network of Medical Faculties MEFANET, Journal of Medical Internet Research, 2013, 15 (7), art. no. e135. http://dx.doi.org/10.2196/jmir.2590.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 801–810
DOI: 10.15439/2018F217
ISSN 2300-5963 ACSIS, Vol. 15

Collective clustering of marketing data— recommendation system Upsaily

Maciej Pondel
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: maciej.pondel@ue.wroc.pl

Jerzy Korczak
International School of Logistics and Transport
ul. Sołtysowicka 19B
51-168 Wrocław, Poland
Email: jerzy.korczak@ue.wroc.pl

***Abstract*—The article discusses the importance of the recommendation systems based on data mining mechanisms targeting the e-commerce industry. The article focuses on the use of clustering algorithms to conduct customer segmentation. Results of the operation of many clustering algorithms in segmentation inspired by the RFM method are presented, and the method of collective clustering using the positive effects of each algorithm is separately presented.**

I. Introduction

THE first seminars and conferences of the 90s on advisory systems [1],[2],[3] were a significant stimulus for the rapid interest in the methods and techniques of automation of recommendations not only in practice, but also by research. In recent years, under the influence of IT development, social networks, and artificial intelligence methods, the concept of the recommendation system and the scope of its main functionalities has significantly expanded. Today, the recommendation system constitutes a complex interactive system that allows one to determine the rank of a product or preferences that the customer should assign to a given product or group of products [4]. In the literature, this system is considered in three main perspectives. From the managerial perspective, the recommendation system is a decision support system that uses large, heterogeneous data and mechanisms generating recommendations related to the sales strategy and promotion of the products offered. From the client's perspective, it is an advisory system facilitating selection of products in accordance with one's interests, needs, and preferences. From an IT perspective, the recommendation system is an interactive computing platform containing a number of data analysis and exploration models, integrated with transactional systems of the online store and the environment. This platform must guarantee not only access to various information resources, but also scalability of applications operating on a large number of information collections.

The specific economic benefits of a personalized recommendation achieved by e-commerce tycoons (Amazon, Alibaba, eBay, Booking, etc.) have proven the increasing effectiveness of recommendations systems. It has resulted not only in increased sales and marketing effectiveness, but also in significant analytical and decision support for marketing managers. Modern recommendation systems are not limited to giving the recommendation *"You bought this product, but others who bought it, bought / watched X, Y, Z products"*. Many of them have based their recommendations on the customer profile, product characteristics, behavioral, and psychological analysis of customers.

Currently, the systems are distinguished by four categories of advisory mechanisms: recommendation by collaborative filtering of information, content-based recommendation, knowledge-based recommendation, and hybrid recommendation [5],[6],[7],[8],[9],[10]. Recommendation by collaborative filtering is the most common method based on recommending products highly rated by clients with similar profile and preferences [11],[12],[13]. The key issues here are: designation of the similarity between clients and choosing the customer segmentation method. These issues will be discussed in more detail later in the article. The content-based recommendation is founded on the analysis and data mining of products purchased by the customer [1],[14]. In contrast to the previous method, the key issue here is to analyze a customer's purchase history and determine the similarity of the products. The third group of methods builds recommendations based on analysis of product features with reference to its usefulness for the client [15]. In order to take advantage and reduce the negative features of the aforementioned methods, hybrid recommendation systems are increasingly being designed [16],[17].

For several years, we have also been observing a growing interest in recommendation systems by owners and managers of online internet shops in Poland. In 2010, every third online shop used a recommendation based on a simple analysis of CBR and Business Intelligence systems [18]. In recent years in Poland, artificial intelligence, personalized recommendation, and digital marketing have dominated the orientation of developers of e-commerce systems which until recently had focused on the efficiency of shopping services [19]. Currently, almost all big online stores use recommendation systems. However, these systems are to a large extent based on a simple business analytics, limited computational intelligence and reduced possibility of dynamic customer profiling.

The aim of the article is to present methods of analysis and profiling of clients available in the Upsaily[1] recommendation system targeting online internet stores. It is a hybrid system combining recommendation techniques through collaborative filtering and through contextual analysis. In the development of recommendations, in addition to transactional data, the system also uses geo-location and social network data. The data is a source of information for many clustering algorithms in the system. These algorithms can work autonomously or collectively, cooperating with each other in order to achieve semantically rich segmentation that is interesting in business interpretations. This second approach is the subject of the article. Although the source data set is the same, the innovativeness of the solution manifests itself in the selection of algorithms; each of them was selected from a different computing class and applies different similarity criteria. Among the algorithms, in addition to the commonly used k-means that uses Euclidean distance measure, we chose for the Gaussian Mixture Model based on probability distributions the DBSCAN algorithm taking into account the density of observation and the RMF involving the manager engagement. The unification of clustering results in our application is specific to the e-commerce applications – not all the clusters are used, but only one or several clusters. The cluster selection criteria include both statistical metrics as well as external, mainly economic, criteria.

The structure of the article is as follows. The next section describes the main functionalities of the Upsaily recommendation system and sketches its functional architecture. The third section defines the problem of individual and collective clustering together with descriptions of the applied algorithms. The concepts of similarity and criteria for unification of clustering results have also been outlined. The last section of the article describes the experiments carried out and further shows the advantages of collective clustering on real marketing data.

II. Functional Architecture and Functionalities of Upsaily System

The Upsaily system, based on the B2C model, is oriented towards current customers of the online internet shops. In the system database, not only all customer transactions are stored, but also basic data about their demographic and behavioral profile. The system is able to record customer reactions to offers directed at them through various contact channels. Functionally, the system can be classified as a Customer Intelligence solution, i.e. the one whose primary interest is current customers, and the aim is to increase customer satisfaction that translate into increasing turnover through the Based on literature [21],[22],[23],[24] and drawing conclusions from the research carried out as part of the RTOM project [25], the schema of advanced data analysis in marketing has been proposed (fig. 2). The schema is helpful in organizing marketing activities. Depending on the specific purpose, a group of clients to be covered by the campaign should be selected. In general, for the defined clients, the subject of the campaign is selected, e.g. product groups that

customers making follow-up purchases, increasing the value of individual orders by cross-selling or more valuable products (up-selling). The immediate goal of the system is not to help in acquiring new customers. The Customer Intelligence approach is related to conducting analytical activities leading to creation of a clear image of the customer so that one can find the most valuable clients and send them a personalized marketing message [20].

The results of research conducted as part of the RTOM project on Polish online stores operating in various industries that showed that in each of them over 75% of all customers are one-off customers, meaning they never returned to the store after making a purchase form the basis of such orientation of the system. Analysis of the average value of the order for a one-off customer shows that it is lower than for customers who make subsequent purchase. Interestingly, it can be noticed that the general trend of an increase in the average value of the order with the increase in customer loyalty expressed in the number of purchases made by them. This observation is presented in Figure 1. The average value of orders have been hidden due to the company's confidentiality. From this observation, it was concluded that it is worth sacrificing the resources of the online store to build customer loyalty, for the simple fact that a loyal customer is ultimately more valuable than a one-off customer.

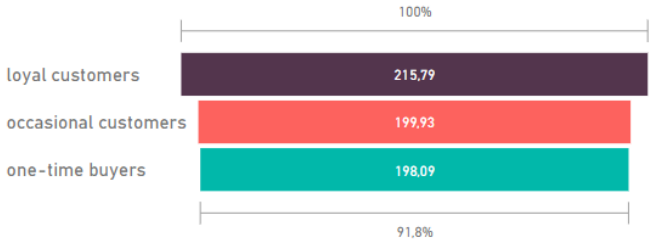


Fig. 1 Graph of the dependence of the average order value on the total number of orders placed by customers.

It should also be pointed out that acquisition of a new customer is always related to the extra cost to be incurred to reach the customer with the marketing message in a selected medium. Usually by acquiring a client then sending them a general message. Without knowing the customer's previous transactions, we are unable to propose an effective offer tailored to client's preferences, therefore in many cases the presentation of a marketing message will not cause projected customer reaction. In case of communication with current clients whose contact details are available and for whom all necessary marketing consents are established - at least at the assumptions level, it can be stated that reaching the customer should cost significantly less and the effectiveness of messages should be definitely higher.

[1]Upsaily system was developed by the Unity S.A., Wrocław, in the framework of the Real-Time Omnichannel Marketing (RTOM) project, RPO WD 2014-2020.

they will potentially be interested in. The final stage is defining the conditions under which customers will be offered participation in the campaign. As the schema shows, cluster algorithms have a wide application in this approach, and this will be shown later in the article.

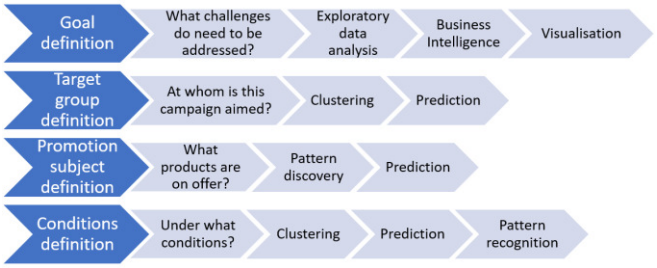


Fig. 2 Stages of building a marketing message with the proposal of using methods and tools for data analysis.

The functional architecture of the recommendation system Upsaily is presented in Figure 3. The Upsaily system collects data from many sources, but the basis of its analysis is transactional data. Data from other sources such as marketing automation systems, social media, systems analyzing activity on the store's website enrich the customer profile and, thus, expand the set of input data for analytical modules that, thanks to them, are able to provide better analyzes and better predictive models. The research platform on which the experiments are carried out has a significant place in the architecture of the system.

These experiments are evaluated in terms of business suitability and when their effects are positive, then they are transformed into regular modules operating in a production manner.

The system information outputs are integrated with:

- Marketing panel or application presenting the results of conducted analyzes, visualizing identified trends, found patterns, and segmentation effects. The recipient of this application are primarily managers and marketing analysts who, in using it, expand their own knowledge on the clients and their behaviors,
- A real-time recommender, an application whose aim is to offer an online store an offer that is as congruent as possible with its needs.
- Module "*campaign for today*", which is based on discovered trends and customer behavior patterns, at the moment of launching it is able to automatically indicate groups of customers, and the product that they may be interested in at that moment.

The results of the Upsaily system will be detailed in the next sections of the article.

III. COLLECTIVE CLUSTERING ASSESSMENT METHODS

There are many algorithms that can be used in collective clustering approach [26], [22], [23]. In the project the composition idea was based on maximum variability and differentiation of clustering paradigms. Therefore the following algorithms were chosen:

- k-means based on the Euclidean distance between observations,
- Bisecting k-means acting on a similar basis to k-means, however, starting with all the observations in one cluster and then dividing the cluster into 2 sub-clusters, using the k-means algorithm,
- Gaussian Mixture Model (GMM), which is a probabilistic model based on the assumption that a particular feature has a finite number of normal distributions,
- DBSCAN identifying clusters by measuring density as the number of observations in the designated area. If the density is greater than the density of observations belonging to other clusters, then the defined area is identified as a cluster.

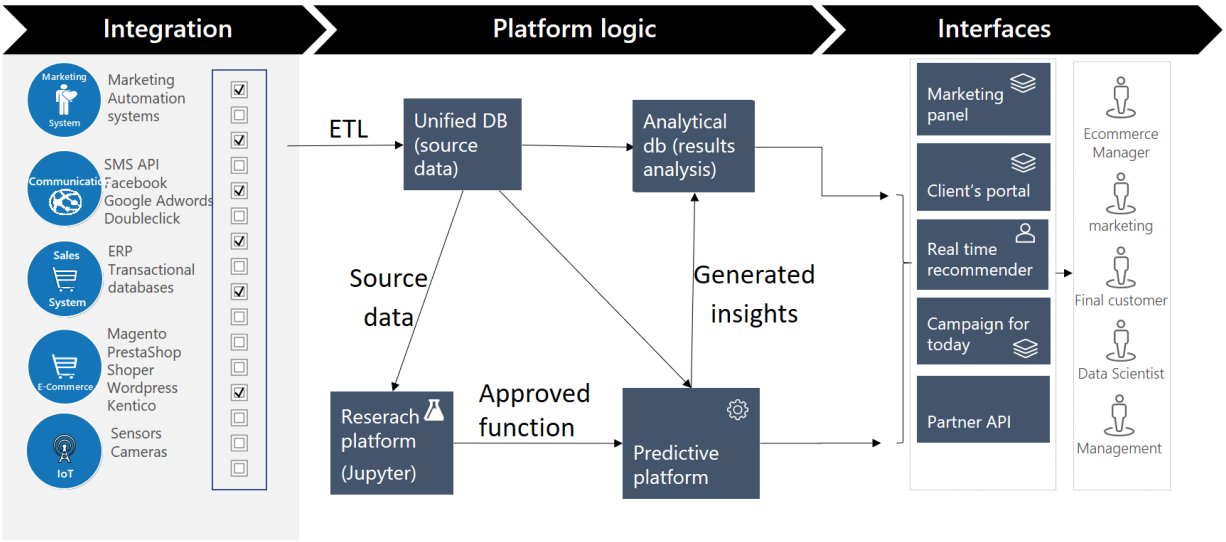


Fig. 3 Functional architecture of the Upsaily system.

Usually the results of clustering algorithms are evaluated according to internal and external criteria. The internal criteria relate to the hierarchy of clusters, taking into account the similarity of observations within clusters and the similarity

between clusters. The Davies-Bouldin[2] and Dunn[3] metrics are usually applied for assessment measures. In addition to the mentioned measures, other functions of assessment are used, such as the silhouette index, measures of cluster cohesion, cluster separation measure, and intra-class scattering matrix [26],[27].

According to external criteria, the results of clustering are evaluated using external data, not considered in the clustering process. Such data are observations which membership in the cluster is assigned earlier by experts. Then the assessment of clustering results from comparing of the content of clusters marked by experts with clusters created by the algorithm. Among the measures used, one should mention the clusters homogeneity index[4] , Jaccard index[5], Rand index[6] . In addition to the specified measures of the assessment, other indicators are also used, such as Kappa, F-score, Fowkes-Mallows index, etc. [26],[27].

In the case of using many clustering algorithms, the obtained results usually differ from each other not only by the number and hierarchy of clusters, but also by the allocation of observations to clusters. In the article, we treat the set of algorithms as a collective of experts whose task is to make the grouping of the set of observations from the business point of view as best as possible. Discrepancies in grouping that appear in the results of the algorithms must be minimized. The solution to this problem is determined by the unification process.

In order to assess the results of clustering, it is often helpful to assign a category to collected observations. In the case of very large data sets, it is not possible to assign all observations by experts. Therefore, it has been proposed to enable the assignment of observation to the clusters through decision rules that define clusters selected by the expert, in the form:

$$X_i \in C_j \mid if\ [(w_{11} \cap w_{12} \cap \ldots \cap w_{1k}) \cup (w_{21} \cap w_{22} \cap \ldots \cap w_{2m})\ \ldots]$$

Where Xi is a given observation, C_j is a cluster in the conditional expression. Attributes used in conditional clauses indicate their importance and usefulness in the characteristics of clusters.

The decision rules are determined by the algorithm of inductive decision tree algorithm C4.5 [28]. These rules make in possible, on the one hand, to interpret the obtained clusters and, on the other hand, to symbolically determine the observations belonging to individual clusters. This solution enable finding of similar semantic clusters generated by different algorithms. The symbolic interpretation of clusters is complemented graphically, which facilitates a quick identification of similar clusters. It should be noted that these works generally require significant involvement of marketing analysts.

In general, in the recommendation systems, the manager is only interested in a few clusters describing similar clients, similar products, or similar transactions. Therefore, for the analyst the first task involves identifying clusters that are still subject to unification. Although the task can be performed algorithmically, our experience has shown that much better results are obtained through selection of clusters by the analyst. If the visual selection is difficult, finding for a cluster C_i, a counterpart among clusters $C_j \in C_k$ obtained from another algorithm, then the formula of similarity between clusters $S(C_i,C_j)$ can be applied:

$$S(C_i,C_j) = max\ (|C_i \cap C_j| / |C_i|).$$

In cases where the cluster's observations C_i are distributed into several clusters from C_k, the assignment should take into account the distribution of S values and the weights of related cluster similarities.

After selecting the clusters obtained from different algorithms, one can start unifying the results. There are many methods of unification [29]. The most commonly used methods are the following:

- Consensus methods [30],[31],[32],[33], which are used more in the first phase of unification to create initial clusterization than to unify the results
- Multi-criterial grouping methods [30],[31] are mainly used to harmonize the criteria of different algorithms,
- Clustering methods supported by domain knowledge [35],[36],.

The last group of unification methods was used in the Upsaily system. The domain knowledge of marketing has been used to direct the unification process of selected clusters. In the system, the earlier created decision rules were used to govern the process of unification, in particular, the conditional expressions of which are treated as grouping constraints. The idea of the proposed method consists in determining semantic relationships-constraints indicating observations that must be included in the cluster (called must-link), and those that

[2] The Davies-Bouldin index is computed according to the formula: $DB = 0.5n\Sigma\ max\ ((si + sj) / d(ci,cj)$ where n is the number of clusters, the cluster centroids, si and sj mean d distances between the elements of a given cluster and the centroid The algorithm that generates the smallest value of the DB indicator is considered the best according to the criterion of internal evaluation.

[3] The Dunn index is calculated according to the formula: $D = min\ (d(i,j) / max\ d'(k)$ where $d(i,j)$ means the distance between clusters i i j and $d'(k)$ the measure of distances within the cluster k. The Dunn index focuses on cluster density and distances between cluster. Preferred algorithms according to the Dunn index are those that achieve high index values.

[4] Cluster homogeneity index is computed according to the formula: $CH = 1/N\ \Sigma max\ |\ m \cup d\ |$ where M is the number of clusters created by the algorithm, D is the number of expert classes.

[5] The Jaccard index measures the similarity between two sets of observations according to the following formula: $WJ = TP / (TP + FP + FN)$, where TP means True Positive error, FP False Positive, FN False Negative. In the case of two identical sets of $WJ = 1$.

[6] The Rand measure is calculated according to the formula: $WR = (TP + TN) / (TP + FP + TN + FN)$. The Rand index, as well as the previous ones, is based on a comparison with the benchmark given by an expert. It informs about the similarity of the assessment of correct decisions between the results of the clustering algorithm and the benchmark.

should not be included in it (called cannot-link). In order to improve the quality of clusters, fuzzy logic is proposed in some works [37],[38],[39] or characteristics of clusters such as values of inter-cluster distances, density [40], [41].

Let us now follow the entire unification process step by step aiming to achieve consensus on the content of the final clusters without a significant loss of quality of the partitions. Let us assume that they were pre-designated as similar two clusters C_j i C_l, each generated by a different algorithm. As indicated, the interpretation of each cluster is given in the form of decision rules, namely:

C_j | *if* $[(w'_{11} \cap w'_{12} \cap ... \cap w'_{1k}) \cup (w'_{21} \cap w'_{22} \cap ... \cap w'_{2m})$ ...]

C_l | *if* $[(w''_{11} \cap w''_{12} \cap ... \cap w''_{1k}) \cup (w''_{21} \cap w''_{22} \cap ... \cap w''_{2m})$ ...]

The final cluster can be created by merging of conditions containing variables (attributes) indicated by the analyst based on domain knowledge. This operation can be called a subsumption according to which the more detailed condition are covered with a less detailed one. However, the resulting cluster may contain too many observations that are too far away from the class sought (as shown in Fig.4). In narrowing the cluster's space, the observations given earlier by the expert might help, defined as a must-link or cannot-link marked in Fig. 4 in green and red respectively.

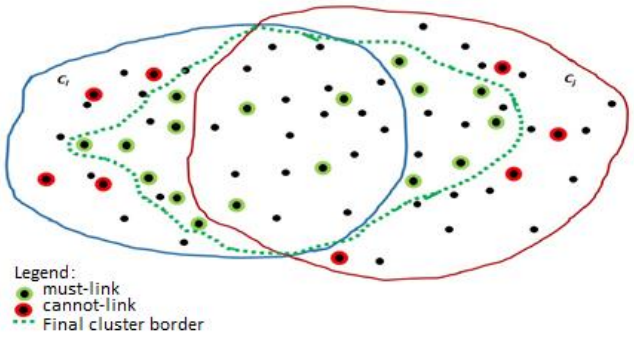


Fig. 4 Example of space of merged clusters.

The boundary of the final cluster (green dotted line) is determined between the sum of observations belonging to two clusters minus the surroundings of ε observations belonging to the *can not-link* relationship and the intersection of the observation plus the surroundings ε observations belonging to the *must-link* relationship. It can therefore be noted that the unified cluster includes observations lying in the space $|C_i \cup C_j|$ - ε *Xi/cannot-link* and $|C_i \cap C_j|$ + ε *Xi/must-link*. The radius of the surroundings ε can be determined based on ½ distance between the closest observations belonging to the *can not-link* and *must-link* relationships.

After the first unification of clusters, the process should be repeated for all similar clusters obtained from all algorithms. It should be noted that the order in which the clusters are selected influences the calculation time. We suggest choosing the most numerous clusters of interest in the first place. The next chapter will show examples of unification of the results of collective clustering.

Due to the thematic orientation of the conference and the restricted volume of the article in the next chapter, we will concentrate only on the business assessment of the results of clustering (domain knowledge). The RFM analysis will be used which is a traditional approach to analyze the customer behavior in the retail industry. Its acronym comes from the words "recency" (period from the last purchase), "frequency", and "monetary value". In this type of analysis, customers are divided into groups, based on information on time which has elapsed from last purchases, how often they make purchases, and how much money they spent (see [42]).

The following observations explain why RFM is interesting for retail companies:

- Customers who have recently made purchases are more likely to make a new purchase soon
- Customers who frequently make purchases are more likely to do more shopping
- Customers who spend a lot of money are more likely to spend more money

Each of these observations corresponds to one of the dimensions of RFM.

In the next section, the usefulness of this approach for assessing clustering algorithms is shown on the real marketing data.

IV. The Results of Experimental Research

In order to show the usefulness of the collective clustering method in specific business conditions, this chapter presents an experiment aimed at finding customer segments with similar behavior on the market. The clustering method should support a process of customer assignment to particular segment, assessment of proposed segments and interpretation of characteristics of these segments. The segmentation example was inspired by the RFM method. The customer is described by the following characteristics: frequency of their purchase (frequency dimension), the number of days which has passed since the last order (recency dimension) and the average order value (monetary value). We extended the customer description by information about the number of orders. Such dimension is essential in the case of an online store in order to determine the loyalty customer. The customers were divided into 6 segments. For each segment, we calculated its value (the sum of all customers' orders from a given segment). The number 6 was chosen arbitrarily. Marketing employees were able to prepare 6 different marketing communication policies addressed to individual customers. With more segments, it would be very difficult for the marketing analyst to interpret segments and subsequently develop a tailored communication policy for selected customers. A larger number of segments will be justified only if the automatic recommendation mechanism uses this segmentation.

The experiment was carried out using three clustering algorithms: bisecting k-means, Gaussian Mixture Model and

DBSCAN[7]. After each experiment, an expert evaluated the results of the segmentation. The analysis covered 56 237 customers who made at least 2 purchases in the online store.

When assessing segmentation, it is very helpful to visualize the data. Having 4 dimensions and ability to present it on surface (with only two dimensions). We used two methods for projecting the multidimensional space into a smaller number of dimensions. In order to prepare the visualization in the experiment, the four dimensions were reduced to two (X and Y), while the color means the segment number to which the given customer was assigned. One of those methods is The Principal Component Analysis[8] (PCA). PCA is a popular technique for reducing multidimensional space [43].

An example of RFM segmentation using the k-means algorithm and visualization using the PCA method is presented in Figure 5. One dot represents one real customer on the visualization (on left hand side of picture). After hovering over the selected dot, one can read the values describing the selected customer. This solution will help the marketing analyst to understand the prepared segments.

In the right part of the report there are funnel charts, presenting the average value of the given dimension attribute in individual segments; for example, average customer from segment 1 purchases with frequency of 14.22 days.

The column chart located in the bottom right corner of the report shows the sum of customers' orders in a given segments. It can therefore be observed that the highest revenue was generated by customers from segment no. 6, while the smallest in segment no. 5.

Another method of reducing dimensions that is useful for visualization is the Uniform Manifold Approximation and Projection (UMAP) [44]. It is a novel manifold learning technique for dimension reduction. UMAP seeks to provide results similar to t-SNE, which is the current state-of-the-art for dimension reduction for visualization, with superior run time performance. A theoretical framework of UMAP is based on Riemannian (a non-Euclidian) geometry and algebraic topology. In overview, UMAP uses local manifold approximations and patches together their local fuzzy simplicial set representations.

It is based on the approximation of the local manifold (local manifold approximations) and fuzzy simplicial sets. In contrast to a simple method such as PCA, where the projection is mainly based on two dimensions, the UMAP method takes all dimensions into account equally. An example of a visualization made using the UMAP method is presented in Figure 6.

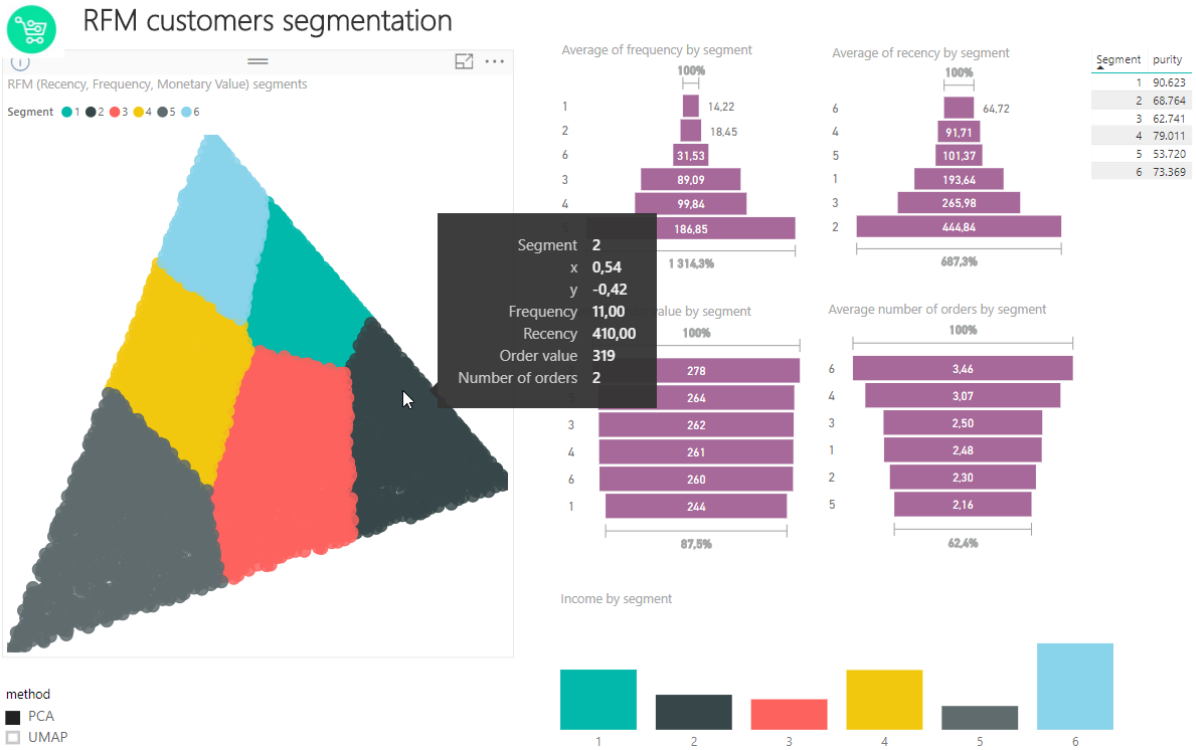


Fig. 5 Segmentation using the k-means algorithm and PCA visualization.

[7] The HDBSCAN algorithm was used, which is an extension of the DBSCAN algorithm. A library available on the GitHub platform was used for this purpose: https://hdbscan.readthedocs.io/en/latest/index.html

[8] The purpose of the PCA method, in brief, is to find a linear subspace (in our case 2-dimensional) in which the variance after projection remains the largest. The PCA method, however, is not to easily reject the dimensions with the lowest variance. It builds a new coordinates system in which the remaining values are the most diverse.

Visualization using two methods as well as presentation of the values of individual dimensions in clusters allow the analyst to better understand the individual customer segments and make an expert assessment of clustering.

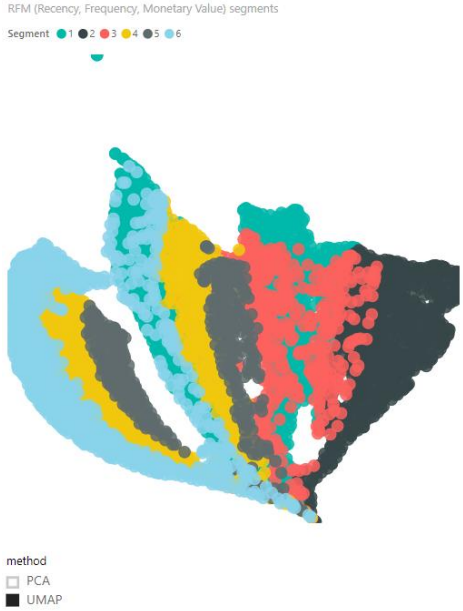


Fig. 6 Visualization of segmentation using the UMAP method.

Clustering using the k-Means algorithm based on the Euclidean distance between observations has many drawbacks. These include the fact that, when assigning customers to segments, the most varied dimensional values have the greatest impact (in our case, recency and frequency). The other dimensions impact less, and this can be observed in the low differentiation in the dimensions of average orders' values. In addition, it should be noticed that the boundaries between individual segments are not sharp. For example, segment 6, with the lowest average value, recency dimensions, includes both customers with a value of 0 and customers with a value of 147, these customers from the perspective of the RFM method, made their purchases relatively long time ago. The main advantage of this algorithm is the fact that the segments are relatively well balanced (their size is relatively similar). It makes those segments worth creating a dedicated marketing policy.

The next algorithm of clustering used in the experiment was the bisecting k-means. In the case of this algorithm, greater diversity was observed in individual dimensions than in the case of k-means. The clusters were again relatively balanced, however, the problem of the slight diversification of the 1 dimension remained, and in some segments there were clients located far away from the average value on a given scale.

Subsequent clustering was performed using the Gaussian Mixture Model algorithm. That method resulted with significant differences in the value of individual dimensions, due to which we can observe interesting cases of outliers (e.g., segment 1 includes customers with a very large number of orders and very high value of orders). Unfortunately, the size of such segments is relatively small (in this case 34 customers), which makes the legitimacy of building a special communication policy targeted to the customers from such a segment questionable. The same experiment was repeated for the DBSCAN algorithm. In case of this algorithm, the number of clusters was defined. Algorithm takes as parameter only the minimum size of the cluster. The disadvantage of this approach is the fact that a large part of the observations were not assigned to any cluster, and also that the majority of clusters are very small. The advantage is that the average values of the dimensions in the indicated segments are very diverse. The use of this algorithm to build communication policies is therefore debatable, but its advantage is the fact that clusters of relatively few but very similar observations are found, which can be used in the automatic recommendation mechanism.

For the marketing analyst, in order to perform the clustering using all the mentioned algorithms, they should observe the boundaries identified by algorithms on individual dimensions, and then those borders to build their own clusters, which will be referred to as according to their interpretation, e.g.

If average order value> 1000 zł
and number of orders > 10 and recency < 300
and frequency < 60
then segment=„active frequent valuable buyers"

If average order value< 200 zł
and number of orders < 3 and recency > 250
and frequency > 200
then segment=„occasional past cheap buyers"

In the platform, client filtering for clustering assignments can be done "manually" using the provided "sliders" presented in the upper right corner in Figure 7.

In the last phase of the experiment, a collective segmentation was proposed, taking into account the results of the three selected clustering algorithms. Because of similar results of the k-means and bisecting k-means algorithms, the k-means was not finally used in the experiment. We created collective segments basing on the results of 3 algorithms. The label of new clusters is constructed with the 3 numbers of clusters generated by the algorithms: bisecting k-means, GMM, and DBSCAN. For example, cluster 326 means that the customer has been assigned originally to clusters with numbers: 3 - bisecting k-means; 2 - GMM; 6 – DBSCAN.

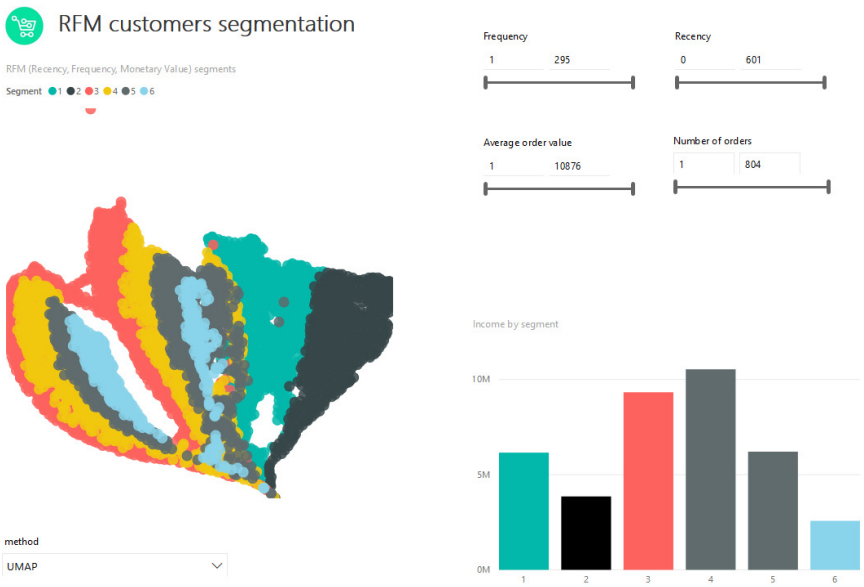


Figure 7 Manual segmentation. Source: Own elaboration in the Upsaily.

As a result, 52 segments were created (on 216 possible combinations), which is presented in Figure 8.

Such a large number of segments, of course, do not allow for an in-depth analysis of each of them and for "manual" preparation of marketing policies. However, these segments can be successfully used in the automated recommendation mechanism.

If the marketing analyst needs to analyze and interpret individual segments, in order to limit the number of clusters, similar segments may be merged. After the analyst decides on the maximum number of clusters or the minimum cluster size, then segments below the thresholds are included in the larger segment meeting the criterion of cardinality. Clusters' merge can be made with the lowest distance between them. The distance of clusters is not determined by the Euclidian measure, as for each of the aforementioned methods, cluster number is just an identifier without any meaning. Such identifiers do not determine similarity of clusters (e.g., cluster 1 doesn't have to be close to cluster 2). Taking the fact into account, distance in this case should be understood as the number of algorithms indicating a different cluster number, e.g., between clusters 525 and 520 the distance is 1 - which means that the clusters differ by the result of 1 method.

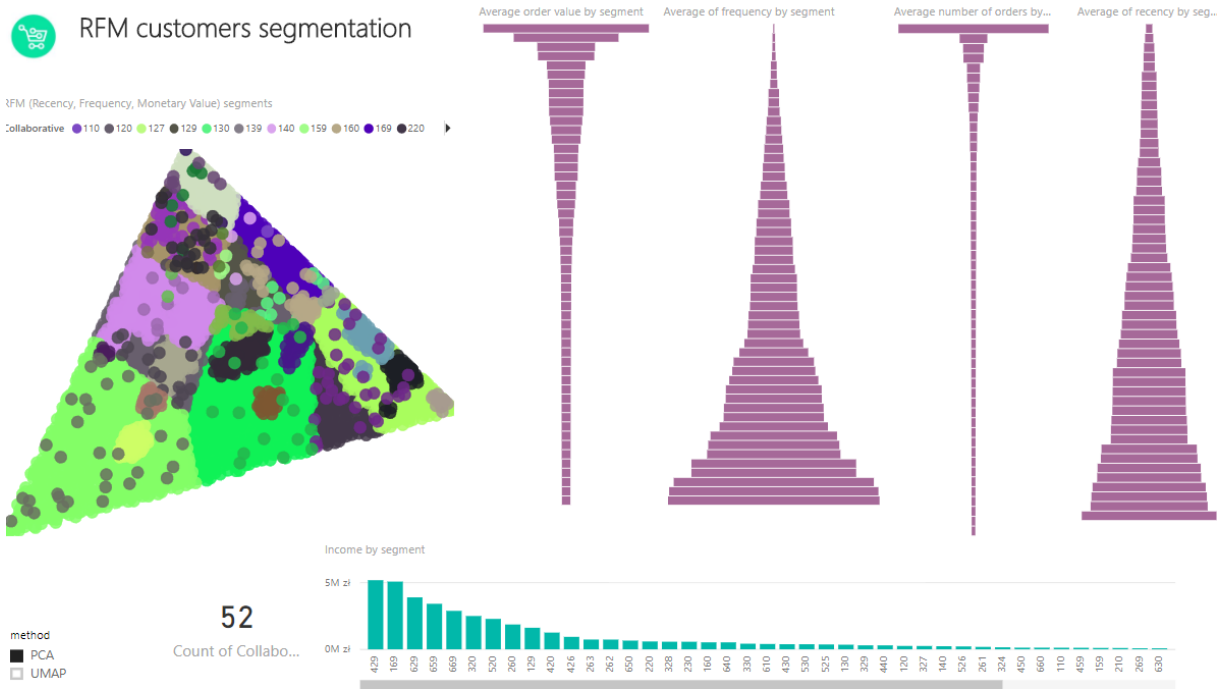


Figure 8 Visualization of 52 segments prepared using a collective approach.

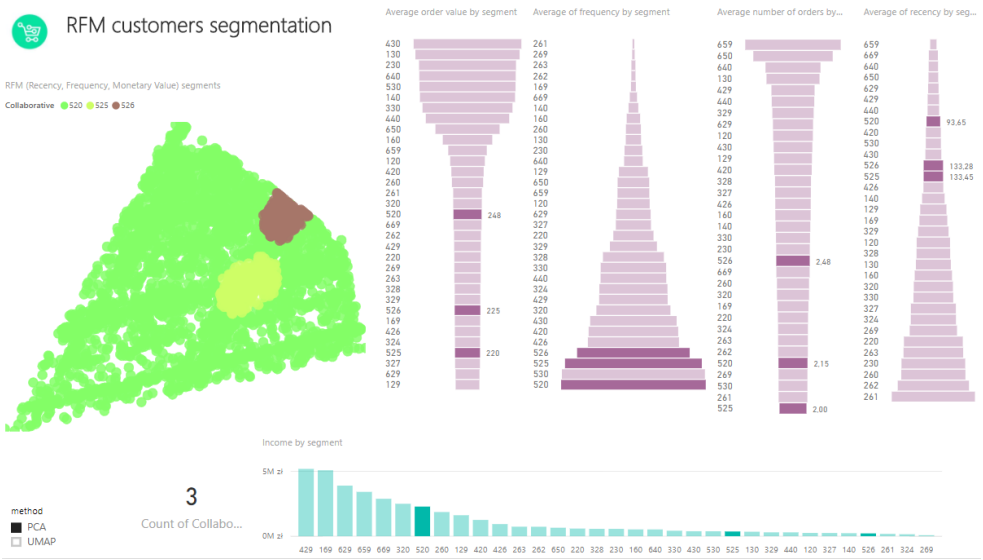


Figure 9 Visualization of the merge of 3 similar clusters prepared using a collective approach.

Between the clusters 320 and 525, the distance is 2. If clusters that should be merged are identified, a number of conflicts is encountered - clusters of the same distance. In this experiment, we will solve the conflict by selecting the highest cardinality cluster to which we attach a cluster that does not meet the criterion of cardinality.

Figure 9 illustrates an example of how to merge clusters 525 and 526 to cluster 520 (as the most numerous).

The k-means, bisecting k-means or GMM algorithms require a pre-determined number of clusters beforehand that we want to receive. The DBSCAN algorithm autonomously selects the number of clusters basing on other parameters, but in its case a large part of the observations are not included in any resultant cluster. We can state that DBSCAN cannot be used in case we would like to define marketing policies covering all clients, but is well suited for identifying smaller groups of observations that are very similar to each other.

Segmentation using a few selected algorithms gives more interesting results from the perspective of the marketing analyst than the segmentation using only one algorithm. First of all, the clusters obtained as a result of collective clustering have better and more useful marketing semantics. In addition, the analyst can decide on their own whether in using the described approach they focus on selecting the optimal number of large clusters, or analyze smaller clusters to identify hidden patterns of customer behavior.

V. Conclusions

The Upsaily system uses clustering as one of the methods for analyzing customer behavior in order to support generation of purchase recommendations. The RFM analysis answers the question when and what value products should be recommended to the customer. Other methods, such as association rules and sequential rules, additionally answer the question of what product / product category to offer to the customer. The Upsaily system also uses classification algorithms to refine the recommendations addressed to the customers.

Segmentation using one algorithm from the marketing analyst's point of view always has disadvantages such as small diversity of segments on particular dimensions or existence of segments with very low cardinality. In order to get rid of these indicated drawbacks and emphasize the advantages of each algorithm, we proposed a collective approach consisting in building a cluster by unification of the segmentation performed by the insights generated by all algorithms. Such segmentation gave us a result of more consistent segments with easier interpretation, however the final number of segments is definitely higher than when using each algorithm individually. Small segments can be useful in situations where we build an automatic mechanism of generating recommendations based on the client's assignment to the segment, where the large number of segments do constitute a problem. Segments consisting of a small number of customers are also useful in the task of identifying atypical clients as outliers.

If we want to provide a marketing analyst with a limited number of segments for the purposes of preparing a tailored marketing policy to each segment separately, then we suggest aggregating segments so that they meet the criterion of cardinality.

In future works, the authors will deal with the subject of collaborative clustering, automatic identification of the optimal number of segments and client clustering based on subsequent dimensions that also take their transactions and purchased products into account.

References

[1] Balabanovic, M., Shoham, Y., Content-based, collaborative recommendation. Com. of ACM 40(3), pp. 66–72, 1997.

[2] Goldberg, D., Nichols, D., Oki, B.M., Terry, D., Using collaborative filtering to weave an information tapestry. Com. of ACM 35(12), pp. 61–70, 1992.

[3] Resnick, P., Varian, H.R., Recommender systems. Com. of the ACM, 40(3), pp. 56–58, 1997.
[4] Konstan, J.A., Adomavicius, G., Toward identification and adoption of best practices in algorithmic recommender systems research. In: Proceedings of the international workshop on Reproducibility and replication in recommender systems evaluation, pp. 23–28, 2013.
[5] Beel, J., Towards effective research-paper recommender systems and user modeling based on mind maps. PhD Thesis. Otto-von-Guericke Universität Magdeburg, 2015.
[6] Jannach, D., Zanker, M., Ge, M., Gröning, M., Recommender systems in computer science and information systems–a landscape of research. In: Proc. of the 13th International conference, EC-Web, pp. 76–87, 2012.
[7] Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds): Recommender Systems Handbook, Springer, pp. 1–35., 2011.
[8] Jannach D., Zanker M., Felfernig A., Friedrich G., Recommender systems – an introduction, Cambridge University Press, 2010.
[9] Lu J., Wu, D., Mao M., Wang W., Zhang W.G.,, Recommender system application developments: a survey, Decision Support Systems, 74, pp. 12-32, 2015.
[10] Said, A., Tikk, D., Shi, Y., Larson, M., Stumpf, K., Cremonesi, P., Recommender systems evaluation: a 3d benchmark. In: ACM RecSys 2012 Workshop on Recommendation Utility Evaluation: Beyond RMSE, pp. 21–23, 2012.
[11] Acilar A. M., Arslan A., A collaborative filtering method based on Artificial Immune Network, Exp Syst Appl, 36 (4), pp. 8324-8332, 2009.
[12] Cornuejols A., Wemmert C., Gançarski P., and Bennani Y. Collaborative Clustering : Why, When, What and How. Information Fusion, 39, pp. 81–95, 2017.
[13] Kashef R., Kamel M.S., Cooperative clustering, Pattern Recognition 43, 6, pp. 2315–2329, 2010.
[14] Konstan J.A., Riedl J., Recommender systems: from algorithms to user experience User Model User-Adapt Interact, 22, pp. 101-123, 2012.
[15] Carmagnola, F., Cena, F., Gena, C., User model interoperability: a survey. User Model. User-Adapt. Interact. 21(3), pp.285–331, 2011.
[16] Burke, R., Hybrid recommender systems: survey and experiments. User Model. User-Adapt. Interact. 12(4), pp.331–370, 2002
[17] Lu J., Wu D., Mao M., Wang W., Zhang G., Recommender system application developments: a survey, Decision Support Systems, 74 , pp. 12-32, 2015.
[18] Kobiela E., Intelligent recommendation systems (pol. Inteligentne systemy rekomendacyjne), Network Magazyn, http://www.networkmagazyn.pl/inteligentne-systemy-rekomendacji, 2011
[19] Gemius 2017, The latest data on Polish e-commerce is now available (pol. Najnowsze dane o polskim e-commerce już dostępne), https://www.gemius.pl/wszystkie-artykuly-aktualnosci/najnowsze-dane-Polish-of-ecommerce-already-dostepne.html.
[20] Nazemoff V., Customer Intelligence. In: The Four Intelligences of the Business Mind. Apress, Berkeley, CA, 2014
[21] Chorianopoulos A., Effective CRM using predictive analytics. John Wiley & Sons, 2016.
[22] Gordon S. Linoff,M., Berry J.A., Data Mining Techniques: for Marketing, Sales, and Customer Relationship, Wiley 2011.
[23] Witten, I. H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[24] Jordan, M. I., MITCHELL, Tom M. Machine learning: Trends, perspectives, and prospects. Science, 349.6245, pp. 255-260, 2015.
[25] Pondel, M., Korczak, J., A view on the methodology of analysis and exploration of marketing data. In Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, pp. 1135-1143, 2017.
[26] Aggarval C. C., Reddy C.K, Data Clustering: Algorithms and Applications, Chapman & Hall / CRC 2013
[27] Gan G., Ma C., Wu J., Data Clustering: Theory, Algorithms, and Applications, SIAM Series, 2007.
[28] Quinlan J., Improved use of continuous attributes in {C4.5}. Journal of Artificial Intelligence Research, 4, pp.77–90, 1996.
[29] Wemmert C., Gancarski P., Korczak J., A collaborative approach to combine multiple learning methods. International Journal on Artificial Intelligence Tools (World Scientific), 9(1), pp.59–78, 2000.
[30] Strehl A., Ghosh J., Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 3, pp.583–617, 2002.
[31] Ayad H., Kamel M. S., Cumulative voting consensus method for partitions with variable number of clusters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(1), pp.160–173, 2008.
[32] Nguyen N., Caruana R., Consensus clusterings. In International Conference on Data Mining, IEEE Computer Society, pp. 607–612, 2007.
[33] Pedrycz W., Collaborative and knowledge-based fuzzy clustering. International Journal of Innovative, Computing, Information and Control, 1(3), pp.1–12, 2007.
[34] Faceli K., Ferreira de Carvalho A.C., Pereira de Souto M. G., Multiobjective clustering ensemble with prior knowledge. Volume 4643, Springer, pp. 34– 45, 2007.
[35] Law M. H., Topchy A., Jain A.K., Multiobjective data clustering. In IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pp. 424–430, 2004.
[36] Wagstaff K., Cardie C., Rogers S., Schroedl S., Constrained k-means clustering with background knowledge. In International Conference on Machine Learning, pp. 557–584, 2001.
[37] Belarte, B., Wemmert, C., Forestier, G., Grizonnet, M., Weber, C. Learning fuzzy rules to characterize objects of interest from remote sensing images. In Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE , pp. 2986-2989, 2006.
[38] Guo, H. X., Zhu, K. J., Gao, S. W., & Liu, T., An improved genetic k-means algorithm for optimal clustering. In Conference on Data Mining Workshops, 2006. ICDM Workshops. IEEE, pp. 793-797, 2006.
[39] Grira N., Crucianu M., Boujemaa N., Active semi-supervised fuzzy clustering. Pattern Recognition, 41(5), pp.1851–1861, 2008.
[40] Bilenko, M., Basu, S., & Mooney, R. J., Integrating constraints and metric learning in semi-supervised clustering. In Proceedings of the twenty-first international conference on Machine learning, ACM, p. 11,2004.
[41] Gancarski P., Cornueejols A., Wemmert C.,Bennani Y., Clustering collaboratif : Principes et mise en oeuvre, Proc. BDA'17, Nancy, 2017
[42] Linoff, G. S., Data analysis using SQL and Excel. John Wiley & Sons, 2015.
[43] Ghodsi, A., Dimensionality reduction a short tutorial, Department of Statistics and Actuarial Science, Univ. of Waterloo, 37, pp. 38, 2006.
[44] McLeland I., Healy J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426, 2018.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 811–815
DOI: 10.15439/2018F347
ISSN 2300-5963 ACSIS, Vol. 15

Utilizing online collaborative games to facilitate Agile Software Development

Adam Przybyłek, Wojciech Kowalski
Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics
Narutowicza 11/12, 80-233 Gdansk, Poland
Email: adam.przybylek@gmail.com, wojkow94@gmail.com

***Abstract*— Effective collaboration and interaction among the development team and between the team and the customer as well as proactive attitude in initiating and implementing improvements play vital roles in the success of agile projects. The challenge is how to address these social aspects since neither the Agile Manifesto nor the Scrum Guide specify techniques that aid the human side of software development. To fill this gap, we developed a web portal which provides 8 collaborative games to be used in agile software development. The feedback received from a Scrum team, who leveraged the games in an industrial project conducted in OKE Poland, indicates that our approach improves participants' communication, motivation, commitment, and creativity.**

Introduction

THE emergence of agile methods has triggered a growing awareness that social aspects play a key role in the success of software projects [1, 15, 27, 28]. Indeed, the Agile Manifesto [11] promotes principles and values such as "face-to-face conversation", close collaboration between developers and stakeholders, "motivated individuals", and regular retrospectives. In addition, in agile software development both developers and stakeholders are expected to be engaged – proactive and creative in identifying problems, envisioning future business practice and shaping solutions that exceed company's expectations [2, 3-5, 9, 14, 16, 18, 19, 23]. Unfortunately, neither the Agile Manifesto nor the Scrum Guide specify techniques to address the human side of software development. Responding to this challenge, in our previous studies [21, 23], we proposed to equip Scrum teams with a set of collaborative games.

Collaborative games refer to several structured techniques inspired by game play, but designed for the purpose of solving practical problems [23]. They involve strong visual or tactile elements that help the participants leverage multiple dimensions of communication, resulting in richer, deeper, and more meaningful exchanges of information [12, 23]. At the same time, they make use of the concepts of teamwork and collaboration, which lead to a variety of measurable societal outcomes.

Our previous studies [21, 23] revealed that playing collaborative games during Scrum meetings improves participants' communication, commitment, and creativity. In this study, we go one step further and make it easier for agile teams to adopt collaborative games. We developed a web portal (http://153.19.52.168) which provides online versions of 8 collaborative games. In these games, a team or a group of stakeholders participates in a collocated session and plays a game to discover requirements, prioritize requirements, or provide feedback related to the development process or the software system being implemented.

Related work

Although there have been hundreds of papers related to the application of serious games for teaching software engineering and software project management [10, 17, 24, 25], the interest in using collaborative serious games has not received so much attention yet. An important cornerstone for this research area were innovation games introduced by Hohmann [12] as market and product research techniques and later adopted by Ghanbari et al. [7] and Przybyłek & Zakrzewski [23] to support distributed requirements engineering and agile requirements engineering respectively. Likewise, Gelperin [6] defined six collaborative games that support requirements understanding. In turn, Trujillo et al. [26] proposed a game-based workshop as an alternative for the software project's Inception phase. Being inspired by their work, Przybyłek & Olszewski [22] proposed an extension to Open Kanban, which contains 12 collaborative games that help inexperienced teams better understand the principles of Kanban. Recreantly, Przybyłek & Kotecka [21] and Mesquida et al. [16] adopted collaborative games to support Agile Retrospectives.

Selection of collaborative games

The first decision to be made was the selection of collaborative games to be implemented. Our main objective when developing the portal was to offer at least one game for each Scrum meeting except the Daily Scrum, which is too short and too well-structured to take advantage of collaborative games. Since there are several games that may be utilized during each Scrum meeting, we chose those that had received the most positive feedback in our previous studies and were easy to implement. Ultimately, our portal provides 8 collaborative games. The assignment of the

games to the Scrum meeting in which the game is applicable is as follows:

- **Product Planning**: Whole Product, AVAX Storming, SWOT Analysis;
- **Sprint Planning**: Buy-a-Feature, How-Now-Wow Matrix;
- **Sprint Review**: Speedboat;
- **Sprint Retrospective**: Mood++, 4Ls.

A. Whole Product

The game helps the team discover new features that can make the product distinct and prioritize the product backlog [12]. The game board comprises four stairs levels that represent four kinds of features:

- Generic – the fundamental features that define the software system;
- Expected – the features that the customer considers absolutely essential;
- Augmented – the features that the customer wishes to have implemented;
- Potential – the features that go beyond the customer expectations.

B. AVAX Storming

The game aims at identifying "needed" and "desired" features of the system to be developed. The final result should be a mind map demonstrating the size of the project [25]. Unfortunately, due to implementation difficulties our version of this game only allows for categorizing features without the possibility of creating a mind map.

C. SWOT Analysis

The game is a strategic planning technique used to help an organization identify the Strengths, Weaknesses, Opportunities, and Threats related to a project. Strengths and weaknesses are internal to the business, while opportunities and threats arise externally. This game can be also employed to discover requirements for a software system [13].

D. Buy-a-Feature

The game is a way of choosing the right set of features to be developed in the next Sprint. In this game, customer representatives collaborate to purchase their most desired features using game money (Fig. 1). Strictly speaking, they jointly prioritize their desires as a group [12]. Each features has a price related to its development cost. Some features may be priced so high that no single player can buy them individually. This motivates negotiations among players because they have to pool their money to buy the feature.

E. How-Now-Wow Matrix

The game helps stakeholders identify features that make the software system unique and distinguish it from its competitors. It should be played in later sprints after the core features are implemented. The game board is a 2×2 matrix with "originality" on the x-axis and "feasibility" on the y-axis as shown in Fig. 2 [23].

Figure 1. Buy-a-Feature

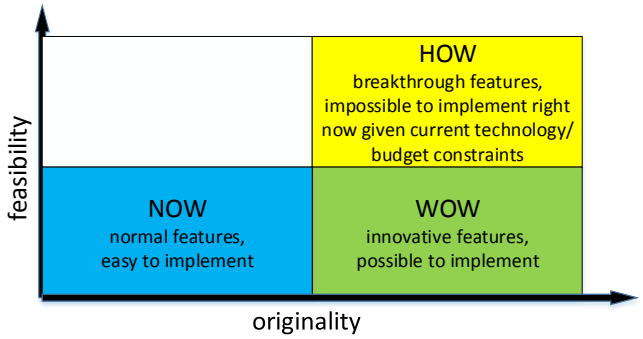


Figure 2. How-Now-Wow Matrix [23]

F. Speedboat

The game explicitly asks customer representatives to say what they do not like about the product. Nonetheless, it lets the facilitator stay in control of how the complaints are stated. The game starts by drawing a speedboat. The speedboat represents the software system. Everyone wants the speedboat to move fast. Unfortunately, the speedboat has a few anchors holding it back [8]. Customer representatives write what they do not like on sticky notes and place them under the speedboat as anchors. The lower an anchor is placed, the more significant the issue is. Customer representatives may also add engines to the speedboat. The engines represent features that can "overpower" the anchors and enable the speedboat to move faster [12].

G. Mood++

The game helps release a heavy emotional steam and gather data about feelings during the Sprint. The game board comprises five areas [21]:

- Mad – frustrations, issues that annoyed the team and/or wasted a lot of time;
- Sad – disappointments, issues that did not work out as well as was hoped;
- Glad – pleasures, issues that made the team happy;
- Flowers – appreciation to colleagues who did something magnificent for the team or a particular team member;

- Ideas – suggestions how to improve the teamwork or the process.

H. 4Ls

The game handles both the positive and negative aspects of the Sprint, but also brings forth the continuous improvement [21]. The game board contains four columns:

- Liked – what did the team really appreciate about the Sprint?
- Learned – what new things did the team learn during the Sprint?
- Lacked – what things could the team have done better in the Sprint?
- Longed For – what things did the team wish for but were not present during the Sprint?

IV. EVALUATION

The evaluation was performed during the second half of 2017 and the first half of 2018 in OKE Poland (oke.pl). OKE Poland is a software development company that provides innovative IT solutions for its partners in Europe and the United States. 6 out of 8 games hosted by our platform were utilized by a Scrum team when they were developing software for a Dutch company. Since the customer was located in a different country, its availability throughout the project was limited. Accordingly, we were not able to evaluate Buy-a-Feature and AVAX Storming, which require the participation of numerous customer representatives. The team consisted of 6 developers, who had experience in all evaluated games due to their participation in our previous research. The second author of the paper facilitated all game sessions. After each session, a questionnaire was issued to collect feedback on game-playing experiences (Fig. 3-10). The responses were on a Likert scale of 5 points. Overall, all games were evaluated positively. The detail results are presented in the succeeding subsection. As the next step, the results were discussed in a focus group. The details about the meeting and its findings are given in Section IV.B.

A. Questionnaire

Figures 3-10 aggregate the number of responses for each Likert level and game for a given question. Although some games required participation of the customer representatives, who varied slightly between the sessions, each game was evaluated by the development team only to ensure the comparability of the results between the games.

The great majority of participants state that each evaluated game produces better results than the standard approach (Fig. 3) and is easy to understand and play (Fig. 4). However, as for Whole Product, Mood++, and 4L's the opinions on whether these games should be permanently adopted by the team, are divided almost equally between supporters, opponents and undecided (Fig. 5). The opponents complain that playing a retrospective game is much more time-consuming than running a traditional retrospective. In turn, the final result of Whole Product was unreadable, because most of the identified features fell into the first category.

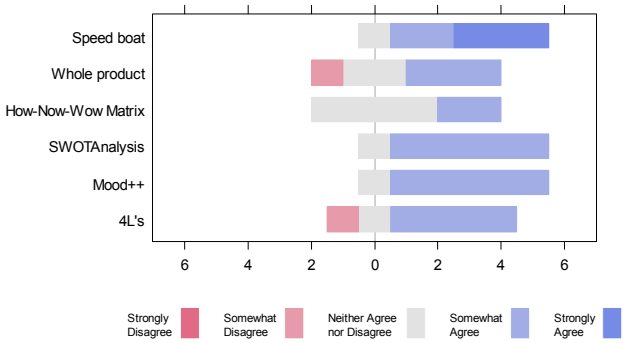


Figure 3. The game produces better results than the standard approach

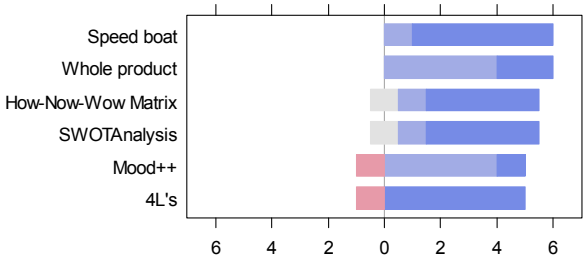


Figure 4. The game is easy to understand and play

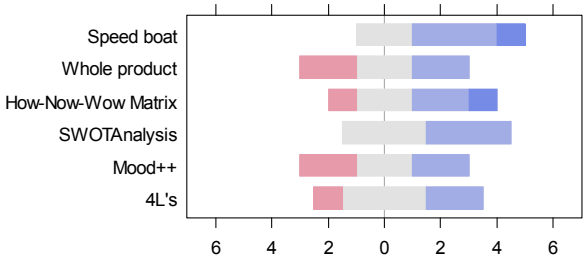


Figure 5. The game should be permanently adopted by the team

The great majority also consider that the games foster participants' creativity (Fig. 6) and improve communication among participants (Fig. 7). Especially, communication between the team and its customer has been improved.

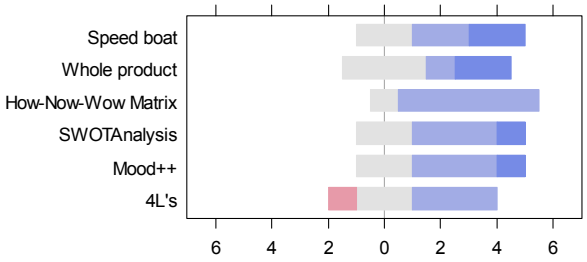


Figure 6. The game fosters participants' creativity

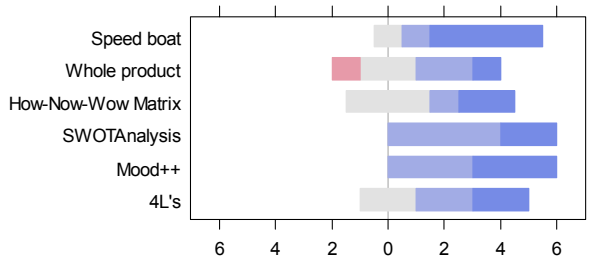


Figure 7. The game improves communication among participants

All games except 4L's are claimed to foster participants' motivation and involvement with only single opposite voices (Fig. 8). As for 4L's, the opinions are divided equally between supporters, opponents and undecided.

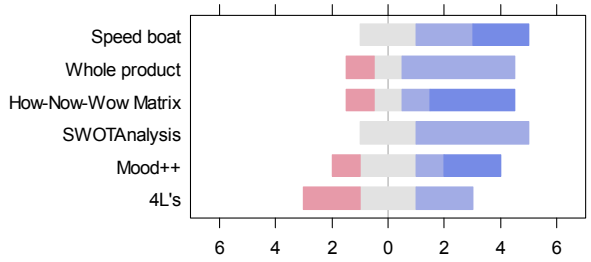


Figure 8. The game fosters participants' motivation and involvement

When it comes to the impact of the games on the willingness to attend the meeting, the responses are dominated by those who purport that it is difficult to unequivocally determine the impact (Fig. 9). Although these respondents see the value in the games, they are afraid that playing a game at each Sprint may be tiring.

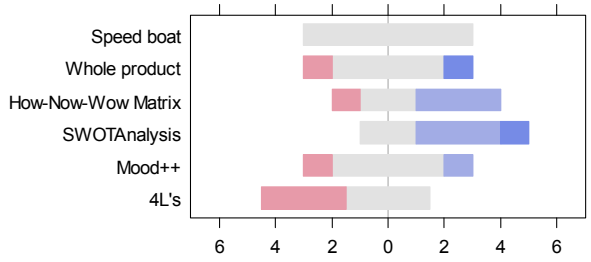


Figure 9. The game makes participants more willing to attend the meeting

The key question for this study is whether the online versions outperform the non-digital ones (Fig. 10). Although the online versions do not perform worse, only the online version of Speedboat, Whole Product, and Mood++ perform significantly better than their non-digital counterparts. As for Whole Product, its digital game board is considered more apparent than the original one (we changed the original game board [12] due to implementation difficulties).

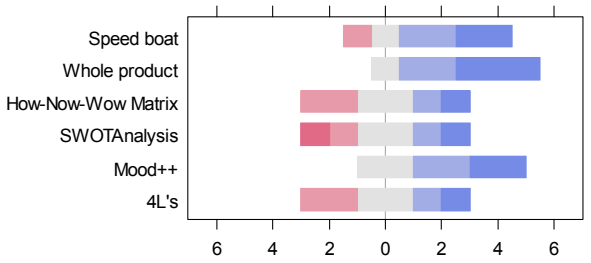


Figure 10. The online version of the game outperforms the non-digital version

B. *Focus group*

We conducted a focus group with the team to analyze and discuss the results presented in the previous subsection. The discussion was structured around four questions:

- What are your comments on the results?
- What are advantages and disadvantages of the online collaborative games over their non-digital counterparts?
- Why did some games perform better than the others?
- Is there something that can be improved in the provided games?

At the end of the day, the findings were as follow. The non-digital versions impose overhead to draw the initial template that participants have to fill in. Accordingly, the more complex the game board, the greater the gain from an online version. One debater noted that our portal cannot be used by distributed participants and suggested that it could be improved by adding a chat facility.

Making corrections (e.g. moving cards/notes between different areas or updating the content) is easier and more flexible in the online versions. Thereby, outcomes generated from the online versions are more readable. Moreover, the non-digital versions require physical game accessories to play a game. Even though most of the games use only simple accessories such as posters, colorful sticky notes and coloring markers, the team encountered situations where there were not enough colors of sticky notes. As for the online versions, there are no problems with missing artifacts.

Joining an existing game session is cumbersome. It would be better if there is a drop-down list of all available game sessions that users can join. Furthermore, the rules of a game should be accessible when the game is running.

V. Conclusions

In this study, we developed a web portal which provides 8 collaborative games to be used in agile software development. The received feedback not only confirms our previous findings that playing collaborative games during Scrum meetings improves participants' communication, commitment, and creativity, but it also suggests that our online collaborative games can substitute their non-digital counterparts. Nevertheless, the intention of this work is not to convince anyone to switch from the non-digital versions into the online versions, but to simplify the adoption of collaborative

games into daily practice by those who have never used them. Our portal lets agile teams try collaborative games without any investments in physical game artifacts. However, we still believe that playing collaborative games in their non-digital form creates a type of glue that bonds participants together and made them more comfortable to participate in the discussion. We hope that our research will inspire practitioners to utilize collaborative games to address the social aspects of software development.

As future work, the provided games need to be evaluated in other settings and contexts. We also hope that new games will be added in our portal in the future, since its source code is publicly available and we invite the community to contribute. Moreover, we want to study the effect of collaborative games on social aspects of software development in a controlled experiment with settings similar to [20].

ACKNOWLEDGMENTS

The authors would like to thank other programmers who developed the web portal, i.e.: Alicja Białous, Monika Czwartosz, Bartosz Stefański, Bartosz Zaborowski, Tomasz Piwowarski, Mateusz Górski.

REFERENCES

[1] Amin, A., Basri, S., Hassan, M.F., Rehman, M.: Software engineering occupational stress and knowledge sharing in the context of Global Software Development. In: National Postgraduate Conference, Kuala Lumpur, Malaysia, 2011

[2] Boehm, B., Turner, R.: Balancing Agility and Discipline: A Guide for the Perplexed, Addison-Wesley, Boston, MA, 2004

[3] Cockburn, A., Highsmith, J.: Agile software development, the people factor. In: IEEE Computer, vol. 34(11), Nov 2001, doi: 10.1109/2.963450

[4] Crawford, B., León de la Barra, C., Soto, R., Monfroy, E.: Agile software engineering as creative work. In: 5th International Workshop on Co-operative and Human Aspects of Software Engineering, Zürich, Switzerland, 2012

[5] Essebaa, I., Chantit, S.: Model Driven Architecture and Agile Methodologies: Reflexion and discussion of their combination. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18), Poznan, Poland, 2018

[6] Gelperin, D.: Increase Requirements Understanding by Playing Cooperative Games. In: INCOSE Inter. Symp., Denver, CO, 2011

[7] Ghanbari, H., Similä, J., Markkula, J.: Utilizing online serious games to facilitate distributed requirements elicitation. In: Journal of Systems and Softwar, vol. 109 (November 2015), pp. 32–49

[8] Gonçalves, L., Linders, B.: Getting Value out of Agile Retrospectives: A Toolbox of Retrospective Exercises. Leanpub, 2014

[9] Hanslo, R., Mnkandla, E.: Scrum Adoption Challenges Detection Model: SACDM. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18), Poznan, Poland, 2018

[10] Hernández, J.A.C., Duarte, M.P., Beardo, J.M.D.: Skill assessment in learning experiences based on serious games: A Systematic Mapping Study. In: Computers & Education, vol. 113, 2017, pp. 42-60, doi: 10.1016/j.compedu.2017.05.008

[11] Highsmith, J., Fowler, M.: The agile manifesto. In: Softw. Dev. Mag. 9, pp. 29–30, 2001

[12] Hohmann, L.: Innovation Games: Creating Breakthrough Products Through Collaborative Play. Addison-Wesley Professional, 2006

[13] Jarzębowicz, A., Połocka, K.: Selecting Requirements Documentation Techniques for Software Projects: a Survey Study. In: 1st International Conference on Lean and Agile Software Development, pp. 1189–1198, 2017, doi: 10.15439/2017F387

[14] Jarzębowicz, A., Ślesiński, W.: Assessing Effectiveness of Recommendations to Requirements-Related Problems through Interviews with Experts. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18), Poznan, Poland, 2018

[15] John, M., Maurer, F., Tessem, B.: Human and social factors of software engineering: workshop summary. In: SIGSOFT Softw. Eng. Notes, vol. 30(4), pp. 1–6, July 2005

[16] Mesquida, A.L., Karać, J., Jovanović, M., Mas, A.: A Game Toolbox for Process Improvement in Agile Teams. In: 24th European System, Software & Service Process Improvement & Innovation, Czech Republic, 2017

[17] Miler, J., Landowska, A.: Designing effective educational games - a case study of a project management game. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS'16), Gdansk, Poland, 2016, doi: 10.15439/2016F434

[18] Nguyen, L., Cybulski, J.: Into the future: inspiring and stimulating users' creativity. In: 12th Pacific Asia Conference on Information Systems, Suzhou, China, 2008

[19] Ossowska, K., Szewc, L., Weichbroth, P., Garnik, I., Sikorski, M.: Exploring an Ontological Approach for User Requirements Elicitation in the Design of Online Virtual Agents. In: 9th EuroSymposium on Systems Analysis and Design, Gdansk, Poland, 2016

[20] Przybyłek, A.: An empirical study on the impact of AspectJ on software evolvability. In: Empirical Software Engineering, vol. 23(4), pp. 2018 – 2050, August 2018, https://doi.org/10.1007/s10664-017-9580-7

[21] Przybyłek, A., Kotecka, D.: Making agile retrospectives more awesome. In: 2017 Federated Conference on Computer Science and Information Systems (FedCSIS'17), Prague, Czech Republic, 2017, doi: 10.15439/2017F423

[22] Przybyłek, A., Olszewski, M.: Adopting collaborative games into Open Kanban. In: 2016 Federated Conference on Computer Science and Information Systems (FedCSIS'16), Gdansk, Poland, 2016, doi: 10.15439/2016F509

[23] Przybyłek, A., Zakrzewski, M.: Adopting Collaborative Games into Agile Requirements Engineering. In: 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'18), Funchal, Madeira, Portugal, 2018

[24] Souza, M.R.A., Veado, L., Moreira, R.T., Figueiredo, E., Costa, H.: A Systematic Mapping Study on Game-related Methods for Software Engineering Education. In: Information and Software Technology, vol. 95, pp. 201-218, 2018, doi: 10.1016/j.infsof.2017.09.014

[25] Trujillo, M.M., García-Mireles, G.A., Maslova, P.: What Can Go Wrong in a Software Project? Have Fun Solving It. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18), Poznan, Poland, 2018

[26] Trujillo, M.M., Oktaba, H., González, J.C.: Improving Software Projects Inception Phase Using Games: ActiveAction Workshop. In: 9th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'14), Lisbon, Portugal, 2014

[27] Wrobel, M.R.: Applicability of emotion recognition and induction methods to study the behavior of programmers. In: Applied Sciences, vol. 8(3), p. 323, 2018, doi: 10.3390/app8030323

[28] Wrobel, M.R., Zielke, A.W.: MaliciousIDE – software development environment that evokes emotions. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18), Poznan, Poland, 2018

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 817–821
DOI: 10.15439/2018F188
ISSN 2300-5963 ACSIS, Vol. 15

Towards a Framework for Semi-Automated Annotation of Human Order Picking Activities Using Motion Capturing

Christopher Reining*, Fernando Moya Rueda†, Michael ten Hompel*, Gernot A. Fink†
*Chair of Materials Handling and Warehousing, †Pattern Recognition in Embedded Systems Group
TU Dortmund, Dortmund, Germany
{christopher.reining, michael.ten.hompel, fernando.moya, gernot.fink}@tu-dortmund.de

***Abstract*—Data creation for Human Activity Recognition (HAR) requires an immense human effort and contextual knowledge for manual annotation. This paper proposes a framework for semi-automated annotation of sequential data in the order picking process using a motion capturing system. Additionally, it introduces proper annotation labels by defining process steps, human activities and simple human movements in order picking scenarios. An attribute representation based on simple human movements meets the challenges set by the versatility of activities in warehousing.**

I. Introduction

ORDER picking is the process of pulling items from a warehouse to satisfy specific customer orders. This basic warehousing process makes up more than half of the total operating expenses [1, p.1-30]. Sub-processes may be partially automatized in high-wage countries. Nevertheless, manual order picking systems remain dominant in practice [2]. To evaluate order picking systems, manual processes need to be quantitatively determinable [3]. Manual assessment of the order picking efficiency is unfeasible as trained specialists would be required to manually gather the necessary information in a highly versatile environment. Due to advancements in sensor technology and data processing, IT-supported approaches of Human Activity Recognition (HAR) gain significance.

HAR is a classification task where time-series segments are assigned to a specific activity class [4], [5], [6]. The authors in [5] provided the first approach of HAR in the order picking process. They recorded multichannel time-series from Inertial Measurement Units (IMUs). IMUs were attached to both arms and the torso of three workers in two warehouses. IMUs provide measurements of three different sensors: accelerometers, gyroscopes and magnetometers for three axes (x, y, z). The authors followed a standard pipeline in pattern recognition; that is, segmenting sequences, extracting handcrafted features, and training a classifier. They used a sliding window approach for segmenting time-series segments. For each of these segments, statistical features were computed and processed by three classifiers. Recently, deep convolutional neural networks (CNN) and recurrent neural networks (RNNs) were successfully used for recognizing human activities [6], [7], [8], [9]. A combination of convolutional layers and recurrent units is proposed in [7] for recognizing activities of daily life. In [8], different deep architectures were deployed to recognize human locomotion activities. In particular, they used a CNN, a long-short term memory (LSTM) network, and a bi-directional LSTM network. The authors in [6] proposed a CNN for solving HAR in the order picking process. In contrast to previous architectures, this CNN contains parallel branches. Each of these branches is composed of two or three convolutional layers and max-pooling operations processing segments per IMU. This architecture, called IMU-CNN, showed the state-of-the-art performance in HAR.

The success of deep architectures in different tasks heavily depends on the amount of data. Nowadays, large collections of data are available for tasks such as image classification, image segmentation and face recognition. However, this is not the case for HAR, which datasets are rather small and scarce. Providing data collections involves recording high quality raw data along with their respective class annotations. Data should be large, variate and correctly labeled. This process in HAR is more challenging in comparison with other tasks. For image classification datasets, label annotations can be carried out using a combination of unsupervised clustering and manual work [10]. However, HAR is diverse involving different type of data sources, e.g. from videos, or multichannel time-series from on-body sensors. HAR faces challenges with regards to environment settings, number of participants and number of sensors [4]. Furthermore, due to the large intra- and inter-class variability of the human movements, a large number of experiments must be carried out, which draw motion repetitions from the same or different persons [7]. These circumstances increase the data collection and annotation efforts. Obtaining and annotating data from videos is computational expensive, and, in the case of multichannel time-series signals, signals are visually hard to interpret. In both cases, annotations are carried out manually, involving the synchronization of the time-series with videos, observing the actions and labeling the sequences. This procedure takes enormous time. For example, annotations of time-series in the order picking dataset in [11] demanded $26min$ in average per minute of annotated data. In addition, annotations are inconsistent among different

The work on this publication has been supported by Deutsche Forschungsgemeinschaft (DFG) in the context of the research project "Adaptive, ortsabhängige Aktivitätserkennung und Bewegungsklassifikation zur Analyse des manuellen Kommissionierprozesses"

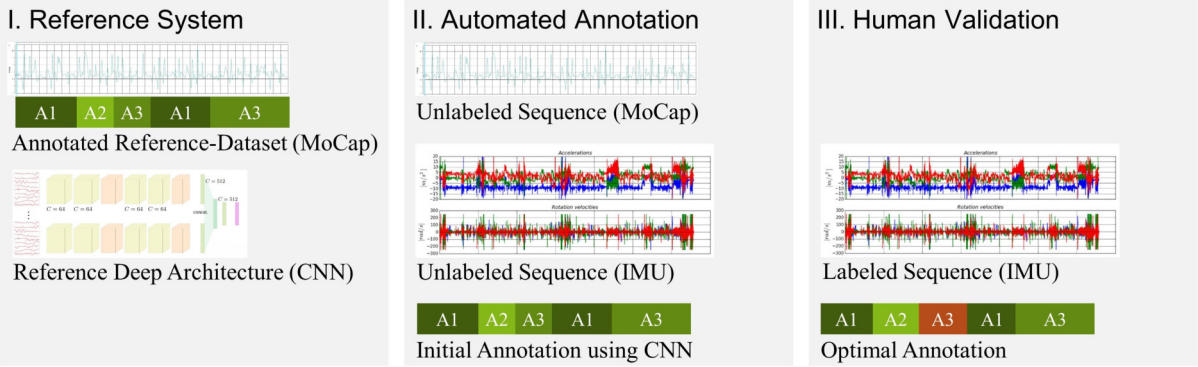


Fig. 1: Framework for semi-automated Annotation

annotators. Repetitions in the annotation process enhance the data quality [11], but escalate the data collection effort.

Apart from the annotation effort, the definition of activities is of high interest. In order picking scenarios, coarse activities like walking, picking and searching are often used [6], [11]. However, these scenarios are highly versatile involving a variety of activities. A possible way out is the definition of more finely subdivided activities. This implies more effort for creation and annotation of datasets. Following [9], activities could be represented by a set of attributes. Attributes are high level semantic descriptions of activities [12], [13]. These attributes are shared among all of the activities. For example, attributes like moving or not moving a foot and the velocity could define walking, running, and standing. Using an attribute based representation, problems like imbalanced data and overfitting are reduced. Sequential data from the most frequent activities could be used for learning attributes that are shared with less frequent activities, as simple human movements are shared among activities. In general, attribute annotations in the context of multichannel time-series HAR are not available. The annotations are related with specific coarse activities, for example standing or walking. However, there are no annotations of attributes describing those coarse activities. In [9], attribute representations for HAR are learned using an evolutionary algorithm, starting from a random combination of attributes. The learned attribute representations are suitable for solving HAR as classification task. However, their semantic interpretation is missing and therefore not understandable by humans.

II. Method

Datasets consisting of multichannel time-series from on-body sensors are of special interest in order picking. Usually, multiple sensors, e.g. IMUs, are worn by a worker gathering recordings in a simple and non-invasive manner. Besides, these sensors are impersonal, i.e. recordings do not portray the identity of the person. In comparison with HAR using videos, they do not suffer from occlusion, as the person's visibility changes along videos. In addition, IMUs are rather economic. Nevertheless, datasets from these devices are hard to annotate manually. As they are difficult to interpret by a human, additional video material is necessary to visualize the respective activity. This paper presents a framework, see Figure 1, to annotate multichannel time-series from on-body sensors using a deep learning model that is trained on highly accurate data. This framework is divided in three parts. First, sequential high quality data are created and annotated from a controlled environment as a reference dataset. Humans are recorded following activities that are commonly seen in order picking scenarios. Proper annotation labels are defined and, in addition, an attribute representation for human activities is introduced. This attribute representation is based on basic human activities and warehousing components. A deep model for solving HAR is learned on the reference dataset. Second, using this model, sequential data from an uncontrolled environment are initially labeled. This initial label includes the computation of uncertainty for the initial predictions. Third, uncertain predictions are revised by human work for final labeling.

A. Controlled Environment

On the one hand, naturalistic, real-life data are desired. On the other hand, data is prune to be disturbed in uncontrolled environments [14]. The primary reason to use a controlled environment set-up is the high accuracy of the available sensors. Interfering signals can be averted, and recording sessions can be conducted and repeated with different settings. The Motion Capturing (MoCap) that has been used for this paper is based on photogrammetry methods for measuring object positions on $2D$ and $3D$ spaces using a string of cameras. As an installation of the motion capturing system in a real warehouse is not practicable, it is located at the "InnovationLab Hybrid Services in Logistics" of the chair of materials handling and warehousing at the TU Dortmund University. The MoCap system consists of 38 cameras that cover a space of approximately $22m \times 10m \times 6m$. It uses passive markers to track rigid and flexible objects, such as drones, robots or humans in real time [15]. The passive markers reflect incoming infrared signals to the cameras, and their $3D$ positions are determined via triangulation.

The purpose of the MoCap system is to construct and record skeleton data from workers performing activities in an

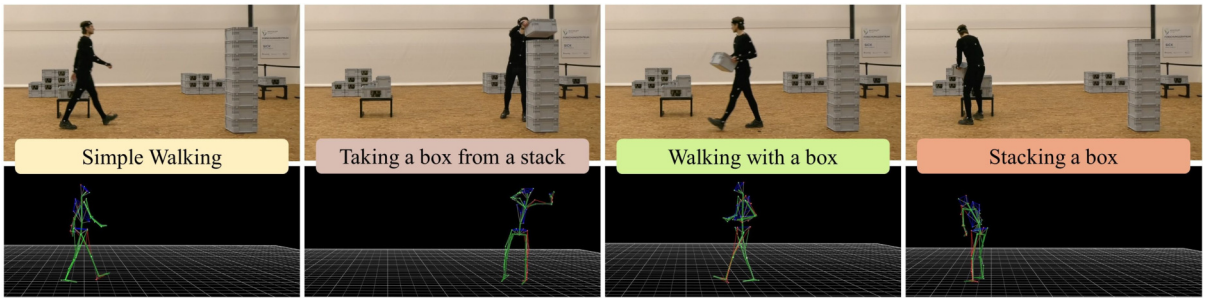


Fig. 2: Exemplary Activities and Motion Capture Data Skeleton

order picking scenario, as shown in Figure 2. Workers wear a specific suit with a set of passive markers. The MoCap computes the global $3D$ positions and it constructs a human skeleton. The MoCap System provides global poses from different parts of the human body, e.g. head, torso, arms and feet. A pose is a combination of position and angular values in $[X, Y, Z]$ of a certain reference system.

B. Annotation of Order Picking Activities

From a macro-level perspective, the human activity in order picking systems can be segregated into basic activities such as locomotion, retrieving and confirming [1, p.1-30]. An obvious approach of HAR would be to interpret each activity as a class. However, this approach is incapable to deal with the versatility of actions in real-world systems. Members of the same class differ significantly in terms of motions and tasks that are executed by the pickers [3]. For example, a warehouse employee can simply walk or walk while carrying a box. A single class cannot account for such distinctions. There is a wide variety of components that influence the human activity, ranging from the type of storage and collecting unit to the information technology [1, p.1-30], [16], [17]. These components and their combinations define coarse order picking process steps, e.g. *putting a box from a shelf onto a cart*. However, process steps can be composed of fine human activities such as *taking a box from a shelf* and *putting a box onto a cart*. Thus, each relevant process step needs to be defined with regards to human activities. This approach offers a high degree of flexibility. On the one hand, the definition of each human activity is fixed so that patterns in the sensor can be recognized and the obtained data is reusable. This is feasible as the definition of human activity is supposed to hold global validity irrespective of a specific context and environment. On the other hand, the definition of process steps is not fixed. Depending on the user's requirements, process steps can be defined very specifically or in more general terms. In addition, following [9], human activities are represented by a set of attributes that describe them semantically. These attributes are simple human movements, for example moving an arm or a foot. As shown in [9], attribute representations boost HAR tasks using deep architectures.

The proposal is to annotate time-series with a respective activity and a set of attributes, see Figure 3. The definition of both the activities and the attributes must be created a priori by a warehousing specialist to ensure that they are semantically understandable. The attributes are the output of the CNN that operates on the sensor data. The combination of attributes implies a specific activity. The activity sequence is then comprehended as a process step of order picking.

C. Creation of Reference Dataset

A reference dataset for order picking scenarios using the MoCap system, see subsection II-A, is created. The closeness to reality within the controlled laboratory environment was ensured by using the same kind of equipment, such as boxes or racks, that are used in real warehouses.

For this reference dataset, eight activities have been recorded: *Standing (none)*, *Walking (none)*, *Standing (box)*, *Walking (box)*, *Reaching forward (none)*, *Lifting (box)*, *Putting down (box)*, *Straighten up (none)*. Here, the words *box* and *none* express whether a worker walks with or without a box. Thus, the sequence of *reaching forward (none)* and *lifting (box)* implies the process step *picking up a box*. The box was a standard small load carrier with the dimensions L 600 mm x W 400 mm x H 220 mm and a gross weight of 4 kg.

The sample recording for this paper was conducted with eight participants of which four have been female and four male. Their height ranged from 161 to 192 cm and the average age was 25. Five participants have been right-handed and three participants have been left-handed. Previous research suggests that the handedness and gender have an impact on the motion [18]. The amount eight participants is equivalent to state-of-the-art approaches [19].

The activities were not recorded in a sequence and subsequently segregated into activities. Rather, they were recorded successively as modular units to ensure the creation of a balanced dataset; that means, all activities have a similar number of recordings regardless of their occurrence in a given scenario. All standing and walking activities have been recorded for five minutes per participant in 5 individual recordings of 60sec each. Both the activities *Reaching forward (none)* and *lifting (box)* were recorded in a single run to reduce the recording effort. The box was picked 10 times from 9 different heights, from the ground level up to a stock of 8 boxes. The participants approached the stack from different starting positions to ensure a natural motion. The boxes had to be lifted with both hands.

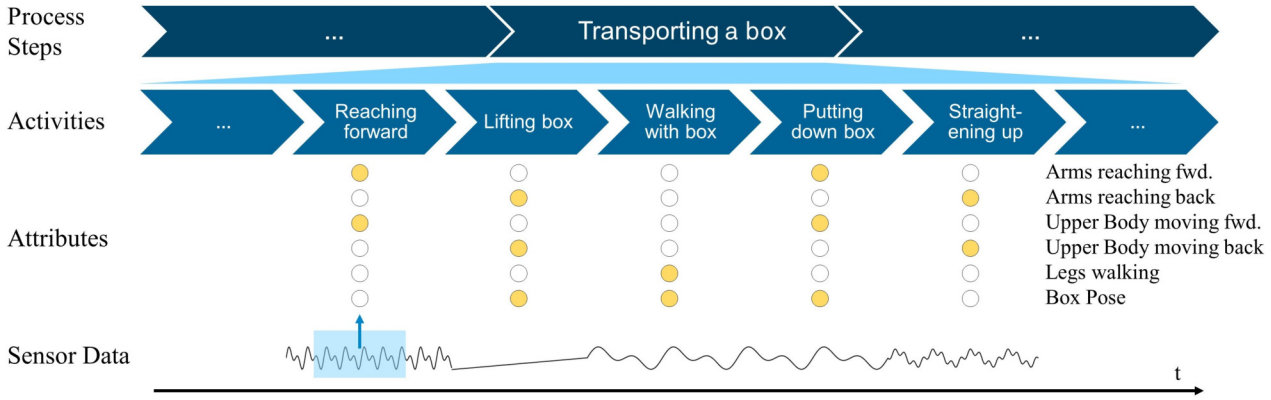


Fig. 3: Attribute based representation of a process step composed of activities that are semantically described

Apart from that, no instructions were given. The total amount of 90 recordings was likewise recorded for the activities *put down (box)* and *straighten up (none)*. A testing data set of 60 sec were recorded for each participant. In the testing data set, the participant conducts a sequence of the previously classified and annotated activities in an arbitrary order and duration. This data set is manually annotated. 202 recordings were conducted with eight participants each, resulting in a total of 1616 recordings. As the data set is based on skeleton poses, one can visualize them easily for annotation purposes. The annotation of walking and standing data sets is simple, as there is no alteration of neither the attributes nor the activity over time. The activities that included the stacked boxes contained not only the two activities *Reaching forward (none)* and *lifting (box)*, as well as *put down (box)* and *straighten up (none)*. The stack was approached and departed by the participants by feet. Therefore, the two *walking* activities and the two *standing* activities were annotated as well. Having this modular recording from activities, the annotation took approximately $2.5min$ per recorded minute.

D. Convolutional Neural Networks for HAR based on Skeleton Datasets

This paper uses the deep architectures, proposed in [6], [9]. These architectures are suitable for multichannel time-series. They are composed of temporal-convolutions and pooling layers, which perform convolution and downsampling operations along the time axis. These architectures extract hierarchical-temporal relations of human movements creating abstract representations of an input sequence. Fully-connected layers connect these representations creating a global one of the input data. The network will compute an attribute representation of an input sequence. This representation is a vector $a \in \mathbb{B}$ containing $1s$ and $0s$ in which 1 for having or not an attribute, the sigmoid activation function is applied to each element of the output layer. Its output corresponds to pseudo-probabilities for each attribute a_i being present in the representation.

The architecture was designed for handling sequences from multichannel time-series, which are measured from m individual portable-devices. These devices are located on different parts of the human body. Convolutional and pooling layers are configured in parallel branches for processing these sequences. Specifically, a single branch processes sequences from a single device increasing the descriptiveness. This architecture is called CNN-IMU. Besides, this configuration allows for more robustness against different and asynchronous devices. This architecture contains m convolutional branches, one per device. Each branch is composed of four temporal-convolution, two max-pooling layers and a fully-connected layer.

Different from [6], [9], the input sequences are not measurements from any portable sensor located on human body parts. Sequences are provided from the MoCap System, see subsection II-A, which provide global poses of human segments. Then, for each of these segments, one has six different measurements. There are in total 22 human segments, e.g. the head, torso, feet, knees and arms. In total, 134 channels have been taken into account. The global pose sequences are normalized with respect to the lower back human-segment. This is necessary to avoid a dependency of the human activity recognition to a global position of warehousing equipment in the laboratory. Each of this measurements is taken as a channel, similar to sequences from portable devices. One considers in total 132 channels and $m = 22$ branches. In the CNN-IMU, convolutions are computed along the time axis, and their filters are shared among the channels.

For training, the following configurations are employed. Sequences from persons $1 - 6$, person 7 and person 8 are used as training, validation and testing sets respectively. The parameters of the networks are updated by minimizing the binary-cross entropy loss using the stochastic gradient descent with the RMSProp update rule as in [7], [9]. Sequence segments, extracted using a sliding window approach, are fed to the networks. These segments are assigned the most frequent ground truth. In general, learning rates are decreased by $\gamma = 0.1$ at a certain epoch or iteration during training. Additionally, we use dropout with probability of 50% on the inputs of the first and second fully-connected layer, and orthogonal initialization [7]. As suggested in [6], [7], input sequences were normalized per channel to a range $[0, 1]$. Moreover, a Gaussian noise of $\mu = 0$ and $\sigma = 0.01$ is

added, simulating inaccuracies on the MoCap System. For a given attribute representation A describing the aforementioned activities in the reference dataset, a nearest neighbour approach is used for predicting a specific activity by measuring the cosine distance from the CNN's output for a certain input sequence $\tilde{a}$ to the set $a \in A$. Different sets A of attribute representations, provided by experts, will be evaluated.

E. Human Validation

Following a sliding window approach with a window size of T and step of s, an unlabeled sequence from the reference dataset and an unlabeled sequence from IMU's measurements are segmented. A set of D sequences of size T are then obtained. These sequences are fed to the CNN-IMU computing their attribute representations. By means of a nearest neighbor, these sequences are assigned to the activity where the distance between their representations is minimal. Following [20], an uncertainty measure can be computed for each of the predictions. This measure give a value of how certain a CNN is with respect to a prediction. Uncertain predictions are then revised by experts for generating the final annotation of the sequence.

III. Discussion and Conclusion

This contribution proposed a framework to reduce the annotation effort for multichannel time-series. An attribute based representation creates a high-level semantic description of activities. This is beneficial to make full use of imbalanced data, avoid overfitting and to recognize unseen activities. The logical connection of activities and process steps has been explained and an exemplary attribute representation has been provided. Motion Capture datasets of eight activities including a training and validation data set have been recorded with eight participants each, resulting in a total 1616 recordings. The recordings have been annotated, normalized and used to train a state-of-the-art CNN. Recording further participants and the manual annotation of the MoCap data requires few manual effort. The attributes used by the CNN can be understood by a human and thus transferred to new activities.

Based on the proposed framework, a large multichannel time series can be annotated with respect to semantics in a semi-automated manner. It is not restricted to IMU data but can be used for other sources, such as video data, as well.

References

[1] R. Manzini, Ed., *Warehousing in the global supply chain: advanced models, tools and applications for storage systems*. Springer, 2012.

[2] E. H. Grosse, C. H. Glock, and W. P. Neumann, "Human factors in order picking: a content analysis of the literature," *International Journal of Production Research*, vol. 55, no. 5, pp. 1260–1276, Mar. 2017.

[3] K. Weisner and J. Deuse, "Assessment Methodology to Design an Ergonomic and Sustainable Order Picking System Using Motion Capturing Systems," *Procedia CIRP*, vol. 17, pp. 422–427, Jan. 2014.

[4] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 1–33, Jan. 2014.

[5] S. Feldhorst, M. Masoudenijad, M. ten Hompel, and G. A. Fink, "Motion Classification for Analyzing the Order Picking Process Using Mobile Sensors," in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, ser. ICPRAM 2016. Portugal: SCITEPRESS - Science and Technology Publications, Lda, 2016, pp. 706–713.

[6] R. Grzeszick, J. M. Lenk, F. M. Rueda, G. A. Fink, S. Feldhorst, and M. ten Hompel, "Deep Neural Network based Human Activity Recognition for the Order Picking Process," in *Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction*. ACM Press, 2017, pp. 1–6.

[7] D. R. Francisco Javier Ordóñez, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. Advances on Data Transmission and Analysis for Wearable Sensors Systems, p. 115, 2016.

[8] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables," *CoRR*, Apr. 2016.

[9] F. M. Rueda and G. A. Fink, "Learning Attribute Representation for Human Activity Recognition," *arXiv:1802.00761 [cs]*, Feb. 2018. [Online]. Available: http://arxiv.org/abs/1802.00761

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *British Machine Vision Conference*, 2015.

[11] S. Feldhorst, S. Aniol, and M. ten Hompel, "Human Activity Recognition in der Kommissionierung – Charakterisierung des Kommissionierprozesses als Ausgangsbasis für die Methodenentwicklung," *Logistics Journal : Proceedings*, vol. 2016, no. 10, Oct. 2016.

[12] H.-T. Cheng, F.-T. Sun, M. Griss, P. Davis, J. Li, and D. You, "NuActiv: Recognizing Unseen New Activities Using Semantic Attribute-based Learning," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: ACM, 2013, pp. 361–374.

[13] J. Zheng, Z. Jiang, and R. Chellappa, "Submodular Attribute Selection for Visual Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 39, Issue: 11,)*, ser. 11, vol. 39. IEEE, Nov. 2017, pp. 2242 – 2255.

[14] A. Vinciarelli, A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal, D. Heylen, R. Kaiser, M. Koutsombogera, A. Potamianos, S. Renals, G. Riccardi, and A. A. Salah, "Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions," *Cognitive Computation*, vol. 7, no. 4, pp. 397–413, Aug. 2015.

[15] A. K. R. Venkatapathy, H. Bayhan, F. Zeidler, and M. t. Hompel, "Human machine synergies in intra-logistics: Creating a hybrid network for research and technologies," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2017, pp. 1065–1068.

[16] J. Haase and D. Beimborn, "Acceptance of Warehouse Picking Systems: A Literature Review," in *Proceedings of the 2017 ACM SIGMIS Conference on Computers and People Research*, ser. SIGMIS-CPR '17. New York, NY, USA: ACM, 2017, pp. 53–60.

[17] D. Battini, M. Calzavara, A. Persona, and F. Sgarbossa, "Additional effort estimation due to ergonomic conditions in order picking systems," *International Journal of Production Research*, vol. 55, no. 10, pp. 2764–2774, May 2017.

[18] R. Müller-Rath, C. Disselhorst-Klug, S. Williams, C. Braun, and O. Miltner, "Einfluss des Geschlechts und der Seitendominanz auf die Ergebnisse der quantitativen, dreidimensionalen Bewegungsanalyse der oberen Extremitäten," *Zeitschrift für Orthopädie und Unfallchirurgie*, vol. 147, no. 04, pp. 463–471, Jul. 2009.

[19] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millàn, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, Jun. 2010, pp. 233–240.

[20] R. Grzeszick, S. Sudholt, and G. A. Fink, "Optimistic and Pessimistic Neural Networks for Scene and Object Recognition," *arXiv:1609.07982 [cs]*, Sep. 2016.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 823–826
DOI: 10.15439/2018F164
ISSN 2300-5963 ACSIS, Vol. 15

The model of local e-administration development

Agnieszka Agata Tomaszewicz
University of Szczecin
Faculty of Management and
Economics of Services
ul. Cukrowa 4, 71-004 Szczecin Poland
Email:agnieszka.tomaszewicz@wzieu.pl

***Abstract*—ICT which constitutes basis for e-administration development, is used more and more often in office-citizen kind of communication. Thus, terms regarding the state and directions of local e-administration development are quite often discussed in polish literature. However, they are seldom held on the local level i.e. local authorities. The issues regarding improvement of service among local communities resulting from applying e-administration solutions are particularly researched. Hence, the article presents the results of enhanced scientific research in the area of impact of local e-administration solutions on service efficiency among citizens which result in proposal of the model of local e-administration development. Its' application shall enable proper usage of local e-administration potential as well as improving service for local communities.**

I. Introduction

IN TODAY'S economy, the Internet plays an important role as a mean of communication and distribution and as such it shall be used intensively by local units in order to improve the service for their communities. The Internet shall be used by offices to improve the whole process of providing services so that it would result in implementation of full offer of e-services available on the offices' web pages. Providing dynamic development of e-administration and wide usage of provided services via electronic channels shall result in better, more efficient and effective functioning of local offices. It will foster both, more efficient customers service and implementing new computerised solutions enabling growing customers' needs and even stay ahead of them. Thus, it may be assumed that e-administration is public administration in which the use of information and communication technologies (ICT) contributes to better administrational service for citizens thereby improving the quality of their life. On the basis of the needs and improvements identified in the research, which result from the implementation of e-administration in municipalities of West Pomeranian Voivodeship, it was concluded that self-governmental administration has the potential which may contribute to improvement of services for local communities. The development of public e-services, particularly at the initial stage of development, depends on the access and the quality of public e-services which specify the advance level of their development. Therefore, in order to increase the access to public e-services in municipalities, the growth of their maturity and improvement of customer service, resulting in better quality of life regarding citizens and better quality of realized business processes regarding entrepreneurs, the model of local e-administration development was created on basis of research held in municipalities of West Pomeranian Voivodeship. The municipalities do not vary from the whole population regarding the researched qualities.

II. Literature review

The innovative approach towards improvement of governmental effectiveness, according to M.A. Abramson, J.D, Breul, J. M. Kamensky, is driven by the technological development which lead to vital changes in functioning of organizations both, in public and private sector contributing to improvement of administration efficiency. They claim that technology shall be seen not only as a basic activity of public administration but as a driving force for its activity [1]. The biggest advantage of that kind of innovation is to create grounds to change the nature of business and interpersonal communication and to establish new relationships between people and organizations [2]. According to H. Izdebski [3], progress connected with the technical development of IT facilitates contact with offices by obtaining desired information or settling the matter via electronic way. Thanks to the development and better access to the technologies, the meaning of electronic administration shall be systematically growing [4]. The influence of the administration on the economy, particularly by using ICT in recent years is seen as a factor which boosts economy and leads to public sector transformation, drastically changing the way of functioning of public institutions [5]. Within the research conducted for the needs of eGovernment Readiness Index, five e-administration models were found [6]:

1. centralized- where the information and public services system is organized around the main national portal and presented information are highly unified;
2. decentralized- based on the individual sites, created separately for particular institutions, initiatives and programs in which collective public platforms play only referential function and presented information are not standardized;
3. network- in which systemic character is obtained by the number and kind of links between particular websites;
4. e-participation oriented- in which the base constitute the tools enabling citizens' engagement in creating the administrational processes and making decisions;
5. e-services oriented- in which computerization of the processes like *back-office* and *front-office* are treated as the most important factor in creating e-administration systems.

The approach to local administration keeps on evolving all the time. Currently, it is expected that it shall implement citizens' participation in the process of creating public services. Additionally, in places where it has not been implemented yet, administration shall pass from top-down hierarchical effectiveness to bottom-up, meaning democratic one. This effectiveness is measured on the basis of the results and in the given context. As far as changes in new administration paradigm are considered, in e-administration focus on the client shall prevail.

III. Research methodology

A. Population of the study

The subject and the main area of the research were local authorities in West Pomeranian Voivodeship (urban, including city with powiat rights, urban-rural and rural) and local communities which are serviced by these authorities. The research includes the subjects: (1) Local authorities in West Pomeranian Voivodeship in which research was conducted on three stages: (Stage I) - online survey regarding the state of local e-administration; (Stage II) - the analysis of the websites of municipalities chosen from stage I of the research; (Stage III) - direct interview with the representatives of the local authorities; (2) Local communities of the West Pomeranian Voivodeship which were included in the survey.

B. Data collection

The model was elaborated on the basis of the used secondary and primary information sources by using the following research methods: CAWI technique *(Computer Assisted Web Interviews)*, survey technique, was used among local communities which was drawn on the basis of random sample; direct interview, applied in order to identify problems connected with the development of local administration; critical analysis method- including observation of the websites' content method and applying e-administration solutions in municipalities in West Pomeranian Voivodeship; analysis method of the service provided for the local communities by local authorities in order to identify needs and expectancies within the scope of e-administration; *Case Study Method*- presenting particular solutions connected with local e-administration.

The data collected from the primary and secondary sourced was used to create the model of improving local e-administration.

C. Reliability and validity

The questionnaire that was used in the present study was rigorously tested for its content and construction validity. A draft of the final questionnaire was shown to two officials and three academics, in order to test whether it met all theoretical and practical requirements.

The research conducted at the second stage regarded mainly functioning of the websites on the task level. The subject of the research were local e-services. The content of the offices' websites was not analyzed multidimensionally but it focused merely on the basic terms within the scope of adjusting the websites to providing services for local communities.

IV. Conceptual framework

The basic approach in the proposed model is process approach. Implementing the model which enables meeting the requirements of the local communities, requires its' constant adjustment to changing citizens' and environment's needs. The model was graphically presented in the Figure 1.

A. Modules of local e-administration

In the structure of the model aiming at creating coherent work of local e-administration, it was necessary to identify particular stages designed to improve the service for local communities by local authorities units. The model includes five basic modules which constitute the core of local e-administration (e-A). They include: (1) inventory of the customers service processes in the local offices, (2) recognizing the needs of local customers (citizens, entrepreneurs, other public institutions), (3) the standardization and improvement of processes and public services (in the traditional and electronic meaning), (4) interoperability of the processes and implemented/developed e-services by development and closer cooperation of the units engaged in the process of providing e-services, (5) digitalization of the public services/ the increase in maturity of the available public e-services. In order to specify the level of citizens' satisfaction, offices may use monitoring. Due to changeable character of local communities' needs, the process of monitoring must be in compliance with the rule of continuous improvement by Deming [7]: constant checking whether implemented solutions still respond to citizens' preferences. This activity will not require financial or material outlays. Specifying the needs and preferences of the final customer of local e-administration influences the increase in adaptability of the proposed model.

Bearing in mind, that the customers' needs are tightly connected with the realization of public services, it is legitimate that the accepted model is targeted at creating new

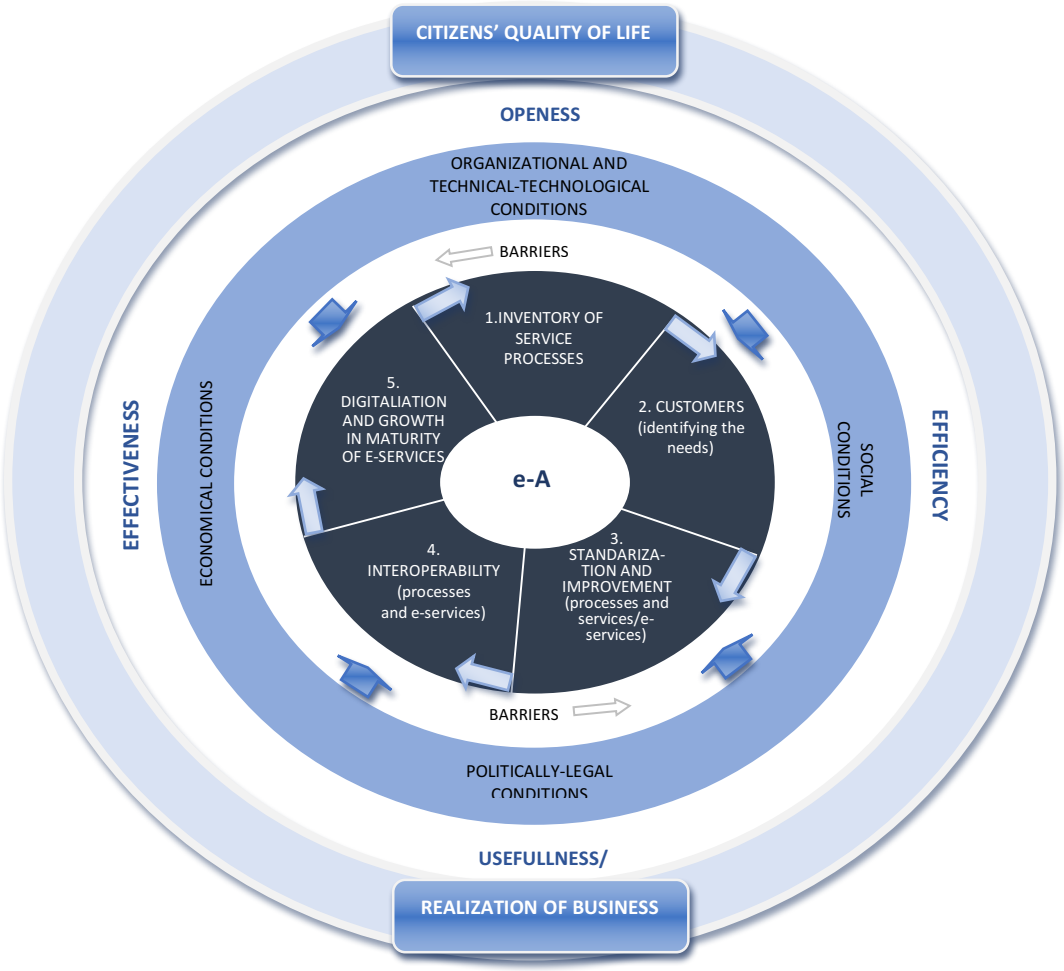


Fig.1 The model of improving local e-administration (e-A)
Source: own elaboration

services and expanding maturity of the existing e-services which main objective is to be adjusted to the local communities' needs. Digitalization constitutes the basic tool in creating social models, including local administration units which shall broadly use its potential, particularly in stimulating social development.

B. Barriers in e-administration development

Implementing the model ay encounter numerous obstacles which may include; organizational barriers, legal, economic, political, technological or social. They may occur on different staged and levels and ay influence slowing the whole process down. One of the most prominent limits are lack of interoperability, lack of unified standards on municipalities' websites, customers' low knowledge on ICT use, necessity of autonomous search of information while handling the matters on the Internet or too complicated service implementation process via the Internet. The access to ICT technology or the willingness to handle the official matters online are believed to be the most crucial obstacles. The obstacle in implementing e-administration in municipalities may be unclear or the websites be not functional. In order to remove this barrier, it is particularly necessary for offices to share services via one website especially designed for it. Currently, councils use even 4 websites (council's own website, BIP, eBOI and ePUAP) having discretion in posting e-services on them. The key barrier which often is finally associated with responsibility for success in e-administration implementation is the cost barrier. EU funds are particularly helpful in this area. From the implementation of the model point of view, particularly important stage is identification and fast elimination of the existing barriers or their removal. These activities shall be consequences of continuous improvement of the e-administration process in communal authority units.

C. Conditions for e-administration development

The success of realizing the model for e-administration development depends on the understanding and knowing the external conditions which may influence it. These conditions may enable it to reach the highest stage of public e-services development, thus, constituting administrational offer for the users. Conditions which are particularly emphasized, include

technically-technological, economic, politically-legal, social and organizational conditions. Besides presenting determinants specified in the model of e-administration development one shall also refer to the theory of critical success factors. In Poland the most important critical success factors (CSFs) of e-administration, as well noticed by E. Ziemba and T. Papaj, are these which are connected with finances, integration and interoperability of public e-services with various systems of public administration institutions, employees' IT skills and top-level management engagement, information safety or implementing innovative teleinformatic infrastructure in public institution, availability of free software [8],[9],[10].

V. Empirical Results

The results of the conducted research enable formulating the following detailed conclusions: 1. E-administration improves communication between citizens and councils. Along with the development of the Internet, this form of communication will be more popularized and intensified. 2. Municipalities of West Pomeranian Voivodeship are characterized with low level of local e-administration development. 3. Low level of local e-services maturity influences rather poor citizens' interest in handling matters with the use of the Internet. In West Pomeranian Voivodeship amounts to 47% (including only 69% of citizens who only download the electronic forms from the websites). However, the vast majority of the participants, up to 74%, feel the need to fully manage at least one out of 12 basic official issues. 4. The increase in advance level of public e-services development and functionality of the websites shall contribute to better service provided for local communities. 5. Higher percentage of citizens who used the Internet while dealing with official matters, claim that it had a positive effect on the quality of the services provided than these people who claimed that there was no improvement. 6. The level of adjusting the websites and their functionality are varied and it is hard to evaluate them clearly. However, mostly customers due to council's use of few internet portals, encounter obstacles in finding the matter they are interested in because it may be provided on different websites. 7. A meaningful problem remains; the lack of coherence and standardization of documents. Local authorities shall establish cooperation with other units, particularly on local level. 8.The factor conditioning wider scope, form and intensity of public e-services use is obtaining proper digital skills both, by employees and customers. IT courses may seem helpful in this area.

VI. Conclusions

The local e-administration has potential to enable, in a considerable manner, contributing to improved service provided for local communities. In conclusion, essential factors which condition the success of implementing the model of local e-administration are as following: 1. Noticing the need for changes by local authorities and their engagement in the process of implementing solutions within the scope of local e-administration on every stage. 2. Accepting efficiency, effectiveness, openness and usability as basic results' measures of implementing the model of e-administration development. 3. Implementing standardization and interoperability as basic factors responsible of work efficiency. 4. Reliable valuation of the project costs and guarantee in financing it. 5. Obtaining proper IT skills by employees and social communities. 6. Using the teleinformatic technology. 7. Participation of local communities and their approval for new tools implemented in municipal administration. 8. Coordination, cooperation, monitoring and evaluation. 9. Constant adjustment to changing needs of the environment.

The benefits of implementing the model of local e-administration are connected with improved skill of satisfying the needs of local communities as a result of the maturity growth (improving the quality of services) or sharing new public e-services.

References

[1] M. A. Abramson, J. D. Breul, J M. Kamensky, Report "*Six Trends Transforming Government*", IBM Center for the Business of Government, Washington 2006, p. 4.

[2] A. Budziewicz-Guźlecka, *Management of changes in enterprises as a form of adaptation to e-economy*, Scientific Journal No. 681 Service Management Vol. 8, Szczecin 2012, p.202

[3] H. Izdebski, *Samorząd Terytorialny. Podstawy Ustroju i Działalności*, Edition II, Wydawnictwo Prawnicze Lexisnexis Sp. z o.o., Warszawa 2003, p. 47.

[4] P. Minkowski, P. Motek, R. Pedał, *Poziom zaawansowania wielkopolskich urzędów gmin w zakresie informatyzacji i rozwoju elektronicznych usług publicznych*, Wydawnictwo M-Druk, Poznań 2009, p. 8.

[5] M. Kowalczyk, *E-Urząd w komunikacji z obywatelem*, Wydawnictwa Akademickie i Profesjonalne, Warszawa 2009, p.7.

[6] G. Curtin, *Global E-Government/E-Participation Models*, Measurement And Methodology: A Framework For Moving Forward, Unipan, New York 2006, p. 21-24, http://unpan1.un.org/intradoc/groups/public/documents/un/unpan023680.pdf.

[7] The Deming Cycle (or Deming Wheel), is referred to as the PDCA Cycle (Plan-Do-Check-Act), is the W.E concept. Deming, containing chronologically ordered actions aimed at continuous improvement. it takes place in the cycle: planning - execution - checking - operation. (http://quality-managemnet.pl/pdca/70-14-zasad-filozofii-deminga.html).

[8] E. Ziemba, T. Papaj, *Determinanty I Bariery Rozwoju E-Administracji W Polsce*, conference materials, http://www.dabrowa-gornicza.pl/portal/download/file_id/29123/pid/1763.html

[9] S. Assar, I. Boughzala, I. Boydens, *Practical studies in e-Government. Best Practices from Around The World*, Springer, 2011, p.1.

[10] L. Alzahrani, W. Al-Karaghouli, V. Weerakkody, *Analysing the critical factors influencing trust in e-government adoption from citizens' perspective: A systematic review and a conceptual framework*, International Businnes Review, Volume 26, Issue 1, 2017, p. 70-71; https://doi.org/10.1016/j.ibusrev.2016.06.004.

24th Conference on Knowledge Acquisition and Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference “Knowledge Acquisition and Management". The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we’ve joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The aim of this event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

TOPICS

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

EVENT CHAIRS

- **Hauke, Krzysztof,** Wroclaw University of Economics, Poland
- **Nycz, Malgorzata,** Wroclaw University of Economics, Poland
- **Owoc, Mieczyslaw,** Wroclaw University of Economics, Poland
- **Pondel, Maciej,** Wroclaw University of Economics, Poland

PROGRAM COMMITTEE

- **Abramowicz, Witold,** Poznan University of Economics, Poland
- **Andres, Frederic,** National Institute of Informatics, Tokyo, Japan
- **Bodyanskiy, Yevgeniy,** Kharkiv National University of Radio Electronics, Ukraine
- **Chmielarz, Witold,** Warsaw University, Poland
- **Christozov, Dimitar,** American University in Bulgaria, Bulgaria
- **Jan, Vanthienen,** Katholike Universiteit Leuven, Belgium
- **Mach-Król, Maria,** University of Economics in Katowice, Poland
- **Mercier-Laurent, Eunika,** University Jean Moulin Lyon3, France
- **Sobińska, Małgorzata,** Wroclaw University of Economics, Poland
- **Surma, Jerzy,** Warsaw School of Economics, Poland and University of Massachusetts Lowell, United States
- **Vasiliev, Julian,** University of Economics in Varna, Bulgaria
- **Zhu, Yungang,** College of Computer Science and Technology, Jilin University, China

ORGANIZING COMMITTEE

- **Hołowińska, Katarzyna**
- **Przysucha, Łukasz,** Wroclaw University of Economics

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 829–838
DOI: 10.15439/2018F136
ISSN 2300-5963 ACSIS, Vol. 15

RC-ASEF: An open-source tool-supported requirements elicitation framework for context-aware systems development

Unai Alegre-Ibarra, Juan Carlos Augusto, Carl Evans
Middlesex University, London NW4 4BT, U.K.
Email: {U.Alegre, J.Augusto, C.Evans}@mdx.ac.uk

***Abstract*—In terms of software engineering, context-aware systems (C-AS) have notably different development needs than those of traditional computing. Yet, there are no established methodologies that uniformly support the development life-cycle of these systems. A key goal of this research is to improve the current state-of-the-art with respect to engineering techniques for the life-cycle of a C-AS. Within the scope of this higher order goal, this paper addresses the lower level order goal of a holistic framework for gathering requirements which is specialised to the creation of C-AS. The framework follows an end-user, stakeholder-centred vision, which guides the analysis of stakeholders towards the discovery of specific stakeholder profiles and their particular needs, preferences, and limitations. It allows the operationalisation of the high level objectives of the system into requirements, which are more tangible and related to the implementation of the system. An evaluation procedure is supported, based on heuristics and rules from the NFR framework and REUBI. All the diagrams introduced for this framework have been developed as part of an open-source tool based on Modelio, which is intended to be developed in the future as part of a framework that covers all the stages of the development process. The proposal is illustrated through the analysis of an application for a European funded project.**

I. Introduction

The creation of context-aware systems (C-AS) can entail a great amount of challenge and complexity [1]. In comparison to the development of traditional systems, C-AS are more expensive, diverse, and prone to change. During its creation, developers might find difficult, or even impossible to identify the situations in which to display services, as well as what services should be exhibited in those situations [2]. Also, developers might find it challenging to describe the information to identify these situations, and make the system aware of them by using a heterogeneous array of sensors, which are likely to provide inaccurate, overlapping, contradictory or missing data. Additionally, advanced reasoning techniques need to be implemented in order to make the system infer that situations are happening, based on sensor information. This intricacy emphasises that there is a substantial difference between developing conventional systems and those that are context-aware. As part of previous research, an extensive analysis has studied the different approaches to the development of C-AS [1]. Although there is lot of research related to the development of these kinds of systems, this is focused on solving particular issues and it is usually scattered and not connected with other development stages. The evidence presented in [1] supports the need of a more holistic and unified approach for the development of C-AS.

A key goal of this research is to improve the current state-of-the-art with regard to techniques and methods to help establish the foundations of a uniform engineering process that covers the entire life-cycle of a C-AS. As part of a lower-order goal of the bigger picture, this paper focuses on the creation of the foundations for the *Requirements for Context-Aware Systems Framework* (RC-ASEF), a holistic framework for the requirements elicitation stage, which takes into account the specific demands of C-AS development. The support provided for the requirements elicitation stage in RC-ASEF is divided into two main foci. During early stages of the requirements elicitation process, the methodology is focused on the generic or non-contextual aspects of the system (F_1), to then iteratively advance towards the requirements which are more related to the identification of situations (context), the way in which they are planned to be detected by the system, and their associated context-aware features (F_2). Previous research towards the high order goal of this work has focused on F_2, creating a deeper analysis into the conceptualisation of context and context-awareness [2], which takes into account the philosophical limitations of C-AS in order to create a perspective for developing more usable C-AS. The aim of the work presented in this paper is focused on a generic methodology for gathering the non-contextual aspects of a C-AS, corresponding to F_1, reusing existing methods and tools to provide a coherent requirements elicitation methodology that can cover the demands of C-AS development. Particularly it has been developed with reference to previous work [3] [4] [5], including the mentioned conceptualisation of context and context-awareness [2]. In particular, the framework is based on a collection of models, presented as a combination of dynamic and static diagrams which collectively define this new requirements elicitation framework, for which in addition, new, open-source tools have been developed. These constructs have been strategically chosen to be based on UML profiles, as this facilitates its use along with other existing standards such as UML [6], SysML [7], and U2TP [8], or other UML-based requirements profiles such as UML-AT [9]. These diagrams have been developed as an extension of the open-source tool

framework RC-ASE, Modelio[1], which is further introduced in [2], and which can be found in [10]. The reminder of the paper is as follows. Section II analyses previous work in relation to the framework. Section III introduces the requirements elicitation framework. Section IV is related to the establishment of a project scope. Section V corresponds to the stakeholder analysis of the methodology. Section VI is related to the objective establishment activity. Section VII corresponds to the identification of functional requirements. Section VIII corresponds to the evaluation activities of the methodology. Finally, Section X concludes the paper.

II. Related work

Previous work related to requirements which are specialised for C-AS can be divided into three main groups: scenario-based, goal-oriented and hybrid approaches. Related work has been reviewed [1] in order to find an approach that can be suitable for creating a more holistic approach, which reuses the most positive aspects of existing methodologies and tools. This paper focuses on an analysis of methodologies for a number of aspects that are considered relevant for the higher order goal of creating a more holistic framework for C-AS development, as shown in Table I. Columns 4, 5 and 6 focus on the coverage of the methodologies for the most common elicitation activities for C-AS development. These activities are covered by most of the analysed methodologies. Columns 8 and 10 represent whether or not the methodology is based on goals, scenarios or a hybrid approach. Column 9 describes if the methodology has support for the partial satisfaction of goals rather than just being binary. Column 11 indicates whether or not the methods provide specific and systematic treatment of non-functional requirements. The most complete approach is that of REUBI [3], which has the potential to cover all these approaches. Column 12 indicates whether or not the methodology takes into account the needs, preferences and limitations of the end-users, or in its absence, they support personalisation to a certain degree. Only three methodologies support this feature, from which PC-RE [4] and R4IE [5] are highlighted. Column 13 shows whether or not the methodologies take into account the influence of contextual aspects. Many methodologies support this, but each has its own particular way to manage this. Column 14 indicates if the methodology has specific support guiding developers into: (a) enumerating the set of contextual states that may exist, (b) knowing what information could accurately determine a contextual state within that set, and (c) stating what appropriate action should be taken from a particular state [2]. Oyama et al. [11] present a series of templates for this purpose which could be reused for other methodologies. Columns 15 and 16 show whether or not the approaches have tool support and if such tool is freely and easily available for other researchers to be extended.

From the point of view of the analysis of those aspects, REUBI [3] is the most complete methodology, as it can be observed in Table I. Nevertheless, there are three main aspects that this method does not completely cover. Namely, the explicit lack of a user-centred perspective, and a lack of a tool which is publicly available. Also, it does not provide guidance for developers to discover context, according to the three main principles to get the context right [12] [2]. From the point of view of guiding the developers towards the discovery of situations and context [2], the data-information-knowledge-wisdom model of Oyama et al. [11] could also be employed for this purpose. Nevertheless, Oyama's model lacks mechanisms for elaborating and modelling requirements. Compared to REUBI [3], it also lacks mechanisms for handling soft goals and non-functional requirements. Additionally, there is no tool support for this approach. For the purpose of this research, the REUBI methodology [3] is the most relevant reference point. Therefore, it is concluded that REUBI will be used as the foundation from which the requirements framework for engineering C-AS will be built. Although REUBI has partial support for scenario based techniques, which can be used to understand and gather the context of the system, scenario-based techniques are not a necessary requirement for this work, and they can be further complemented with other techniques. A necessary aspect that needs to be covered for this work, is that of the user-centred perspective. This gap can be addressed by combining other existing methodologies. The R4IE [5] and PC-RE [4] approaches have some synergies that can be used to complement this characteristic. Additional techniques for analysing stakeholders and their needs can also be useful for this purpose. In order to address the shortcomings related to guiding developers into context discovery, a set of guidelines which are based on the perspectives of [2] will be included as part of the methodology. Finally, and significantly, whilst the REUBI approach has no explicit open-source tool support, the work described here has a specific goal of developing an open-source tool to support the proposed software development framework, which includes specific support for requirements engineering.

III. RC-ASEF: Requirements for Context-Aware Systems Engineering Framework

Figure 1 presents the six main activities of a coherent methodology out of the most relevant approaches identified for the purpose of creating a framework for supporting the non-contextual aspects of the requirements elicitation, influenced by R4IE [5]. The main enhancement is that the identification of system performance qualities, used for gathering non-functional requirements, is now part of the objective establishment. A new activity group, corresponding to the evaluation of the objectives and requirements, which is partially based on the harmonisation activity from R4IE, is introduced. The activities in Figure 1 are divided into different sub-activities, as shown in Figure 1, which are mainly influenced by the works presented in [3] [5]. The method gives great importance to the exhaustive analysis of the stakeholders of the systems, as part of the identification of their needs and preferences in further stages. The sub-activities constitute an enhancement

[1]https://www.modelio.org/

No.	Name	Year	Reference	Acquisition	Elaboration	Modelling	Goal-oriented	Softgoals	Scenario based	NFR treatment	User-centred	Context-awareness	Developer Guidance for context	Tool support	Open source
	(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(a)	-	2001	[13]	-	✓	✓	✓	-	-	-	-	✓	-	-	-
(b)	PC-RE	2006	[4]	✓	✓	✓	✓	-	✓	-	✓	✓	∼	-	-
(c)	RE-CAWAR	2007	[14]	-	✓	✓	✓	-	✓	-	-	✓	∼	-	-
(d)	-	2008	[11]	✓	-	-	✓	-	-	-	✓	✓	✓	-	-
(e)	-	2008	[15]	✓	✓	∼	-	-	✓	-	∼	✓	∼	✓	-
(f)	FLAGS	2010	[16]	-	✓	✓	✓	✓	-	-	-	-	-	-	-
(g)	REUBI	2013	[3]	∼	✓	✓	✓	✓	∼	✓	∼	✓	∼	✓	-
(h)	R4IE	2014	[5]	✓	✓	-	∼	-	∼	∼	✓	✓	-	-	-

✓ = The property is completely satisfied. ∼ = The property is partially satisfied. "-" = The property is not satisfied at all.

TABLE I
COMPARISON OF CURRENT METHODOLOGIES FOR REQUIREMENTS ENGINEERING IN TRADITIONAL AND CONTEXT-AWARE SYSTEMS.

of the R4IE methodology, where the first sub-activity of the stakeholder analysis is inspired by [17], and the second sub-activity is impacted by the profiling of users [4] [5], the ethical analysis recommendation in [17], and the e-FRIEND ethical framework [18]. It is also influenced by the conceptualisation presented in [2]. The last activity in the stakeholder analysis, and the sub-activities related to the establishment of objectives, have been adopted from [3]. Finally, those sub-activities corresponding to the identification of functional requirements and the application of the evaluation procedure are inspired by those activities in [3], and influenced by the heuristics and rules from the NFR Framework [19] as well as the SySML [7] standard.

IV. ESTABLISH SCOPE

The central activity of the methodology during F_1, is to establish the scope of the system in terms of the system boundaries (*i.e.*, what is inside the system and what is immediately external to it) . As it can be observed in [5], this activity is influenced by the remaining core activities in F_1, which help to determine the objectives, resources, budget and schedule to be included within the scope statement.

V. STAKEHOLDERS ANALYSIS

The initial step consists of a stakeholder analysis, which allows documenting and modelling the outcome from the array of techniques proposed in [17], using a UML profile for the creation of Stakeholder Diagrams. The stakeholders are identified, and their different relevant relationships to the project are analysed. The outcome of this activity is used as part of the scope statement and part of the models. Finally, the aim is to identify different user profiles, in order to pave the way for discovering useful Situations of Interest in F_2. Using the information gathered during the stakeholder analysis, it focuses on the identification of activities.

A. Identify stakeholders

This activity is initiated by a small group, and later reviewed with a larger group of stakeholders. After the review with a larger group of stakeholders, the participants should think about those stakeholders who are still not included. If there are more interested parties, a bigger group should be assembled to review the stakeholders [17]. This process iterates until a consensus has been arrived at such that it is considered that all relevant stakeholders have been accounted for. A set of techniques are recommended to guide this process [17]. Each of which can build on the previous technique, and it includes the listing of stakeholders, its basic analysis, the power versus interest grids as well as the stakeholder influence diagrams. The stakeholder identification task can also be complemented with a stakeholder analysis, as further explained in [17].

B. Determine stakeholder profiles

The aim of this activity is to identify stakeholder profiles, by establishing personal goals and setting different levels of achievement. The user profiling is attained by setting certain achievement levels and monitoring progress towards those personal goals [4]. In order to set the achievement levels, three main dimensions are analysed during F_1, which include the cultural aspects of the stakeholders, their quotidian activity, and their relevant ethical aspects. Finally, the information obtained from this analysis is used to customise the requirements, as well as the system set-up and training. In activities related to stakeholder profiling corresponding to F_2, other dimensions are analysed, namely, the interaction modalities, and the mechanisms for monitoring the achievement of personal user goals. The user profiling activity is mainly based on the activity with the same name in R4IE [5], but it also includes the cultural analysis and profiling guidelines from PC-RE [4] and the ethical analysis mechanisms from [17] and [18]. The main enhancement is that the *task subset* and *context-interaction requirements* sub-activities of R4IE [5], and the *monitoring mechanism specification* related activities of PC-RE [4] have been moved to the context-aware specialised stage, F_2. Also, a new sub-activity has been proposed, to analyse the activity of stakeholders in order to prepare the situation of interest identification in F_2.

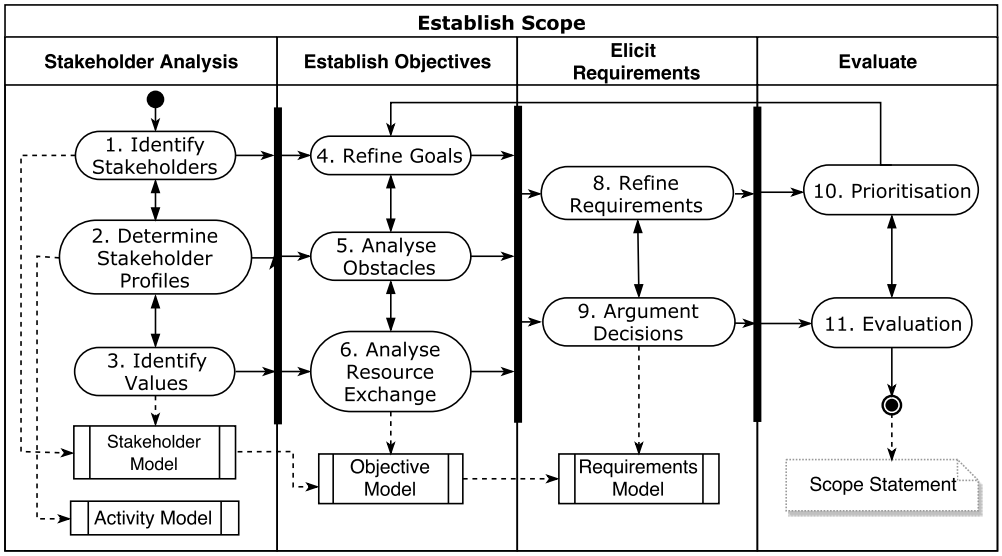


Fig. 1. Activity model representing the core activities and corresponding sub-activities in the early requirements elicitation stage, corresponding to F1.

1) Cultural analysis: The first sub-activity of stakeholder profiling deals with the system from an international point of view, where the different effects of culture are analysed in order to influence the definition of requirements for localising systems and specifying how it will be tailored for its different cultural profiles. During this activity, scenarios are sourced from users who belong to the cultures, nationalities and linguistic groups inside the intended market. The four[2] main steps of the guide proposed in [4] can be applied for this purpose.

2) Ethical analysis: An ethical analysis can contribute to ensure the ethical appropriateness of actions are ultimately taken in a project. For this purpose the use of Ethical Analysis Grids is recommended [17]. This grid can aid the satisfaction of both deontological (duty-based) and teleological (results-oriented) obligations. It consists of classifying some characteristics of each stakeholder into: High, Medium, Low and None. The characteristics are the vulnerability and gravity of the stakeholder, her/his dependency on the government, likelihood remedy, risk to fundamental value and policy impact. Although the ethical analysis proposed in [17] is useful for general purpose systems, it is not focused on C-AS. Context-awareness is the essence of different areas[3] that typically raise some ethical concerns which are different to those of traditional systems. For this reason, this sub-activity also adopts the eFRIEND ethical framework [18]. In order to apply it, it is recommended to carefully evaluate and discuss with the end-user stakeholders the different ethical concerns that might arise, until there is an agreement between all parties (*e.g.,* increasing user safety at the expense of giving up some privacy). The discussions can be complemented by questionnaires or interviews. The outcome of those discussions at a conceptual level can be used to modify or create different objectives and requirements.

3) Activity analysis: This stage consists of analysing the activity of end-user stakeholders, and is especially focused on that activity of end-user stakeholders. The purpose is to facilitate (for the benefit of developers) the identification of the meaning behind the behaviour of the end-user stakeholders. Particularly, by analysing how they usually behave in their quotidian tasks, and by thinking about how the stakeholders could use the proposed system to improve the way in which they achieve these tasks. This gives more opportunities to identify services that can be provided to them according to their particular needs, preferences, and limitations. Techniques such as observation, prototyping, scenarios or *wizard of oz* [20] can be used. Other approaches such as ethnomethodology can be adopted to understand the meaning of the actions of the end-user stakeholders. On the other hand, data analysis techniques such as classification or pattern-recognition could also help in revealing unexpected relations in the behaviour of the stakeholders.

4) Determine customisation, set-up, and training: The method proposed in Figure 1 is iterative. Once developers have defined some requirements, it is time to use the information gathered during this activity to customise existing requirements. The main dilemma is to specify C-AS that suit the requirements of individual users, while delivering a general system that can be used by many (individually different) users [4]. Not only this, but requirements can also evolve for the same user. For instance, as users become more experienced using the system, they require less help and supportive dialogues, and can access more sophisticated features. Also, the requirements engineering process should take into account aspects of maintenance and bespoke tailoring (to different stakeholders) after the system is deployed [5]. In

[2]Note that the fifth step has been moved to the user profiling activity in F_2.

[3]Particularly referring, in this case, to Ubiquitous & Pervasive Computing, Intelligent Environments, Ambient Intelligence and Ambient Assisted Living.

order to help the identification of different stakeholder profiles, there is a need to think about how the system will be set-up by/for the different stakeholders, trying to distinguish the different common needs of stakeholders that can be classified into profiles. As well as how the different stakeholders will want to customise the system, what type of training will they receive, and how will they receive it. In order to enable the customisation of the system, an individual user profile is defined first. The requirements are frequently set by another expert stakeholder (*e.g.,* a teacher sets certain requirements for a student's learning abilities). Individuals directly elicit (and own) personal goals. For both personal goals and user profiles, attainment targets can be set which become benchmarks for monitoring processes. A trade-off analysis might help to identify any conflicting user profile goals set by the expert stakeholders with the personal ambitions of the end-user stakeholder.

C. Identify values

During this sub-activity the aim is to identify the different values to be produced by the system and consumed by the stakeholders. This model has been adopted from the value model introduced in REUBI [3]. It takes existing stakeholders, humans or agents, and identifies the different values that are expected to be exchanged. Particularly, stakeholders which produce, consume a value, or are interested in a value or in its acquisition quality. Also, developers analyse what values are interchanged by the system and other stakeholders. Then developers reflect on how these values can be enhanced. Aspects such as how to improve the value, what is the expected quality of the value, what time restrictions exist in the provision of the value are taken into account. Other enhancement aspects apply, such as the flexibility in the value acquisition, precision or reliability restrictions, as well as cost or security restrictions applicable to the value.

D. Stakeholder diagram

Inspired by the techniques of the sub-activities explained in this section, the Stakeholder Diagram is introduced, which can model relevant stakeholder related information, as it is shown in the meta-model of Figure 2. In addition to the meta-models of this diagrams, each stakeholder can be associated to a *userProfile* stereotyped element. This user profile is composed of *profileFeature* stereotyped elements that can have *profileFeatureInstances*. This is better illustrated in the example from Section IX-A.

VI. Establish objectives

Taking into account the value analysis, the objectives of the system are declared. Then, from the higher order objectives, a refinement process is applied in order to obtain and decompose them into sub-objectives. This step is followed by an analysis of the adverse conditions that may impede the satisfaction of a goal. Following this, the analysis focuses on the resources required for the satisfaction of goals.

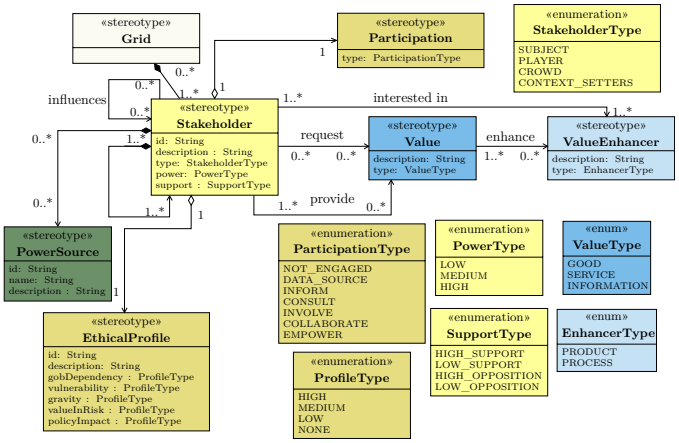


Fig. 2. UML meta-model for the Stakeholder Diagram.

A. Refine goals

Once the system boundaries and the higher-order objectives are identified, in the form of *values* and *value enhancers*, it is necessary to derive more specific objectives, and progressively refine them in order to obtain more knowledge about the system under development. Objectives act as a bridge between the system values and the final requirements of the system, providing specific guidance during the requirements elicitation process. Objectives are divided into *goals* and *soft-goals*, according to the identified *values* and *value enhancers*. Goals have a clear criteria of satisfaction. Soft-goals do not have a clear criteria of satisfaction, which means that they can be used for identifying and modelling non-functional requirements. The objectives can also be progressively decomposed using *inclusive* (AND), *alternative* (OR), or *exclusive* (XOR) relationships between them. Such relationships can help in determining if their corresponding parent objectives are satisfied or not, as further explained in [3].

B. Analyse obstacles

The dynamic nature of C-AS is closely related to the existence of multiple adverse conditions which can make it difficult for system objectives to be met. This sub-activity consists of identifying obstacles which may affect meeting a specific goal, in the same way as described in [3]. The main objective is to determine those situations which are likely to be inconvenient for meeting the objectives, even if obtaining a complete set of adverse conditions can be a difficult achievement.

C. Analyse resource exchange

Sometimes, there exist restrictions on the way in which sub-objectives need to be satisfied in order to satisfy the parent objective; such as not satisfying an objective until other objectives are satisfied, mainly because these require access to certain resources which are generated as a result of satisfying other objectives. The objectives relate to the resources through two different relationships: *provision* and *demand*, as further explained in [3].

D. Objective diagram

With the purpose of facilitating the objective, obstacle and resource exchange analysis explained in this section, the *Objective Diagram* is introduced, which has been adopted from the Interdependency Graph in [3]. The meta-model of this diagram and its corresponding example can be observed in [3].

VII. Elicit requirements

Once the objectives of the system are defined, they need to be operationalised into requirements. Then, an analysis of the contribution that the requirements have to objectives should be performed. This stage is inspired in the *task/function* and *system performance qualities* identification activities of R4IE [5]. Following this, requirements are refined, decomposed into sub-objectives, and related to other model elements. All the decisions taken need to be documented as rationales, in order to facilitate requirements tracing, by modelling the reasons which developers are following to make decisions.

A. Refine requirements

Once the engineers agree upon the representation of values, objectives, their decomposition, obstacles and resources; the next step is to discover alternatives which can satisfy the objectives, finding their possible operationalisations, in the form of requirements. In the previous sub-activity, higher-order requirements are identified, as well as their contribution to the objectives. In this sub-activity, those requirements are refined into more precise requirements, and are related to other elements of the system. In addition to those relationships (RefineObj and Contribute) introduced by the Interdependency Graph in REUBI [3], the Requirements Diagram inherits five different types of relationships from SysML (Derive, Refine, Satisfy, Verify and Copy) [7]. This sub-activity mainly consists of applying these seven relationships between requirements, objectives and other elements of the system. More information on the application of these relations can be found in [3], [7].

B. Argument decisions

Once the requirements are operationalised and refined, the aim is to model the decisions taken during the previous activities. The Requirements Diagram enables this through the use of the Argumentation stereotype, which is related to other elements in the Requirements Diagram. SysML already provides a means to argument relationships via the Rationale stereotype. Nevertheless, the Argumentation stereotype provided in the Interdependency Diagram of REUBI [3], facilitates the specialisation of the rationale into support and rejection arguments.

C. Requirements diagram

For the purpose explained in this section, the Requirements Diagram is introduced, which inherits the stereotypes of the OMG SysML Requirements Diagram [7], and the object and justification meta-models from the REUBI Interdependency Graph [3]. The Operationalisation stereotype from REUBI, has been substituted by the SysML Requirement, which gives several advantages. The requirements traceability relationships help keep track of what happens to a requirement during system modelling and specification by identifying sources, destinations and links between requirements and models. Additionally, the SysML requirements enable a mapping to evaluation constructs such as the test case, providing a way of documenting how the requirements will be tested, which can be used along with other UML-based standards such as the UML 2.0 Testing Profile [8], to facilitate the design and automation of test runs [21]. SysML also offers two requirements visualisation mechanisms to identify, prioritise and improve requirements traceability through requirements tables and requirements traceability matrixes. Although the exclusive use of Use-case diagrams might be limited for the requirements engineering process, the use of SysML requirements to complement them represents an advantage and improves standardisation [22]. SysML requirements can also be related to use-cases with the *refines* relationship. Finally, the approach also inherits some advantages from the use of REUBI objectives, as these are used to facilitate the discovery of requirements and non-functional requirements and act as a bridge between the stakeholder analysis and the requirements. Finally, it is also important to mention that the operationalisation of objectives into requirements can be evaluated using the evaluation procedures from REUBI [3].

VIII. Evaluate

The last activity of the framework consists of an evaluation of the objectives and requirements, which is guided by a set of heuristics, that have been adapted to their application to the framework presented in this paper from [19] [3]. Then, a plan for evaluating the requirements is created, setting the criteria for how each requirement will be evaluated once the system is implemented. For this, the objectives need to be prioritised, in order to enable developers to focus on the development efforts on the most important objectives first. Then, an evaluation helps engineers to determine if the current modelled operationalisation of objectives into Requirements satisfies the objectives. For this the evaluation procedure for the NFR framework is adopted, as presented in [19] and [3].

IX. Case study

The case study introduced in this section is based on the insights gained during the development of the EU funded POSEIDON[4] project [23]. The project name stands for PersOnalised Smart Environments to increase Inclusion of people with DOwn's syNdrome, and is particulary focused on using smart assistive technologies in order to foster the independence of people with this condition. The example presented in this work is constrained to an outdoors navigation application, which is bespoke to this particular disability [25]. More specifically, the case study focuses on a mobile application that uses a real-world representation of maps along with location

[4]http://www.poseidon-project.org/

services to support outdoor journeys that might be walking or by bus. Due to space restrictions, this example is further limited to bus displacements happening in London, United Kingdom. The application uses routes with tailored directions, notifications, reminders, and other services which will be triggered depending on the context. The navigation system described in this case study can be found in [26], and it has been developed using the open-source framework RCASE [10] developed as part of the contribution presented in this paper. All the figures appearing in the remaining of the paper are screenshots of the RCASE tool.

A. Stakeholder analysis

The stakeholder identification activity presents a set of techniques that build on the previous activity. The following list of stakeholders are identified: 1) Primary Users (PU), people with Down's Syndrome; 2) Secondary Users (SU), parents or carers of people with Down's Syndrome; 3) POSEIDON Managers, the management team of the POSEIDON project; 4) POSEIDON Development Partners, POSEIDON project partners which work in creating code or libraries that are to be reused by this application. 5) Developers, the developers of the navigational system; 6) Bus driver, the person(s) that drive(s) the bus in which the PU will get on; 7) Bus company, the company in charge of the bus line; 8) Calls and Internet provider, referring to the company that provides phone calls, SMS and internet to the mobile device; 9) Device Manufacturer, company that manufactures the device; 10) Operating System Developers, group involved in the development of the operating system of the device; 11) Maps Library Developers, group involved in the development of the maps libraries. The list is further refined into the power versus interest grid, which evolves through iterations into the stakeholder influence grid, as it is shown in Figure 3. The stakeholder

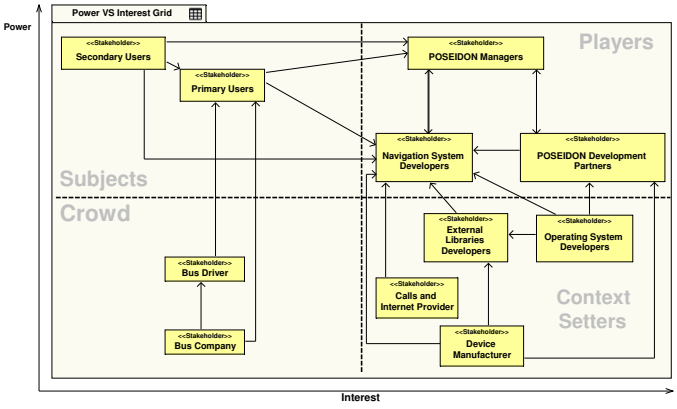


Fig. 3. Power vs Interest Grid representation, created with the Stakeholder Diagram from the RC-ASE Tool.

diagrams introduced give better insights about who are the stakeholders of the system and their relevant aspects to the project. The analysis on the stakeholders and their profiles can provide relevant information of the stakeholders which can be later reused for identifying their needs and preferences in the context related requirements. The stakeholder profiling activity follows. The POSEIDON project, involved a total of three different countries, namely, United Kingdom, Germany and Norway. These three cultures are similar in the sense of avoiding uncertainty, having similar work patterns, and responding similarly to authority, initiative and responsibility, since they all share the same continent. As expecting users with possibly low-skills with technology [27], the use of visual or symbolic representations of context is preferred, as well as the one task at a time approach. Another relevant aspect to take into account is the language difference between these three countries. Additionally, the United Kingdom has a different currency, representation of metrics, and driving direction than Germany and Norway. This might affect the payments of users for public transport, the location of bus stops, as well as the distance representation in the maps. The discovery of personas revealed that some of the particular users have visual or auditive impairments, and the questionnaires revealed that different skill levels using information technologies [27]. There are five different user profile features for the primary user stakeholder: Culture, visual impairment, skills with technology, independence degree, and auditive impairment. Each of the profile features is divided into its corresponding user profile feature instances. For example, the independence degree can be classified into three profile feature instances: *independent*, *moderately dependent*, and *dependent*. Following, the activity of the end-user stakeholders when displacing is analysed. A user will typically walk to the bus station, wait for a bus, take the bus, press the stop button one destination before the stop, get off the bus and walk again if necessary. This information will be used to identify situational interests in F_2, as it is shown in Table 1 from [2].

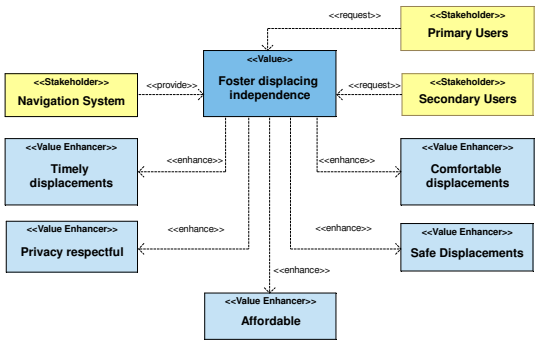


Fig. 4. Value model representation, Stakeholder Diagram, RC-ASE Tool.

The last activity in the stakeholder analysis consists of creating a value model, as shown in Figure 4. The first actor to take place is the navigational system itself, which offers the value of fostering the independence of both primary and secondary users. That value is a service, which is requested by the primary and secondary users. Five main aspects enhance the value provided by the navigation system. These are: that primary and secondary users can afford the system, that the system can preserve the privacy of the users, that the primary users can displace safely when using the system, that the

primary users can reach their destination on time, and that the instructions given by the navigation system are understandable by primary users.

B. Establish objectives

The main goals and soft-goals of the system are derived from the value model shown in Figure 4. In this way, the goal *Guide displacements*, is related to the *Foster displacing independence* value, as shown in Figure 5. Since this value is still too generic, it needs to be refined. The goal can be decomposed into two sub-goals: *Walking displacement guidance* and *Bus displacement guidance*. Note that for satisfying the high level objective, both lower level objectives must be satisfied. It equally happens with the value enhancers. *Walking displacement guidance* is refined into the objective "*Time-based guidance*", which proposes that the guidance received by the stakeholders will take into consideration time constraints. This goal is divided into another two lower level goals, which are to *provide guidance about when to start the displacement*, and to *provide guidance according to the waking speed*. The value enhancer *Affordable*, is distilled into the *Low-cost* soft-goal, which at the same time is divided into *Low-cost hardware* and *Low-cost software* soft-goals. The value enhancer *Privacy respectful* is also refined into the soft-goal *User privacy*, that is divided into the two soft-goals *Anonymity/pseudonimity* and *User intimacy*. The value enhancer *Safe displacement* is refined into the soft-goals *Displace through safe environments*, and *Support lost users*. Finally, the value enhancer *Comfortable displacement*, is distilled into the goal *Guide on required objects*, that supports the user with a list of objects that can make more comfortable the displacement or the activity to do where the user is displacing. Also, this value enhancer is refined into the soft-goal *Provide understandable guidance*.

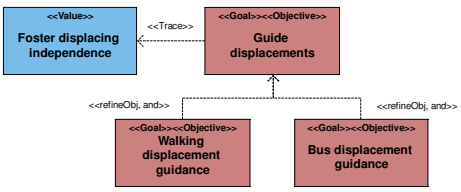


Fig. 5. Goal Decomposition I, Objective Diagram, RC-ASE Tool.

Following, an obstacle analysis over the objectives proceeds. The main obstacle found is due to the interruption of the service, caused by a lack of power. The battery may run off, and the user is left without instructions to follow. In order to mitigate, the soft-goal *Availability* is added. Next, is the Resource analysis. Here, the goal of guiding displacements refines from the *Start instructions* resource, generated from the time-based guidance goal. Additionally, the goal for guiding displacements also requires from the *Personal object list* resource, generated from the *Guidance on object list* goal.

C. Elicit requirements

At this stage of the method, the different goals of the system are refined into requirements, that represent a condition or capability that the system needs, and which contribute to the satisfaction of objectives. These design decisions, as well as the positive or negative contributions of the decisions are studied. For this, the lowest-level goals are considered (*i.e.,* those goals which do not have any sub-goal). In the previously introduced goal models, there are 5 low-level goals, which are used to define the functional requirements of the system, and 7 low-level soft-goals, which are used to define the non-functional requirements. For simplicity, the Requirements Diagrams of this example have been divided into three parts: requirements related to navigation, as shown in Figure 6; requirements related to reminders of the system; and non-functional requirements. Navigational requirements are based on a main requirement, *Navigation Map*, that specifies that the user will be able to observe a map that represents the real-world surroundings. Note that this requirement is a positive contribution towards two low-level goals: *Walking displacement guide* and *Bus displacement guide*. However, since just

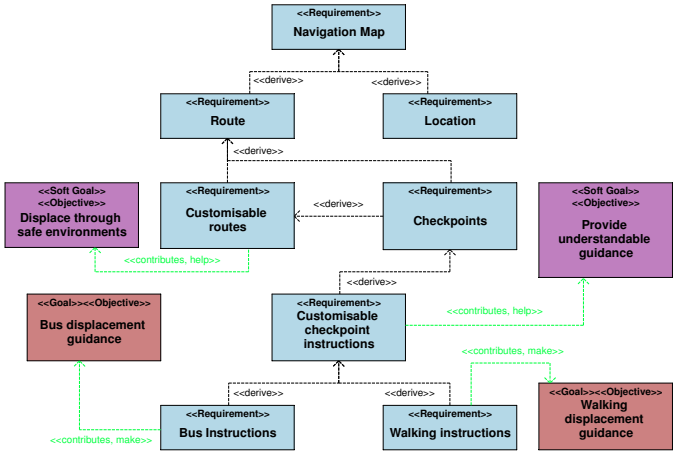


Fig. 6. Requirements model I, Requirements Diagram, RC-ASE Tool.

showing a map can not be considered as providing enough guidance, the contribution relationship can not be considered as a *Make* contribution. To keep the diagram simple, the *help* relationships between requirements and these two goals have been omitted in Figure 6. Since the *Navigation Map* requirement is not enough proof for satisfying the two previously mentioned goals, this main requirement is divided into another two additional requirements which are to show specific instructions on the next movements that users need to do in order to ultimately arrive at their destination. The navigation map will have a *route*, indicating the path that the user has to follow in order to arrive at her/his destination. Additionally, the navigation map will display the *location* of the user in the map in real-time. Although these two new requirements also provide a positive contribution towards the satisfaction of the two main guidance goals of this diagram, they are still not enough proof for providing adequate guidance to the users when walking and displacing by bus. Taking into account the low level objective of *Displacing through safe environments*, the requirement *Customisable routes* is included, where it is

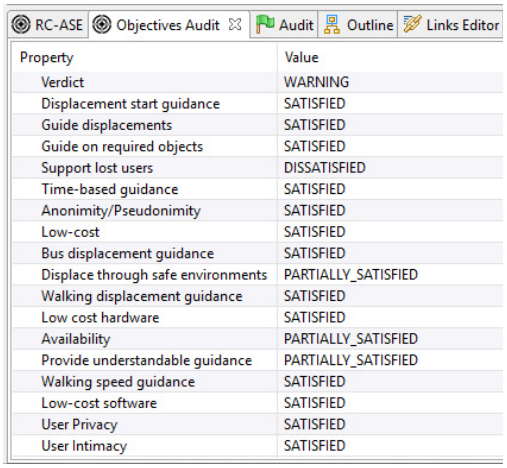
specified that the secondary users will be able to create their own routes for the primary users. The difference between the application under development and other navigation applications, is that this option increases the security of the primary users, as parents are expected to send them through safer and easier routes, instead of the most complicated ones. This requirement also satisfies the needs of users with different skill levels. As it can be observed in Figure 6, this new requirement is considered as a positive contribution towards the soft-goal for safe environments. Routes, will also have *checkpoints* that divide the route into more manageable smaller parts. Although this requirement by itself does not provide any contribution to the objectives, it is necessary to understand the next requirement that derives from it: *Customisable checkpoint instructions*. The checkpoints of the routes, will not only be located by the secondary users, but they will include a set of personalised instructions about the next movement. For example, it could be " *When you see the blue house with a white door, turn left, using the crosswalk*". An additional picture of the blue house can be included for making the instruction more clear. This requirement positively contributes to the soft-goal of *Provide understandable guidance*. These customisable checkpoint instructions can map to *walking* or for *bus displacements*. These last two instruction types *make* the main guidance goals. Therefore, the requirements engineers can consider this diagram as finished, and continue with the following diagram. Due to space reasons, reminder-related requirements and non-functional requirements have been omitted from this example. The operationalisation into requirements from objectives will occur similarly to that of navigational requirements.

After completing the operationalisation of objectives into requirements, the next step is to personalise or create new requirements according to the different user profiles. As it can be observed in the first figure, the cultural profile affects the existing communication with the users. Therefore, this figure illustrates the different requirements that are created to satisfy the demands of a British profile. The project supports English, German and Norwegian languages, British pounds (BGP) and Euros (EUR), as well as the Imperial and Metric systems. On the other hand, the second figure enables a different communication with the users. For those users with visual impairments, audio based communications will be present, and for those with auditive impairments visual communications will be enabled. The sub-activity for personalisation introduced in Section V-B4 it also includes a specification of the set-up and training. The users of the navigation application, will have to their disposition a training tool for letting them acquire navigation skills in a virtual environment, without exposing themselves to unnecessary risks. For space reasons, the further explanation on the training framework is out of the scope of this example, but the Reader is referred to [28] for more information about how users can train using this system.

Finally, an ethical analysis of the stakeholders is done. For example, an ethical analysis of the primary user stakeholder can be conducted against the *Login*, *Mobile platform*, *Reminders*, *Navigation map*, and *Communication with the users* requirements. For this profile, there is no *dependency* of the stakeholder on the government regarding the mentioned requirements. Also, there is a medium level of *vulnerability* from the users, in case they can get lost by misinterpreting indications. Nevertheless, the *gravity* of this stake is low. There is a high *likelihood* that there will be a remedy for this which will be addressed when creating the context-awareness specialisation of the requirements methodology. There is a medium *risk* to the integrity of the stakeholders, and the *policy impact* is high.

D. Evaluate

First of all, if it has not been done already, all the objectives defined in IX-B need to be prioritised. Then, the R-CASE module will automatically give a verdict on the satisfaction of the objectives of the system, according to the algorithm and rules explained in Section VIII. For this example, the result of this evaluation can be observed in Figure 7. The current verdict of the example is *WARNING*, as the *Support lost users* objective is *DISSATISFIED*, and the *Availability* objective is *PARTIALLY_SATISFIED*. This means that in order to improve the verdict to *PASS*, special attention should be payed to the completion of these objectives when eliciting requirements related to the context-awareness of the system.



Property	Value
Verdict	WARNING
Displacement start guidance	SATISFIED
Guide displacements	SATISFIED
Guide on required objects	SATISFIED
Support lost users	DISSATISFIED
Time-based guidance	SATISFIED
Anonimity/Pseudonimity	SATISFIED
Low-cost	SATISFIED
Bus displacement guidance	SATISFIED
Displace through safe environments	PARTIALLY_SATISFIED
Walking displacement guidance	SATISFIED
Low cost hardware	SATISFIED
Availability	PARTIALLY_SATISFIED
Provide understandable guidance	PARTIALLY_SATISFIED
Walking speed guidance	SATISFIED
Low-cost software	SATISFIED
User Privacy	SATISFIED
User Intimacy	SATISFIED

Fig. 7. Screenshot of the evaluation of objectives using the RC-ASE module in Modelio.

X. Conclusions and Future Work

This paper proposes a framework for facilitating the systematic treatment of requirements, and which is specialised for the non-contextual aspects of C-AS. The framework proposes a guide for developers that spans from the identification of stakeholders, to the identification of objectives and its operationalisation of goals, and introducing a UML/SysML profile for supporting the documentation and modelling of the process. The process is based on the strong points of different methodologies which are gathered as a coherent framework, helping to cover the gaps in the development of C-AS that current requirements elicitation methodologies have (Table I).

Additionally, the framework has been implemented as part of an open-source tool which supports the Diagrams introduced in this paper, as well as other SysML features to increase the traceability of elements throughout the models. A novel module for Modelio has been developed, namely Requirements for Context-Aware Systems Engineering (RC-ASE) [10], which implements not only the Diagrams introduced during this section, but also the missing SysML features that the free version has, including *traceability matrixes* and *requirements tables*, as well as other relevant functionality such as partial documentation generation. The approach has been applied to a navigation system of the POSEIDON project. Currently, there is undergoing work to create a more specialised UML/SysML profile that is more focused on the contextual aspects, related to F_2, as introduced in [2]. More work is being dedicated to the creation of another framework that facilitates the design and automatic code generation, aimed for the management of context information for context-aware rule-based reasoning support in both mobile [29] and stationary [30] platforms. The aim is not only to create services that can create C-AS that are more related to the preferences and needs of the users, but to create more reliable services by automating the verification of reasoning rules.

ACKNOWLEDGMENT

The research leading to these results has been partly supported by the POSEIDON project funded by the European Union (FP7/2007-2013) under grant agreement no. 610840.

REFERENCES

[1] U. Alegre-Ibarra, J. C. Augusto, and T. Clark, "Engineering context-aware systems and applications: A survey," *Journal of Systems and Software*, vol. 117, pp. 55–83, 2016. doi: 10.1016/j.jss.2016.02.010

[2] U. Alegre-Ibarra, J. C. Augusto, and C. Evans, "Perspectives on engineering more usable context-aware systems," *Journal of Ambient Intelligence and Humanized Computing*, 2018. doi: 10.1007/s12652-018-0863-7

[3] T. Ruiz-López, M. Noguera, M. J. Rodríguez, J. L. Garrido, and L. Chung, "Reubi: A requirements engineering method for ubiquitous systems," *Science of Computer Programming*, vol. 78, no. 10, pp. 1895–1911, 2013. doi: 10.1016/j.scico.2012.07.021

[4] A. Sutcliffe, S. Fickas, and M. M. Sohlberg, "Pc-re: a method for personal and contextual requirements engineering with some experience," *Requirements Engineering*, vol. 11, no. 3, pp. 157–173, 2006. doi: 10.1007/s00766-006-0030-0

[5] C. Evans, L. Brodie, and J. C. Augusto, "Requirements engineering for intelligent environments," in *Intelligent Environments (IE), 2014 International Conference on*. IEEE, 2014. doi: 10.1109/IE.2014.30 pp. 154–161.

[6] OMG, "OMG Universal Modeling Language (UML), Version 2.5," Object Management Group, Tech. Rep., 2015. [Online]. Available: http://www.omg.org/spec/UML/About-UML/

[7] ——, "OMG Systems Modeling Language (OMG SysML), Version 1.3," Object Management Group, Tech. Rep., 2012. [Online]. Available: http://www.omg.org/spec/SysML/1.3/

[8] ——, "UML 2.0 Testing Profile, Version 2.0," Object Management Group, Tech. Rep., 2017. [Online]. Available: http://www.omg.org/spec/UTP/

[9] R. Fuentes-Fernández, J. J. Gómez-Sanz, and J. Pavón, "Understanding the human context in requirements elicitation," *Requirements engineering*, vol. 15, no. 3, pp. 267–283, 2010. doi: 10.1007/s00766-009-0087-7

[10] U. Alegre-Ibarra, "Requirements for context-aware systems engineering (rcase) tool," https://github.com/ualegre/rcase, [Online; Last accessed 19-February-2018].

[11] K. Oyama, H. Jaygarl, J. Xia, C. K. Chang, A. Takeuchi, and H. Fujimoto, "Requirements analysis using feedback from context awareness systems," in *Computer Software and Applications, 2008. COMPSAC'08. 32nd Annual IEEE International*. IEEE, 2008. doi: 10.1109/COMPSAC.2008.239 pp. 625–630.

[12] S. Greenberg, "Context as a dynamic construct," *Human-Computer Interaction*, vol. 16, no. 2, pp. 257–268, 2001.

[13] A. Finkelstein and A. Savigni, "A framework for requirements engineering for context-aware services," in *In Proc. of 1 st International Workshop From Software Requirements to Architectures (STRAW)*, 2001. doi: 10.11648/j.ajsea.20150406.11 pp. 200–1.

[14] W. Sitou and B. Spanfelner, "Towards requirements engineering for context adaptive systems," in *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, vol. 2. IEEE, 2007. doi: 10.1109/COMPSAC.2007.223 pp. 593–600.

[15] N. Seyff, F. Graf, P. Grünbacher, and N. Maiden, "Mobile discovery of requirements for context-aware systems," in *Requirements Engineering: Foundation for Software Quality*. Springer, 2008. doi: 10.1007/978-3-540-69062-7_18 pp. 183–197.

[16] L. Baresi, L. Pasquale, and P. Spoletini, "Fuzzy goals for requirements-driven adaptation," in *International Requirements Engineering Conference (RE), 2010*. IEEE, 2010. doi: 10.1109/RE.2010.25 pp. 125–134.

[17] J. M. Bryson, "What to do when stakeholders matter: stakeholder identification and analysis techniques," *Public management review*, vol. 6, no. 1, pp. 21–53, 2004. doi: 10.1080/14719030410001675722

[18] S. Jones, S. Hara, and J. Augusto, "e-friend: an ethical framework for intelligent environment development," in *Ethics and Information Technology*, vol. 17. Springer, 2015. doi: 10.1186/2192-1962-3-12 pp. 11–25.

[19] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, *Non-functional requirements in software engineering*. Springer Science & Business Media, 2012, vol. 5.

[20] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through wizard of oz," in *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. ACM, 1993. doi: 10.1145/169059.169215 pp. 277–284.

[21] M. Hause, A. Stuart, D. Richards, and J. Holt, "Testing safety critical systems with sysml/uml," in *Engineering of Complex Computer Systems (ICECCS), 2010 15th IEEE International Conference on*. IEEE, 2010. doi: 10.1109/ICECCS.2010.59 pp. 325–330.

[22] M. dos Santos Soares and J. Vrancken, "Requirements specification and modeling through sysml," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, 2007. doi: 10.1109/ICSMC.2007.4413936 pp. 1735–1740.

[23] J. C. Augusto, T. Grimstad, R. Wichert, E. Schulze, A. Braun, G. M. Rødevand, and V. Ridley, "Personalized smart environments to increase inclusion of people with down's syndrome," in *International Joint Conference on Ambient Intelligence*. Springer, 2013. doi: 10.1007/978-3-319-03647-2_16 pp. 223–228.

[24] "Official website of the poseidon project," http://www.poseidon-project.org/, [Online; Last accessed 19-February-2018].

[25] D. Kramer, A. Covaci, and J. C. Augusto, "Developing navigational services for people with down's syndrome," in *Intelligent Environments (IE), 2015 International Conference on*. IEEE, 2015. doi: 10.1109/IE.2015.26 pp. 128–131.

[26] D. Kramer and Tellu, "Poseidon application," https://play.google.com/store/apps/details?id=no.tellu.poseidon, [Online; Last accessed 19-February-2018].

[27] POSEIDON Project, "Poseidon deliverable 2.1 - report on requirements," PersOnalized Smart Environments to increase Inclusion of people with DOwns syNdrome, Tech. Rep., 2015. [Online]. Available: http://www.poseidon-project.org/wp-content/uploads/2015/12/D2.1-Report-on-requirements-revised-after-pilot-without-interviews.pdf

[28] A. Covaci, D. Kramer, J. C. Augusto, S. Rus, and A. Braun, "Assessing real world imagery in virtual environments for people with cognitive disabilities," in *Intelligent Environments (IE), 2015 International Conference on*. IEEE, 2015. doi: 10.1109/IE.2015.14 pp. 41–48.

[29] D. Kramer and J. C. Augusto, "Supporting context-aware engineering based on stream reasoning," in *International and Interdisciplinary Conference on Modeling and Using Context*. Springer, 2017. doi: 10.1007/978-3-540-88479-8_37 pp. 440–453.

[30] U. Alegre-Ibarra, J. C. Augusto, and A. Aztiria, "Temporal reasoning for intuitive specification of context-awareness," in *Intelligent Environments (IE), 2014 International Conference on*. IEEE, 2014. doi: 10.1109/IE.2014.44 pp. 234–241.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 839–844
DOI: 10.15439/2018F170
ISSN 2300-5963 ACSIS, Vol. 15

Performance evaluation of trading strategies in multi-agent systems – Case of A-Trader

Jerzy Korczak*, Marcin Hernes**
*International University of Logistics and Transport, Wrocław, Poland
**Wrocław University of Economics, Poland
e-mail: {jerzy.korczak, marcin.hernes}ue.wroc.pl

***Abstract*—The article presents the problem related to evaluation of Forex trading strategies in multi-agent systems. The ratios based on financial measures cannot be assumed to be only evaluation criteria because other aspects determining effectiveness of the strategies, such as, for instance, investment risk, statistics on winning, and lost transactions, transaction costs, should also be taken into consideration. The aim of this paper is to review the general financial investments performance measures in relation to the performance analysis of trading strategies. The characteristics of the commonly used performance measures are outlined. The discussion will be illustrated by solutions developed in the trading support system, called A-Trader system. The performance analysis in A-Trader is detailed on real FOREX quotations.**

I. Introduction

TODAY the multi-agents systems are increasingly being used as trading support on Forex market [1, 2]. Agents operating in such systems provide strategies for open/close long/short positions with the use of various investments methods and techniques. There arises the need of constant evaluation of the agent performance to guarantee satisfactory benefits to the trader. Many of the performance evaluation methods [e.g. 2, 3, 4] and practical solutions (e.g. MetaTrader [5], Plus500 [6], MetaStock [7] or NinjaTrader [8]) are based on ratios of return (e.g. rate of return, gross profit, and the number of unprofitable transactions).

Investment decisions are made under conditions of risk and uncertainty. In the case of a trading decision, it is difficult to talk about the optimal decision as the decision bringing the highest possible rate of return for the investor, rather should be said about the decision bringing a satisfactory rate of return for the trader under a given level of other factors (e.g. risk). Many authors [e.g. 9, 10] have drawn attention to the fact that making an optimal decision is in practice very difficult in a situation of risk and uncertainty. If, on the other hand, we adopt the principle of satisfactory benefits, referred to in the literature as the principle of subjective expected utility [11], decision-making process becomes less complicated. According to this principle, for each alternative, its expected usefulness can be determined, and then the alternative that has the most usability is to be selected. The idea is to "set the bar" at such a level that it would not be too low, because then the result would be unsatisfactory, and not too high, because it might be unattainable for a trader. Note that in this case, the trader can set the "bar" higher and higher in sequence, which will bring him closer to the optimal value. Therefore, the ratios based measures cannot be assumed as the only evaluation criterion because other aspects having influence on the effectiveness of the strategies, such as, for instance, investment risk [12], statistics on winning and lost transactions as well as transaction costs should also be taken into consideration.

The aim of this paper is to review the general financial investments performance measures in relation to the performance analysis of trading strategies. The discussion will be illustrated by solutions developed and provided by the A-Trader system [13]. A-Trader system is composed of the agents capable of generating independent trading decisions on FOREX market. It allows for High Frequency Trading (HFT). It is realized in near real time (the notion of "near real time" can be understood as a very short delay between the last quote and the time of generated trading decision; usually 10-250 ms) and characterized by high speed, short-term positions, it concentrates attention on price formation process using sophisticated algorithms based on efficient and robust indicators and modern IT [14]. High frequency traders take decisions on the basis of real-time quotes changes in order to achieve satisfactory rate of return.

In the first part of this paper, the characteristics of performance measures are outlined. Next, the methods of performance analysis in the trading support system, called A-Trader, are detailed. In the final part of the article, conclusions and future works are presented.

II. Performance measures for financial investment

There are many performance measures in related works. These measures was developed in economics, management, and finance, both by researchers and practitioners. The related works [15, 16] divide performance measures for financial investments into three main groups:

- measures based on ratios of excess returns (ratios-based),
- measures based on systematic risk measured by factor models (risk-based),
- measures based on endogenous benchmarks derived from portfolio theory (benchmarks-based).

The major differences between these groups, but also between specific measures refer to the definition of risk. The next part of the section presents characteristics of particular groups.

A. *Measures based on ratios of excess returns*

Ratio-based performance measures specify the return per unit of risk. Ratio-based performance measures are usually easy to compute and have only low data requirements. These measures are of high relevance in practical applications and are frequently used in publications [12, 15, 17,18].

All ratios of return-based measures follow a similar schema: a measure of the return of asset in excess of the return on the benchmark is divided by a measure of the investment risk of asset. The most popular are the following::

- arithmetic rate of return,
- logarithmic rate of return,
- the number of transaction,
- gross profit,
- gross loss,
- total profit,
- the number of profitable transactions,
- the number of profitable transactions in a row,
- the number of unprofitable transactions in a row.

All of them are available in A-Trader.

B. *Measures based on systematic risk measured by factor models*

Risk-based performance measures adjust for risk by computing the spread between actual returns and a hypothetical benchmark return which is determined [17].

These measures indicate whether the trader was able to beat the benchmark, strictly speaking, they do not allow for comparison of different investment products because risk-based performance measures are subject to manipulation by leverage. For identification of relevant and meaningful risk factors and computation of "fair" or expected returns, risk-based performance measures draw heavily from the asset pricing literature. This group contains measures, such as [17, 19]:

- *Sharpe Ratio*

$$S = \frac{E(r)-E(f)}{|O(r)|} \cdot 100\% \quad (1)$$

where:
$E(r)$ – arithmetic average of the rate of return,
$E(f)$ – arithmetic average of the risk-free rate of return,
$O(r)$ – standard deviation of rates of return.

- *Treynor Ratio*

$$T = \frac{E(r)-E(f)}{\beta(r)} \quad (2)$$

where:
$E(r)$ – arithmetic average of the rate of return,
$E(f)$ – arithmetic average of the risk-free rate of return,
$\beta\,(r)$ – beta coefficient of rates of return.

- *The Kappa ratio*

$$T = \frac{E(r)-E(f)}{\sqrt[n]{LPM(r)}} \quad (3)$$

where:
$E(r)$ – arithmetic average of the rate of return,
$E(f)$ – arithmetic average of the risk-free rate of return,
$LPM(r)$ – lower partial moments of rates of return.

- *Omega ratio*

$$\Omega = \frac{\sqrt[n]{HPM(r)}}{\sqrt[n]{LPM(r)}} \quad (4)$$

where:
$HPM(r)$ – higher partial moments of rates of return.
$LPM(r)$ – lower partial moments of rates of return.

- *Average coefficient of variation*

$$V = \frac{s}{|E(r)|} * 100\% \quad (5)$$

where:
V – average coefficient of variation,
s – average deviation of the rates of return,
$E(\mathrm{r})$ – arithmetic average of the rates of return.

- *Jensen Model*

$$JM = E(r) - (E(f) + \beta\,(r) \cdot \big(E(m) - E(f)\big)). \quad (6)$$

where:
$E(r)$ – arithmetic average of the rate of return,
$E(f)$ – arithmetic average of the risk-free rate of return,
$E(m)$ – arithmetic average of the realized return of the appropriate market index,
$E(f)$ – arithmetic average of the risk-free rate of return,
$\beta\,(r)$ – beta coefficient of rates of return.

- *Value at Risk*

  The measure known as a value exposed to the risk - that is the maximum possible loss of the market value that a financial instrument can bear in a specific timeframe and at a given confidence level.

$$VAR = P * O * k \tag{7}$$

where:
P – the initial capital,
O – volatility - standard deviation of rates of return during the period ,
k – the inverse of the standard normal cumulative distribution (assumed confidence level 95%, the value of k is 1,65).

Some of these measures are available in A-Trader (Sharpe ratio, average coefficient of variation and Value et Risk).

C. *Measures based on endogenous benchmarks derived from portfolio information*

This group of measures usually compares the return of each security in the portfolio to the return of a "benchmark" security in order to determine abnormal performance. The comparable securities are selected based on characteristics, or they are derived from portfolio in another time period. Consequently, data requirements are higher for these models, while the statistical concepts are relatively simple [15, 16].

- *Characteristic-Based Models*

  Characteristic-Based Models are interpreted as portfolio-weighted sum of the differences in returns between the stocks and the benchmark portfolios, and can be calculated, for example by following equation:

$$CM_t = \sum_{j=1}^{m} w_{jt} \left(r_{jt} - r_{jt-1}^{b}\right) \tag{8}$$

where:
w_{jt}– weight of asset j at time t,
r_{jt} – corresponding excess return of asset j,
r_{jt-1}^{b} – the return on a benchmark portfolio that is matched to asset j measured in $t-1$.

- *Holdings-Based Models*

  These models define managerial skill as a co-variation between portfolio weights and returns of single stocks taking into consideration an omega ratio.

$$HM = \sum_{j=1}^{m} Cov\left(w_{jt,} r_{jt} | \Omega_t\right) \tag{9}$$

where:
w_{jt}– weight of asset j at time t,
r_{jt} – corresponding excess return of asset j at time t,
Ω_t – omega ratio at time t.

- *Trade-Based Models*

  These models define managerial skill as a co-variation between portfolio weights and returns of single stocks.

$$TM = \sum_{j=1}^{m} Cov\left(w_{jt,} r_{jt}\right) \tag{10}$$

where:
w_{jt}– weight of asset j at time t,
r_{jt} – corresponding excess return of asset j at time t,

Performance measures presented in this section allow for a wide range of evaluation of investment strategies on Forex. There are many other measures in the related works, however we have tried to select such as are more often used in practice and which can be implemented in High-Frequency trading systems due to low computational complexity.

In A-Trader, a characteristic based models are available (based on buy and hold and random walk benchmarks).

Next part of paper presents method for performance evaluation in A-Trader multi-agent system.

III. PERFORMANCE EVALUATION METHOD IN A-TRADER

In general, A-Trader system is composed of the agents capable of generating independent trading decisions on FOREX market. It should be noted that decisions can be consistent or contradictory, e.g. two independent agents may generate buy and sell decision at the same time [20, 21]. The trading opportunities are provided by consensual advice, generated by multiple software agents that use technical and fundamental analysis as well as behavioral sentiments [22]. Trading agents in the A-Trader form the investment strategies, which advise recommended open and closed positions for online FOREX traders. There are many strategies implemented as Supervisor Agents, such as:

- Basic Strategy,
- Consensus,
- Candle genetic algorithm,
- Kohonen network,
- Growing neural gas,
- Fundamental back propagation network,
- Evolutionary algorithm.
- Deep Learning

Supervisor Agent is the most important agent in A-Trader. Its goal is to generate profitable trading advice, on the basis of three groups of The Supervisor Agent coordinates functioning of the other agents (which form a given strategy) presented, and to provide the final advice to the trader. Its other task include resolving conflicts between agents [23]. The strategies are permanently evaluated by Supervisor Agents, and those with the highest evaluation value can be taken by default or chosen by the trader.

The performance analysis in A-Trader is carried out with the consideration of the following measures (ratios):

- rate of return (ratio x_1),
- number of transactions,
- gross profit (ratio x_2),
- gross loss (ratio x_3),

- total profit (ratio x_4),
- number of profitable transactions (ratio x_5),
- number of profitable consecutive transactions (ratio x_6),
- number of unprofitable consecutive transactions (ratio x_7),
- Sharpe ratio (ratio x_8)
- average coefficient of variation (ratio x_9)
- Value at Risk (ratio x_{10})
- the average rate of return per transaction (ratio x_{11}).

There are many ways of defining the performance evaluation function. For the purpose of comparison of the agents' performance, the following simple evaluation function has been proposed:

$$y = \left(a_1x_1 + a_2x_2 + a_{3(1-}x_3\right) + a_4x_4 + a_5x_5 + a_6x_6 + a_7(1 - x_7) + a_8x_8 + a_9(1 - x_9) + a_{10}(1 - x_{10}) + a_{11}x_{11} \quad (11)$$

where x_i denote the normalized values of particular performance measures from x_1 to x_{11}. It was adopted in the test that coefficients a_1 to a_{11}=1/11.

It should be mentioned that these coefficients may be modified with the use of, for instance, an evolution-based method, or they could be determined by the trader in accordance with their preferences (for instance the trader may determine whether they are interested in higher rate of return with accompanying higher risk level or lower risk level but accepting a lower rate of return).

The output of the function is a value in the range [0..1], and the agent's efficiency is directly proportional to the function value.

Figure 1 presents performance evaluation panel in A-Trader. The upper part of window presents information related to open/close positions generated by strategy in a given period. The profitable ones are marked on green and unprofitable ones are marked on red. The bottom part of window presents performance evaluation values related to selected positions (it is possible to mark all positions, or only selected positions).

Referring to the evaluation analysis related to particular measures performed in other systems (mentioned in section 1), as previously underlined, these systems only offer the functions calculating the rates of return based ratios. It should be noted that evaluation, in most cases, is performed "manually" by the trader. This work has many inconveniences. Due to its time consumption, the trader can use only selected measures of performance, and choice of these measures may be narrow. Also the trader acting under time pressure may select inadequate measures, and, in consequence, important financial losses may be generated. In addition, it is very difficult to have valid current knowledge on online trading (the trader's knowledge, in very turbulent market conditions, may be outdated or/and incomplete). These issues imply that systems operating in real time are very limited.

As has been mentioned, the evaluation function used in A-Trader enables the evaluation of performance of specific strategies. These operations are made automatically, in time close to real time, by the Supervisor Agent which may then

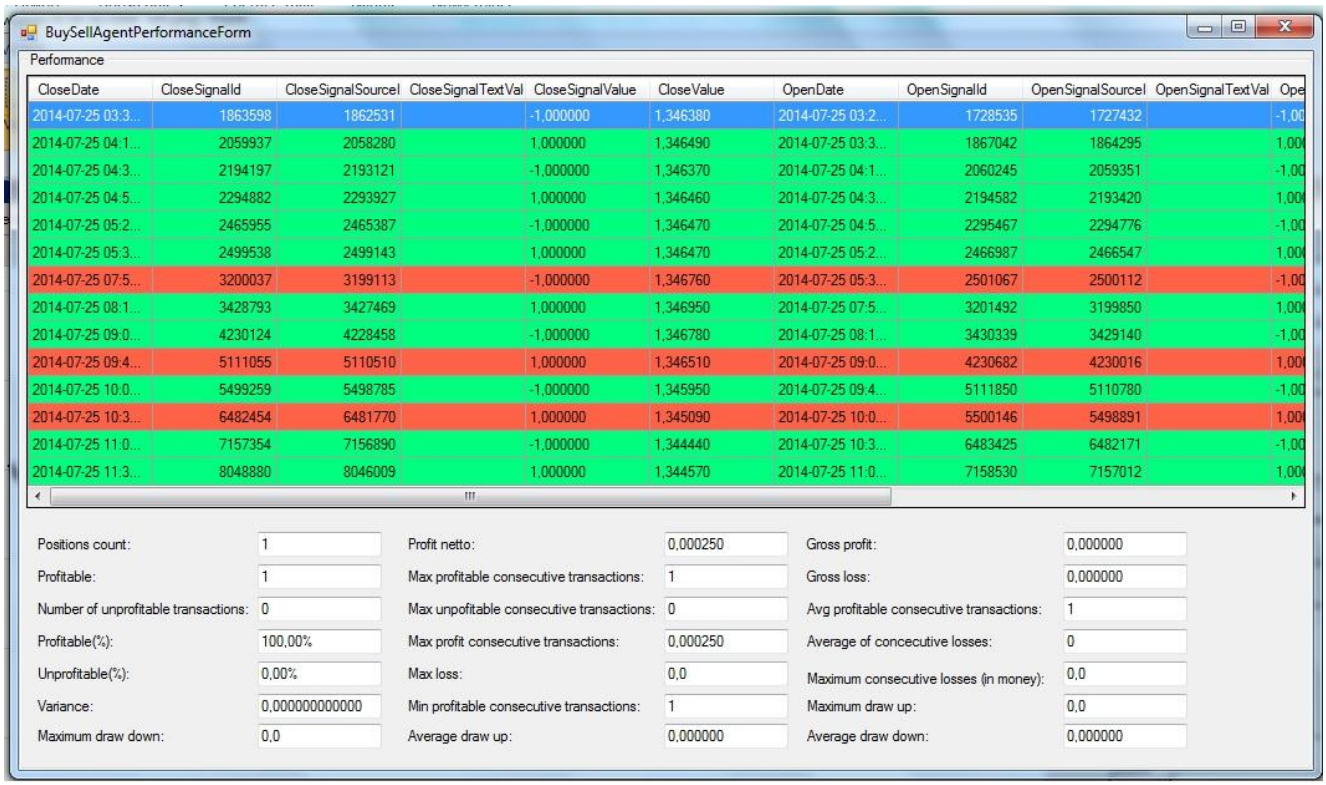


Fig. 1. Performance evaluation panel in A-Trader.
Source: Own work.

suggest to the investor taking final decisions on the basis of strategy with the highest level of performance. In addition, enabling the user to change coefficients a_i and ratios x_i parameters of the evaluation function allows for considering their preference concerning the criterion of importance of particular evaluation ratios. The performance evaluation function also considers the transaction costs with the assumption that this reflects the relationship between the number of transactions and the average rate of return from the transaction. However, this simple principle cannot be adopted because a large number of transactions has impact on the reduction of the strategy's efficiency level, especially for the transactions with a high rate of return.

A-Trader uses only selected measures because using largest number of ratios require higher computing resources than are now available (in HFT this measures must be calculated near real time). In future, we plan to use the cloud computing resources, then we will be in a position to implement a larger number of measures. It should be noted that the number of performance measures in A-Trader is not limited and they can be added to evaluation function in an easy way.

IV. Conclusion

The strategies in the A-Trader system open/close independent long/short position use multiple criteria of trading performance, which belong to three groups of performance measures: ratios-based, risk based, and benchmark-based. As a consequence, this enables the trader to compose an evaluation function according to their preferences and to apply to the strategies of the best Supervisor Agents. The results presented in our previous research [13, 14, 21, 22, 24, 25] allow us to come to the conclusions that there is no universally accepted evaluation function nor universal measures. The choice of measures and the composition of the evaluation function is highly dependent on trader preferences and trading market. We have already demonstrated that the level of performance of particular strategies changes depending on prevailing FOREX market situation. Based on the obtained results, there is no one strategy which definitely dominates over the others.

The use of this performance evaluation function allows for automatic setting of the best strategy in time close to real time, which has, in turn, a positive influence on investment effectiveness.

Future works should concern, among others, implementation of other performance evaluation measures (presented in section 2), development of an evolution method for determining ai coefficients into the A-Trader system, and implementation of cognitive agents performing analysis experts' opinions in the scope of forecasts referring to quotations on the FOREX market..

References

[1] M. Aloud, E.P.K. Tsang and R. Olsen, “Modelling the FX Market Traders' Behaviour: An Agent-based Approach”, [in] Alexandrova-Kabadjova B., S. Martinez-Jaramillo, A. L. Garcia-Almanza & E. Tsang (eds.), Simulation in Computational Finance and Economics: Tools and Emerging Applications, IGI Global, 2012, pp. 202-228. DOI: 10.4018/978-1-4666-2011-7.ch015.

[2] R.P. Barbosa and O. Belo, “Multi-Agent Forex Trading System”, [in] Agent and Multi-agent Technology for Internet and Enterprise Systems, Studies in Computational Intelligence Volume 289, 2010, pp. 91-118. https://doi.org/10.1007/978-3-642-13526-2_5.

[3] L. Mendes, P. Godinho and J. Dias, “Forex trading system based on a genetic algorithm”, J. J Heuristics 18(627), 2012, https://doi.org/10.1007/s10732-012-9201-y

[4] A. Shmilovici, Y. Kahiri, I. Ben-Gal, et al., “Measuring the Efficiency of the Intraday Forex Market with a Universal Data Compression Algorithm”, Computational Economics 33(131), 2009, https://doi.org/10.1007/s10614-008-9153-3

[5] MetaTrader5, https://www.metatrader5.com

[6] Plus500, https://www.plus500.com/

[7] MetaStock, https://www.metastock.com/

[8] NinjaTrader, https://www.ninjatraderbrokerage.com/get_started/forex

[9] F. Gul and W. Pesendorfer, “Hurwicz expected utility and subjective sources”, Journal of Economic Theory, V 159, Part A

[10] R. Pettigrew, “Risk, rationality and expected utility theory”, Canadian Journal of Philosophy, 45:5-6, 798-826, DOI: 10.1080/00455091.2015.111961, 2016.

[11] J. Shanteau and A. Pingenot, “Subjective expected utility theory”, [in] M. W. Kattan (Ed.), Encyclopedia of medical decision making (pp. 1084-1086). Thousand Oaks, CA: SAGE Publications Ltd., 2009, doi: 10.4135/9781412971980.n312.

[12] K. Jajuga and T. Jajuga, “Inwestycje: Instrumenty finansowe, ryzyko finansowe, inżynieria finansowa”, PWN, Warszawa 2000.

[13] J. Korczak, M. Hernes and M. Bac, “Risk avoiding strategy in multi-agent trading system”, [in] Proceedings of the 2013 Federated Conference on Computer Science and Information Systems / Ganzha Maria, Maciaszek Leszek, Paprzycki Marcin (red.), IEEE, 2013. DOI: 978-1-4673-4471-5/$25.00.

[14] J. Korczak, M. Hernes and M. Bac, “Collective Intelligence Supporting Trading Decisions on FOREX Market”, [in] Nguyen N., Papadopoulos G., Jędrzejowicz P., Trawiński B., Vossen G. (eds) Computational Collective Intelligence. ICCCI 2017. Lecture Notes in Computer Science, vol 10448. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-67074-4_12.

[15] P. Lückoff, “Mutual Fund Performance and Performance Persistence: The Impact of Fund Flows and Manager Changes”, Wiesbaden: Gabler Verlag 2011.

[16] Y. Yao and R. Zhang, “Empirical Research on Efficiency Measure of Financial Investment in Education Based on SE-DEA”. [in] Cao BY., Liu ZL., Zhong YB., Mi HH. (eds) Fuzzy Systems & Operations Research and Management. Advances in Intelligent Systems and Computing, vol 367. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-19105-8_35

[17] O.K. Hussain, T.S. Dillon, F.K. and Hussain, E.J. Chang, “Risk Assessment Phase: Financial Risk Assessment in Business Activities”, [in] Risk Assessment and Management in the Networked Economy. Studies in Computational Intelligence, vol 412. Springer, Berlin, Heidelberg, 2013. https://doi.org/10.1007/978-3-642-28690-2_6

[18] Z. Qiu, “Discussion of Investment Analysis Method in the New Round of the China Stock Bull Market”, [in] Li M., Zhang Q., Zhang J., Li Y. (eds) Proceedings of 2015 2nd International Conference on Industrial Economics System and Industrial Security Engineering. Springer, Singapore, 2016. https://doi.org/10.1007/978-981-287-655-3_40.

[19] D. Hu, , G. Schwabe and X. Li, “Systemic risk management and investment analysis with financial network analytics: research opportunities and challenges”, Financial Innovation (2015) 1: 2. https://doi.org/10.1186/s40854-015-0001-x.

[20] J. Korczak, M. Bac, K. Drelczuk and A. Fafuła, “A-Trader - Consulting Agent Platform for Stock Ex-change Gamblers”, [in] Proc. FedCSIS, Wrocław, pp.963-968, 2012.. DOI: 978-83-60810-51-4/$25.00.

[21] J. Korczak and M. Hernes, "Deep Learning for Financial Time Series Forecasting in A-Trader System", [in] Proceedings of the 2017 Federated Conference on Computer Science and Information Systems / Ganzha Maria, Maciaszek Leszek, Paprzycki Marcin (eds.), Annals of Computer Science and Information Systems, vol. 11, Polskie Towarzystwo Informatyczne, Institute of Electrical and Electronics Engineers, 2017. DOI: 10.15439/2017F449.

[22] M. Hernes and J. Sobieska-Karpińska , "Application of the consensus method in a multi-agent financial decision support system", *Information Systems and e-Business Management* 14 (1), Springer Berlin Heidelberg, 2016, DOI: 10.1007/s10257-015-0280-9.

[23] J. Korczak, M. Hernes and M. Bac, "Fuzzy Logic as Agents' Knowledge Representation in A-Trader System", [in] Information Technology for Management. Federated Conference on Computer Science and Information Systems, ISM 2015 and AITM 2015, Lodz, Poland, Revised Selected Papers / Ziemba Ewa (red.), Lecture Notes in Business Information Processing, vol. 243, 2016, Springer, ISBN 978-3-319-30527-1, ss. 109-124, 2015.

[24] J. Korczak, M. Hernes and M. Bac, "Risk avoiding strategy in multi-agent trading system", [in] Proceedings of the 2013 Federated Conference on Computer Science and Information Systems / Ganzha Maria, Maciaszek Leszek, Paprzycki Marcin (red.), IEEE, 2013.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 341–348
DOI: 10.15439/2018F325
ISSN 2300-5963 ACSIS, Vol. 15

Mining e-mail message sequences from log data

Paweł Weichbroth
Faculty of Management, WSB University in Gdansk,
Grunwaldzka 238A, 80-266 Gdansk, Poland
pawel.weichbroth@hotmail.com

***Abstract*—Communication by electronic mail (e-mail), once extravagant, is now the usual way to exchange data and information. Widely accepted by Internet users, business and governments, it is claimed to be the key part of the e-revolution. E-mail systems have been successfully implemented in almost all computer-aided domains of human interest, providing efficient, effective and permanent mechanisms of transmission. However, to date, the capability to exhibit an ordered list (sequence) of e-mail message senders and recipients, with the respective duration time between receiving and answering is still lacking. To fill this gap, in this paper we introduce the SOMF algorithm for mining such sequences from server log data. We specified a three-stage approach to comprehensively target the problem. The first stage concerns a data preparation task in order to assemble the input for the algorithm. The second, known as data mining, is the automatic analysis of data input performed in an unsupervised model by the SOMF algorithm. The third embraces output (knowledge) visualization, interpretation and evaluation. The given case study is based on the log data from an operational STMP server. By design, this simplified example brings about a better understanding of the solution, indicating one of its potential applications to identify and eliminate deadlocks in the realization of business processes. We also tested the efficiency of the implementation of the algorithm in five independent experiments on seven datasets, ranging in size. The results show that mining even 1 million rows is performed in approximately less than 6 minutes.**

I. Introduction

IN THE last decade, we have observed that the approach to information system design is evidently shifting from data to processes. Effective identification, construction, evaluation and deployment of mission-critical processes can give a competitive and strategic advantage to an organization [1]. Some argue that knowledge is still the most important asset to influence overall performance and the ability to innovate [2], [3], [4].

Knowledge workers utilize information technologies and their productivity depends on particular computer applications. The McKinsey Global Institute (MGI), through interviews with 4200 managers from companies in different businesses, in July 2012, reported that 28 percent of total work includes time reading, writing or responding to e-mails [5]. In 2015, the number of business e-mails sent and received per business user per day totalled 122 e-mails, and by the end of 2019 is expected to average 126 messages [6]. Today, there is rising concern that for some workers the volume of e-mail has grown to the size which in turn has negative effects on well-being and performance [7]. The perception of an individual being unable to find, organize or process his/her e-mails effectively is defined as the feeling of e-mail overload [8].

Knowledge management has long been a valuable part of business process architectures and models of various complex and collaborative domains [9]. A business process can be defined as a sequence of activities located and bounded in the frame of a particular organization [10], describing steps and corresponding tasks assigned and performed by particular participants (humans or other physical beings), the objective of which is to achieve a desired result [11]. Speaking from the margins, Adam Smith's theory of labour division is still up-to-date.

Business process design presents assumptions and beliefs in compliance with specified goals [12]. However, in real-life scenarios, an empirical business process can suffer from the following burdens: (1) some participants may not respect assumptions, their roles or tasks, and in consequence, act in a different way; on the other hand, even if they do, (2) some of them may delay performing assigned tasks, intentionally or not (for instance due to e-mail overload). Thus, one can ask for an objective method that provides data-driven evidence that shows how the process is empirically achieved.

In this paper, we investigate one particular activity which concerns e-mail correspondence between participants, and the primary focus is to introduce an algorithm, namely SOMF, for discovering e-mail message sequences from server log data. In this narrow extent, the obtained knowledge can be used for conformance checking, which aims to detect inconsistencies between the model of processes and their corresponding execution. In particular, it can be a way to detect communicated burdens on the one hand, and act as a starting point to discuss possible improvements on the other.

We devote our contribution in the field of data mining algorithms, to extracting sequences from data. From a broader perspective, the elaborated approach can be seen as an autonomous component of competitive intelligence [13], which, being a strategic tool producing actionable intelligence, in turn supports organizations in the decision-making process [14], improves their performance [15], and eventually fosters a competitive advantage [16].

The rest of the paper is organized as follows. The next section provides the problem statement for mining sequences from data. In Section III, the knowledge discovery process is outlined and specified in the frame of the research agenda, and divided into three subsections. In the next section, the process is exemplified by the case study. The last section closes the paper by presenting and discussing the obtained results from testing the efficiency of the algorithm.

II. The Problem Statement

As stated in the previous section, this research study introduces the concept of mining e-mail message sequences from server log data. The problem can be epitomised by three main aspects: (1) data preparation, (2) data mining, and (3) knowledge visualization, interpretation and evaluation.

This study mainly focuses on solving the first and second, by explicitly formulating algorithms devoted to each one. To achieve this goal, firstly, we formulate and provide all the relevant definitions.

Definition 1. A dataset $D = \{t_1, t_2, \ldots, t_n\}$ is a set of transactions, where each transaction is described by four attributes: message-id, time, sender and recipient.

Definition 2. A message m_i is an abstract term that may represent a document or e-mail uniquely identified by the message-id.

Definition 3. The time of the transaction tt_i is the recorded execution time of the performed action by the sender.

Definition 4. A set $P = \{p_1, p_2, \ldots, p_m\}$ is the finite collection of participants, which can be both message senders or recipients.

Definition 5. An event e is a pair $(x \rightarrow y)$ of the sender x and the participant y, where $x \neq y$.

Definition 6. A weight w_i is the difference between execution times tt_i and tt_{i-1} of two subsequent transactions, where $w_1 = 0$, and for $i = 1, 2, \ldots, n$.

Definition 7. A sequence $s \leq (e_1) : w_1, (e_2) : w_2, \ldots, (e_k) : w_k)$ of events is an ordered list of nonempty events; for each integer $k = 1, 2, \ldots, n$; a sequence of the length k is called a k-sequence.

Definition 8. A directed, weighted and labelled graph G is a tuple (V, E, w) consisting of a finite set V, together with a subset $E \subseteq VxVxR$. The elements of V are the vertices of the graph, and the elements of E are the arrows of the graph. An arrow of a graph is an ordered pair $[x, y]$, where x and y are the vertices of the graph, and w is the associated real number of the pair, called its weight, where $x \neq y$. The labels for vertices are a subset L of P.

To visualize knowledge, a user can use any application capable of processing data, which as a result, displays the adequate drawing. The remaining two tasks are usually associated with the specific context of the problem domain, and therefore should not be generalized.

III. The Knowledge Discovery Process

In our approach, the process of knowledge discovery is divided into three stages, followed one by one and independently performed ($1 \rightarrow 2 \rightarrow 3$), in a similar way to well-recognized and accepted models, such as KDD or CRISP-DM.

A. Data preparation

The data preparation stage concerns sorting objects. In the first step, a new message list is initiated. Next, a data set D is scanned, and the total number of rows n is determined. In the body of the loop $(4-6)$, where n is the termination condition, four attributes are selected from each row and form a new object, inserted into the message list. Next, a message list is grouped by the message-id, and sorted in ascending order by a time stamp. The corresponding pseudocode is given below.

Input: D

```
1 Initiate New List(Message-List);
2 Read(D);
3 for i := 1 to n
4  select MessageId, Time, Sender, Recipient from D;
5  create Object(Object-Message);
6  insert to List(Message-List);
7 end;
8 group List(Message-List) by MessageId and
9 sort ASC List(Message-List) by Time;
```

Output: a Message-List.

B. Data mining

The message list is now the input with no parameters for the SOMF algorithm. The pseudocode below shows the main idea lying behind its construction.

Input: List(Message-List)

```
1  while i < Count List(Message-List) do
2   CheckMessage:= read Object.MessageId[i];
3   MessageTime:= read Object.Time[i];
4   for each unique object from List(Message-List)
5    MessageTime:= read Object.Time[i];
6     add vertex(sender) to Sender List(Adjacency-List[0]);
7     add vertex(sender) to Sender List(Adjacency-List[1]);
8     weight:= (MessageTime[i] – MessageTime[1]);
9     add vertex(recipient) & weight to Recipient List(Adjacency-List[0]);
10  end;
11  for each object from List(Message-List)
12   if vertex(sender) exists in List(Adjacency-List[i]) then
13    weight:= (MessageTime[i] – MessageTime[1])
14    add vertex(recipient) & weight to Recipient-List(Adjacency-List[i+1]);
15   if vertex(sender) does not exist in List(Adjacency-List[i] then
16    add vertex(sender) to Sender-List(Adjacency-List[i+1]) and
17    weight:= (MessageTime[i] – MessageTime[1])
18    add vertex(recipient) & weight to Recipient List(Adjacency-List[i+1]);
19   if vertex(recipient) does not exist in List(Adjacency-List[i]) then
20    add vertex(recipient) to Sender-List(Adjacency-List[i+1]);
21   end;
22  if Object.MessageId[i] ≠ Object.MessageId[i+1] then
23     create Graph(MessageId);
24  i++
25 end.
```

Output: A set of graphs.

In the body of the first loop (2–9), for each unique object from the message list, the time of the first sent message is determined; next, a new vertex, representing the message sender, is created and added to the adjacency list on the left side; a new vertex, representing a message recipient, is created and added to the adjacency list on the left side; the weight between the top vertex and the vertices one level down equals zero.

In the body of the second loop (12–22), if the vertex of the sender exists in the adjacency list, then the weight is calculated, and a new vertex, representing a recipient, is created and added along with the weight to the adjacency list on the right side; if the vertex representing the sender does not exist, then a new vertex is created and added to the adjacency list on the left side; the weight is calculated and a new vertex for each recipient is created and added along with the weight to the adjacency list on the right side; if the vertex representing a recipient does not exist, then a new vertex for each recipient is added to the adjacency list on the left side. Finally, if the identifier of the next object is different, then a graph representing a unique e-mail message sequence is created.

We used an adjacency list as the graph representation. This was the most suitable form for us to use. Our goal is neither to prove its appropriateness or efficiency nor to investigate different representations. Having said that, however, if we take into account the results obtained from implementation feasibility and performance testing of the algorithm, the correctness of choice should not be a subject for long discussion.

C. Knowledge visualization, interpretation and evaluation

Knowledge visualization aims to use visual representation to facilitate the understanding of discovered complex data structures [17]. Knowledge interpretation is an arbitrary construct of the information perceived by an individual who aims to specify its meaning by incorporating context, logic and experience [18]. Knowledge evaluation is a subjective judgment on its applicability and validity to solve a particular problem [19]. These three tasks are wider discussed in the next section, having the input, i.e. visualized knowledge, already generated.

IV. Case Study

Let us consider a dataset $D = \{t_1, t_2, t_3, t_4, t_5\}$ of five transactions and a dataset $U = \{a, b, c, d, e\}$ of five users that are both senders and recipients, whose e-mail account names equal their names in the *gdansk.com* domain. Let $M = \{m_1\}$ be a set of one message m_1, represented by the unique identifier $m.1$. Each transaction is described by four attributes, i.e. *message-id, time stamp, sender* and *recipient*. The first transaction t_1 of the input for the data preparation stage is shown below.

```
Message-ID: <m.1@gdansk.com> Date: Fri, 19 Aug 2016
11:30:00 From: =?ISO-8859-2?Q?a?=<a@gdansk.com> To:
=?ISO-8859-2?Q?b?= <b@gdansk.com>
```

To simplify the log data, on the same day, in the one message domain, for the first user, it takes 5 minutes to pass over the message, and for each subsequent user, five minutes more. The message list is depicted in Table 1. Each transaction can be interpreted analogously to the first one: “at 11:30 a.m. the user a (sender) sends the e-mail to the user b (recipient)”.

In the second stage, the algorithm scans the message list, and for the first transaction t_1 determines the execution time of the message sent; next, the sender vertex a and recipient vertex b are added respectively to the first and second position on the left side of the adjacency list; only for t_1 the weight is not calculated and equals zero; the sender vertex b is added to the first position on the right side of the adjacency list and the weight (0) is assigned. The sender vertex b exists on the list; the weight is calculated and the vertex recipient c is added with the weight (5) assigned. The sender vertex c is added to the third position on the left side of the adjacency list; the weight is calculated, and vertices d and e are added on the right side with the weight (10) assigned. The sender vertex e is added to the fourth position on the left side; the weight is calculated and the recipient vertex a is added on the right side with the weight (15) assigned. The sender vertex a exists, the weight is calculated and the recipient vertex e is added as the second on the right side with the weight (20) assigned.

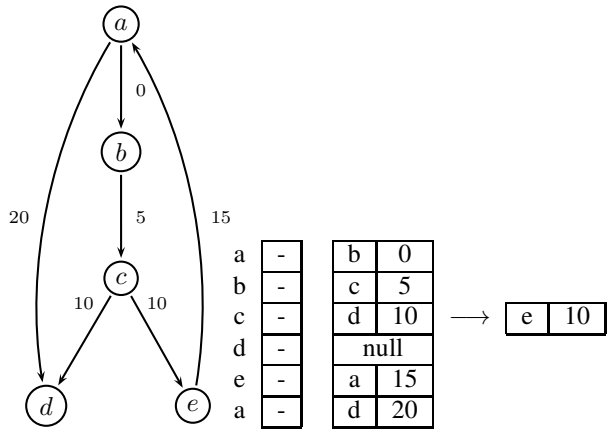


Figure 1. The graph G and the adjacency list

The discovered graph G can be interpreted literally as the following sequence of events:

- (e_1): at a given time, a message was sent from participant a to b;
- (e_2): five minutes later, a message was sent from participant b to c;
- (e_3): ten minutes later, a message was sent from participant c to d and e;
- (e_4): fifteen minutes later, a message was sent from participant e to a;
- (e_5): twenty minutes later, a message was sent from participant a to d;

TABLE I. THE MESSAGE LIST

transaction (t_i)	**time stamp** (hh:mm)	**sender** (name)	**recipient** (name)
t_1	11:30	a	b
t_2	11:35	b	c
t_3	11:45	c	d, e
t_4	12:00	e	a
t_5	12:20	a	d

TABLE II. THE ALGORITHM EFFICIENCY ACCORDING TO DATASET SIZE

D no. of rows	**file size** (KB)	**test1** (mm:ss.ms)	**test2** (mm:ss.ms)	**test3** (mm:ss.ms)	**test4** (mm:ss.ms)	**test5** (mm:ss.ms)
10	2	01.00	01.00	01.03	01.01	01.01
100	18	01.01	01.01	01.01	01.02	01.01
1 000	182	01.06	01.06	01.06	01.06	01.07
10 000	1827	01.59	01.59	01.59	01.59	01.61
100 000	18262	07.52	07.44	07.43	07.50	07.51
500 000	91309	38.96	46.19	40.30	38.99	39.49
1 000 000	182618	05:51.89	05:32.95	05:32.75	05:32.95	05:24.26

V. EFFICIENCY EVALUATION

The SOMF algorithm has been implemented in C#. Firstly, to verify its correctness, we prepared and used a dataset in such a manner that allowed us in advance to determine the output. Secondly, we implemented and executed a stand-alone script which randomly generated seven datasets, ranging in size from 10 to 1 million records. Finally, we separately performed five tests on each dataset, using a mobile computer equipped with an Intel Core I5 (3230M @ 2,6 GHz) processor, 4 GB (DDR3) RAM, and Microsoft Windows 8.1 (x64). The obtained results are summarized in Table 2.

If we take into account only the first four datasets, then the mining duration does not exceed 2 seconds and is comparably the same. Differences can be noticed in the last two datasets; however, the standard deviation is again relatively low, respectively 2,87 and 8,95 seconds. Now, if we consider the largest dataset, such a total number of rows cannot represent one hypothetical message sequence because it is simply too complex to be realized in any real-life scenario. Yet, conversely, in our opinion, the duration of less than 6 minutes is relatively low, if compared to other typical domains of the application of data mining algorithms. To sum up, the algorithm efficiency is at an acceptable level.

REFERENCES

[1] L. Mancilla-Amaya, C. Sanin, C., and E. Szczerbicki, "Using Human Behavior to Develop Knowledge-Based Virtual Organizations". *Cybernetics and Systems: An International Journal*, 41(8), pp. 577–591, 2010.

[2] M. Owoc, and K. Marciniak, "Knowledge management as foundation of smart university". Federated Conference on Computer Science and Information Systems. IEEE, pp. 1267–1272, 2013.

[3] M. Hernes, "Knowledge Integration Method for Supply Chain Management Module in a Cognitive Integrated Management Information System". In: International Conference on Computational Collective Intelligence, pp. 81–89. Springer, 2016.

[4] M. Pondel, and J. Korczak, "A view on the methodology of analysis and exploration of marketing data", 2017 Federated Conference on Computer Science and Information Systems. IEEE, pp. 1135–1143, 2017.

[5] M. Chui et al., "The social economy: Unlocking value and productivity through social technologies". McKinsey Global Institute, pp. 46, 2012.

[6] The Radicati Group. A Technology Market Research Firm. "Email Statistics Report 2015-2019", p.3. London (UK) 2015.

[7] K. Reinke, and T. Chamorro-Premuzic, "When email use gets out of control: Understanding the relationship between personality and email overload and their impact on burnout and work engagement". Computers in Human Behavior, 36, pp. 502–509, 2014.

[8] L. A. Dabbish, and R. E. Kraut, "Email overload at work: An analysis of factors associated with email strain". In Proceedings of the ACM conference on computer supported cooperative work (CSCW), pp. 431–440. New York, ACM Press 2006.

[9] P. Wang, C. Sanin, and E. Szczerbicki, "Prediction based on integration of decisional DNA and a feature selection algorithm RELIEF-F". *Cybernetics and Systems*, 44(2–3), pp. 173–183, 2013.

[10] A. Przybyłek, "The Integration of Functional Decomposition with UML Notation in Business Process Modelling". In: Advances in Information Systems Development, pp. 85–99. Springer 2007.

[11] B. Marcinkowski, and M. Kuciapski, "A business process modeling notation extension for risk handling". In: A. Cortesi, N. Chaki, K. Saeed, and S. Wierzchoń (Eds): Computer Information Systems and Industrial Management, pp. 374–381, Springer 2012.

[12] A. Przybyłek, "A Business-Oriented Approach to Requirements Elicitation". In: 9th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'14), Lisbon, Portugal, 2014.

[13] R. Pellissier, and T. E. Nenzhelele, "Towards a universal competitive intelligence process model". South African Journal of Information Management, 15(2), 1–7, 2013.

[14] D. Heppes, and A. Du Toit, "Level of maturity of the competitive intelligence function: Case study of a retail bank in South Africa". Aslib Proceedings: New Information Perspectives 61(1), 48–66, 2009.

[15] B. Huijbrechts, M. Velikova, S. Michels, and R. Scheepens, "Metis1: An integrated reference architecture for addressing uncertainty in decision-support systems". Procedia Computer Science, 44, 476–485, 2015.

[16] R. Brody, "Issues in defining competitive intelligence: An exploration", Journal of Competitive Intelligence and Management 4(3), 3–16, 2008.

[17] J. Korczak, H. Dudycz, and M. Dyczkowski, "Design of financial knowledge in dashboard for SME managers". In: Computer Science and Information Systems, pp. 1123–1130, IEEE 2013.

[18] M. Owoc, P. Weichbroth, and K. Żuralski, "Towards better understanding of context-aware knowledge transformation". In: Computer Science and Information Systems, pp. 1123–1126, IEEE 2017.

[19] M. L. Owoc, "Wartościowanie wiedzy w inteligentnych systemach wspomagających zarządzanie". Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Seria: Monografie i Opracowania. Wrocław 2004.

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution. Events that constitute SSD&A are:

- MDASD'18—5th Workshop on Model Driven Approaches in System Development
- MIDI'18- 6th Conference on Multimedia, Interaction, Design and Innovation
- LASD'18—2nd International Conference on Lean and Agile Software Development
- SEW-38 & IWCPS-5—Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

AREA SUPERVISORY COMMITTEE

- Hinchey, Mike, SEW-38-IWCPS-5
- Kornecki, Andrew J., MMAP'18
- Luković, Ivan, MDASD'18
- Marasek, Krzysztof, MIDI'18
- Przybyłek, Adam, LASD'18

5th Workshop on Model Driven Approaches in System Development

FOR many years, various approaches in system design and implementation differentiate between the specification of the system and its implementation on a particular platform. People in software industry have been using models for a precise description of systems at the appropriate abstraction level without unnecessary details. Model-Driven (MD) approaches to the system development increase the importance and power of models by shifting the focus from programming to modeling activities. Models may be used as primary artifacts in constructing software, which means that software components are generated from models. Software development tools need to automate as many as possible tasks of model construction and transformation requiring the smallest amount of human interaction.

A goal of the proposed workshop is to bring together people working on MD languages, techniques and tools, as well as Domain Specific Languages (DSL) and applying them in the requirements engineering, information system and application development, databases, and related areas, so that they can exchange their experience, create new ideas, evaluate and improve MD approaches and spread its use. The intention is to target an interdisciplinary nature of MD approaches in software engineering, as well as research topics expressed by but not limited to acronyms such as Model Driven Software Engineering (MDSE), Model Driven Software Development (MDSD), Domain Specific Modeling (DSM), and OMG's Model Driven Architecture (MDA).

1st Workshop on MDASD was organized in the scope of ADBIS 2010 Conference, held in Novi Sad, Serbia. From 2012, MDASD becomes a regular bi-annual FedCSIS event.

TOPICS

- MD Approaches in System Design and Implementation – Problems and Issues
- MD Approaches in Software Process Models
- MD Approaches in Databases and Information Systems
- MD Approaches in Software Quality and Standards
- Metamodeling, Modeling and Specification Languages
- Model Transformation Languages
- Model-to-Model, Model-to-Text, and Model-to-Code Transformations in Software Process
- Transformation Techniques and Tools
- Domain Specific Languages (DSL) and Domain Specific Modeling (DSM) in System Specification and Development
- Design of Metamodeling and Modeling Languages and Tools
- MD Approaches in Requirements Engineering and Business Process Modeling
- MD Approaches in System Reengineering and Reverse Engineering
- MD Approaches in HCI development
- MD Approaches in GIS development
- MD Approaches in Document Engineering
- Model Based Software Verification
- Theoretical and Mathematical Foundations of MD Approaches
- Organizational and Human Factors, Skills, and Qualifications for MD Approaches
- Teaching MD Approaches in Academic and Industrial Environments
- MD Applications and Industry Experience

EVENT CHAIRS

- **Luković, Ivan,** University of Novi Sad, Serbia

STEERING COMMITTEE

- **Gray, Jeff,** University of Alabama, United States
- **Mernik, Marjan,** University of Maribor, Slovenia
- **Ristić, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Tolvanen, Juha-Pekka,** MetaCase, Finland

PROGRAM COMMITTEE

- **Amaral, Vasco,** The New University of Lisbon, Portugal
- **Bryant, Barrett,** University of North Texas, United States
- **Budimac, Zoran,** Faculty of Sciences, Univ. of Novi Sad, Serbia
- **Chen, Haiming,** Chinese Academy of Sciences, China
- **Erradi, Mohammed,** ENSIAS, Mohammed-V University, Morocco
- **Fertalj, Krešimir,** University of Zagreb, Croatia
- **Gray, Jeff,** University of Alabama, United States
- **Härting, Ralf-Christian,** Hochschule Aalen, Germany
- **Ivanović, Mirjana,** University of Novi Sad, Serbia
- **Janousek, Jan,** Czech Technical University, Czech Republic
- **Karagiannis, Dimitris,** University of Vienna, Austria
- **Kardaş, Geylani,** Ege University International Computer Institute, Turkey
- **Kern, Heiko,** University of Leipzig, Germany
- **Kollár, Ján,** Technical University of Kosice, Slovakia
- **Kosar, Tomaž,** University of Maribor, Slovenia

- **Krdzavac, Nenad,** Michigan State University, United States
- **Liu, Shih-Hsi Alex,** California State University, United States
- **Maćoš, Dragan,** Beuth University of Applied Sciences, Germany
- **Melo de Sousa, Simão,** University of Beira Interior, Portugal
- **Mernik, Marjan,** University of Maribor, Slovenia
- **Milosavljević, Gordana,** University of Novi Sad, Serbia
- **Porubän, Jaroslav,** Technical University of Kosice, Slovakia
- **Rangel Henriques, Pedro,** Universidade do Minho, Portugal
- **Ristić, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Seidl, Martina,** Johannes Kepler University, Austria
- **Selic, Bran,** Malina Software Co., Canada
- **Sierra Rodríguez, José Luis,** Universidad Complutense de Madrid, Spain
- **Slivnik, Boštjan,** University of Ljubljana, Slovenia
- **Suvajdžin-Rakić, Zorica,** University of Novi Sad, Serbia
- **Tolvanen, Juha-Pekka,** MetaCase, Finland
- **Vangheluwe, Hans,** University of Antwerp, Belgium
- **Varanda Pereira, Maria João,** Instituto Politecnico de Braganca, Portugal
- **Wimmer, Manuel,** Vienna University of Technology, Austria

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 853–862
DOI: 10.15439/2018F218
ISSN 2300-5963 ACSIS, Vol. 15

Model-driven Query Generation for Elasticsearch

Berkay Akdal*†, Zehra Gül Çabuk Keskin*, Erdem Eser Ekinci*, Geylani Kardas†
*Galaksiya Information Technologies, Ege Technopark, 35100, Bornova, Izmir, Turkey
Email: {berkayakdal, zehragulcabuk, erdemeserekinci}@galaksiya.com
†International Computer Institute, Ege University, 35100, Bornova, Izmir, Turkey
Email: geylani.kardas@ege.edu.tr

***Abstract*—Elasticsearch is a distributed RESTful search engine, capable of solving growing number of use cases and can handle petabytes of data in seconds. However, Elasticsearch comes with a complex query language which causes a steep learning curve for the developers and, therefore, creation of queries can be difficult and time-consuming in many cases. Hence, in this paper, we introduce a Domain-specific Modeling Language (DSML), called Dimension Query Language (DQL), to support the model-driven development of Elasticsearch queries. Elasticsearch queries can be automatically generated from DQL models and DQL's IDE is capable of executing these auto-generated Elasticsearch queries on remote repositories. An evaluation of using DQL has been performed at the industrial level with the participation of a group of developers. The conducted evaluation showed that the use of the language significantly decreases the development time required for creating Elasticsearch queries. Finally, qualitative assessment, based on the developers' feedback, exposed how DQL facilitates the development of Elasticsearch queries.**

I. Introduction

ELASTICSEARCH is a distributed RESTful search engine, which is based on Lucene information retrieval software library [1] and is capable of solving growing number of use cases. Many types of searches (e.g. structured, unstructured, geo, metric) can be prepared and combined. It works in clusters, and according to some tests performed by its developers (namely, Elastic Team), it is reported that Elasticsearch can handle petabytes of data in seconds [2].

Elasticsearch differs from classical relational database management systems (RDBMS) in many ways: Elasticsearch's primary database model is a search engine and it stores documents instead of key-values. Each document in Elasticsearch is a JavaScript Object Notation (JSON) object, and hence it does not use Scripted Query Language (SQL). Queries are provided with its own language based on JSON. A given search can be performed not only in a form of a query; filters can also be used for document search which is faster than the queries. Finally, it is schema-free, i.e. two documents of the same type can have different sets of fields [3]

However, such kind of powerful engine comes with a very complex query language which causes a steep learning curve for the query developers. Moreover, there are numerous types of queries and scripts combinable with each other whose creation and use can be difficult and time consuming in many cases.

There exists a tool for visualizing Elasticsearch data, called Kibana, which is also developed by the Elastic Team [4]. It works on top of the content indexed on an Elasticsearch cluster and it can directly connect to an Elasticsearch server to be used for generating visualizations and reports; but again, the users must have prior knowledge about how Elasticsearch works and need to be experienced in dealing with its complex query language.

The paradigm shift introduced by model-driven development (MDD) [5, 6] in which the focus changes from code to models, leverages the abstraction level and promotes the software development for various application domains (e.g. [7-13]). Moreover, domain-specific languages (DSLs) / domain-specific modeling languages (DSMLs) [14-18] which have notations and constructs tailored toward a particular application domain, assist to the developers during execution of MDD processes by providing first a user-friendly syntax for modeling systems (mostly in a visual manner) and then a translational semantics for generating application software and any other artifacts automatically [19].

Abovementioned features and benefits of applying MDD and using DSMLs in other domains conduce toward producing a MDD framework also for Elasticsearch. Hence, in this paper, we introduce a DSML which can be used inside this MDD framework to facilitate the query writing process required for the Elasticsearch. Although many efforts exist in model-driven database processing and query generation (e.g. [20-23]), they do not consider the specifications of Elasticsearch and do not support generating queries, structured according to Elasticsearch which differs from the traditional databases.

Originating from a metamodel of Elasticsearch, which is also derived in this study, the proposed language provides a graphical concrete syntax for modeling queries within its integrated development environment (IDE). Models of the queries, visually prepared in this IDE, are automatically translated into corresponding Elasticsearch structures which are ready to be executed. If the developer requests execution

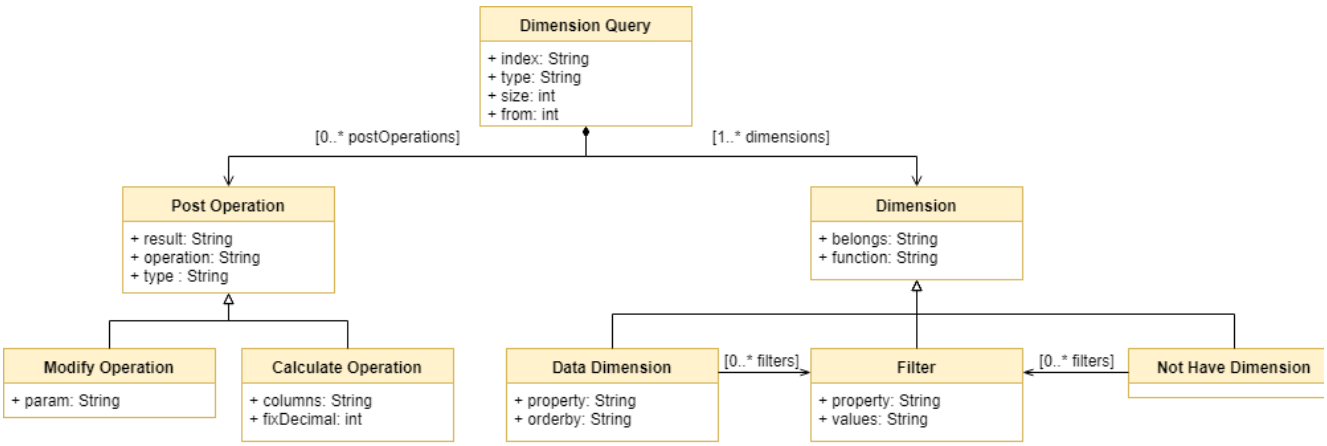


Fig. 1. DQL Metamodel

of these queries, it is also possible to execute those modeled and automatically generated queries on Elasticsearch storages. In this paper, we also discuss the use of this DSML for the industrial applications and give the results of evaluating its use inside a software company specialized for developing commercial Big Data applications.

The rest of the paper is organized as follows: In Section 2, the proposed query language is discussed with including its metamodel, fundamental elements and built-in query transformation process. Section 3 demonstrates the use of the language. Evaluation of the language and results of this evaluation are discussed in Section 4. Related work is given in Section 5. Finally, Section 6 concludes the paper.

II. Dimension Query Language

When we think of the three-dimensional space we are in, every object has coordinates to locate their position and hypothetically, it is possible to list and create reports for each object's or living creature's position on earth. Such report would have three fields for the coordinates linked with the name of the related object or person. To find an entry on the report, we would have needed to know the related entry's name and coordinates.

Mathematics and physics define dimension as the minimum number of points required to know an object's position and velocity on the space they belong. By this definition, we can say that our hypothetical report is a four-dimensional space, containing entries with four dimensions. Originated from this, we named our Elasticsearch query model as Dimension Model (DM) and the proposed Elasticsearch DSML as Dimension Query Language (DQL). In DM, each dimension corresponds to a field of data that must be included within the query.

In the following subsections we first define the fundamental elements and the relations inside DQL which compose the abstract syntax on the DSML for Elasticsearch. Then, we discuss how constraint checks and query validations are performed inside DQL's IDE before automatically transforming prepared query models into Elasticsearch queries. Finally, query transformation process is discussed.

A. *Fundamental DQL Elements*

Our transformation service (that will be discussed later) accepts Dimension Query (DQ) instances and generates Elasticsearch queries. The users can choose to view the transformed queries or directly execute them to view a table report over the underlying Elasticsearch storage. These DQ instances are created by conforming a metamodel which defines fundamental elements and their relations required for Elasticsearch queries. The metamodel, which leads to the generation of DQL syntax, is depicted in Fig. 1. Elements and properties of the metamodel are written in bold in the following text.

Elasticsearch storages, namely indexes, are a collection of documents that have similar characteristics [24]. Documents belonging to the same index may have relations with each other. On Elasticsearch, there are two types of relations. "Parent-child relation" links two documents by marking one as parent while marking the other one as child. "Nested relation" simply writes the whole document into another one.

On query transformation, one of the required properties is the name of the document on the top level, defining the document without a parent document. This needs to be specified as the **type** field in the **Dimension Query**. Another required field is the **index**, which is the name of the index to execute the query on. Finally, on the **dimensions** field, requested dimensions are expected. Along with these, there are some optional fields that a query can uphold, such as expected result size as **size** and offset as **from**.

B. *Dimension Types*

Dimensions vary in three types; **Data Dimension**s, **Filter** and **Not Have Dimension**s. Data dimensions are used to represent the fields to be retrieved upon the execution of the query and filter dimensions, by their name and hence they are used to filter the retrieved data based on some conditions. However, independently from its type, each

dimension must have two main fields; **function** and **belongs**.

The **function** field is used to specify which operation must be performed on the data. With this field; dimensions can be grouped, summed, counted and their average or percentage over their sum may be calculated. **belongs** field represents the name of the Elasticsearch document of the index in which the dimension data are located.

Data and filter dimensions must also have a field called **property**. This field represents the name of the data to be retrieved or filtered. The data dimensions may also have an additional **orderby** field to indicate which dimension must be used on ordering the query results.

If the query designers want to filter data on certain fields, filter dimensions may be used to meet this requirement. These dimensions will filter the data instead of creating another field on the result set. They are different from the data retrieval dimension by having additionally one field called **values**, indicating the values to apply with the filter. Filters can also be applied to specific data retrieval dimensions as well as they can be applied on the query. When used in this way, they affect only the dimension they are getting applied to.

The filtering criterion does not always have to be based on some values. For example, on a customer-invoice database, we may want to list the customers who did not place an order between some dates. To handle this case, there is an additional type of dimension called “Not Have Dimension”. This dimension can be used to filter certain fields, which does not have any relations to the given Elasticsearch document. Considering the customer-invoice example, let us think we have an index with two documents, customer and invoice, and assume that there is a parent-child relation, customer as parent and invoice as child. Creating a “not have dimension” with “belongs” field as `{belongs: 'invoice'}` will allow us to list the customers without an invoice on the whole Elasticsearch index. However, we may want to see the results based on another filter, such as a date interval. On this case, since the “not have dimension”s can also have filters, we can define a filter and add it to our “not have dimension”.

C. Post Operations

Calculations and value formatting is a common thing to do on report generation. When needed, Post Operations may be used to tinker with the retrieved data and they can be elaborated in two types; one for calculating new data, called **calculate operation**, and one for modifying existing data, called **modify operation**.

A post-operation must have the following fields: **result**, **operation** and **type**. They are common for each operation type, where the **result** field is the name of the data field which the post operation will be affecting. For calculation operations, this field will be used as the calculated field’s name. For modify operations, it is the name of the field on which the modification operation works on. **operation** field is the name of the operation to process, such as sum, divide, absolute, floor, ceil. Finally, the **type** field is the indicator of the type of the post-operation itself. It can be whether "*calculateOperation*" or "*modifyOperation*".

Calculate operations have two more additional fields. The first one the "*columns*" which holds the dimensions involved in the calculation operation and the second one is the "*fixDecimal*" which is to specify the number of the digits that should be displayed if the calculated value is a decimal number. Usable calculate operations are essential arithmetic functions; sum, subtract, multiply, divide.

Modify operations have only one additional field called "*param*"; which is the additional required parameter(s) to apply the operation and may not always be necessary. Modify operations, which can be used by the developers, are listed below:

Fix Decimal: To limit the number of the decimals to show of a decimal number. The param field must be the number of the digits.

Floor: To get the floor value of a decimal number.

Ceil: To get the ceiling value of a decimal number.

Abs: To get the absolute value of a number.

Replace: To replace some specific values of a dimension in the result set. A serialized JSON array string, which has objects as elements containing "from" and "to" values, is required as the "param" field.

D. Constraints and Query Validations

Our motivation is to simplify the query generation for the Elasticsearch without needing to know its query formulation details. However, there are also lots of syntactic controls and additional semantic constraints which should also be taken into account while writing Dimension Queries. Based on the metamodel elements and their relations discussed in the previous section, a modeling environment has been developed to use language constructs and features of DQL. Query developers may use our DSML’s graphical syntax and all required constraint checks and hence query validations can be realized automatically according to the Elasticsearch specifications.

Fig. 2 shows a screenshot taken from web-based IDE of the proposed Elasticsearch DQL. On the left side of the screen, the indexes on the system are listed with a combo box. After an index gets chosen, its metadata are shown directly under the index selector. The users then can start to create a DQ by simply dragging and dropping the fields they want to include in the query. The DQ gets automatically updated on the backstage each time the user drops a new field, updates or removes an existing one. The query can be tracked from the query panel at the top side of the screen dynamically.

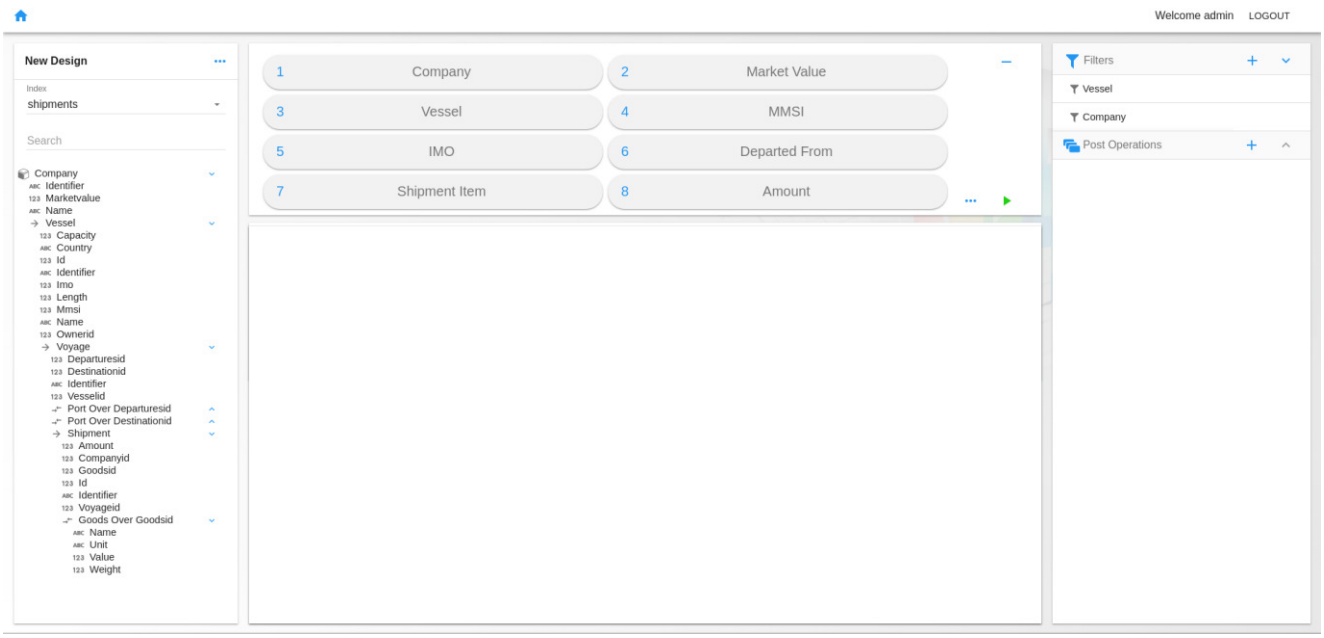


Fig. 2. IDE for the proposed Elasticsearch DSML

Filters and post operations are listed on the panel at the right side of the screen. Users can create new filters and post operations by clicking the add button (+) near them and can include filters to the query by simply dragging and dropping them into either on a data dimension or into the query directly.

After the users finish choosing the fields, the DQ can be sent to the backing server in order to be transformed to the Elasticsearch query. At this point, the users can choose to simply view the transformed query or the results generated with the execution of the transformed DQ.

There are lots of constraints needed to be followed while writing an Elasticsearch query. To be able to generate valid, executable Elasticsearch queries, we have also put some constraints on DQL and hence the IDE warns the user or prints an error message if a constraint gets violated.

Filters are for filtering data; they do not cause a field to be included within the result set. Therefore, filters cannot contain filters. Appending filters to other filters will have no effect on the generated query.

Since they affect the set of all results, the “not have dimension”s can only be used within the query, not within other dimensions.

Fields from different documents with parent-child relation cannot get queried without performing an operation over them. Because, for a member of the parent document, it is possible to have more than one value on the child document and it will not possible to create a result set without making some groupings on the parent document.

Mathematical processes such as number formatting, numerical calculations and digit rounding, can only be performed on numeric dimensions as well as date format operation can only be performed on date dimensions.

Applying group function to a dimension causes an aggregation to get started on Elasticsearch query. On Elasticsearch queries, when an aggregation gets started, all remaining fields must be included into that aggregation in some way. So, when the grouping function is applied over a dimension in the DQ, all remaining dimensions need to have a function value.

Each dimension of the query must be unique. If there is more than one dimension created with the same field of the same document -having also the same function-, one of the dimensions must have a filter different than the other one's filters at least. Semantic definitions on DQL, make all above constraint checks possible inside the IDE.

E. Query Transformation Stage

There are four stages of a query transformation which are all automatized within DQL’s IDE. First one is called the **Reducing** stage where the dimensions in the DQ get inspected and grouped by their respective nested documents on the Elasticsearch index. By doing this, it is possible to make fewer aggregations on the Elasticsearch query, therefore, it increases query execution performance by preventing same nested documents to get aggregated over and over. The requirement for two dimensions to be grouped is that they must be in the same document on Elasticsearch and they must have exactly same filters getting applied to them.

After dimensions are reduced, they get **Sorted** according to their documents and the relations between their documents on the index.

Two rules are applied while sorting the dimensions:

1) For multiple dimensions on the same document, if the document has a nested object, its own dimensions have priority than the nested object's dimensions. For instance,

considering the metadata given in Fig. 3, dimensions of Document A have a higher priority than dimensions of the Nested Object B. Likewise, Document C over Nested Object C1 and Nested Object C1 over Nested Object C1.1 have priorities.

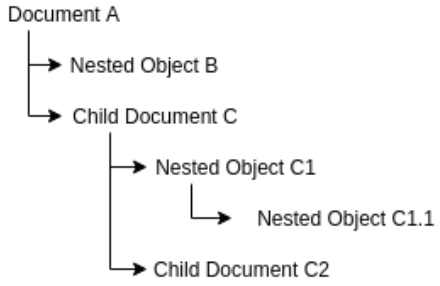


Fig. 3. Sample Index Metadata

2) Finally, calculation dimensions like sum, average, percentage, and count have the lowest priority so they take place at the end of the sorted dimensions list.

Analyzing stage is the one where the Elasticsearch query gets started to be created in pieces. On this stage, each dimension is converted to proper Elasticsearch query fragment and gathered up on a temporary list. This phase is crucial because, during aggregation generation, it is decided whether the aggregations will be linked with a nested relation or parent-child relation. "Not have dimension"s also will be included to the query on this stage.

The final stage of the query transformation is the **Generation** stage. On this stage, the query at hand is already has been reduced (for optimization), sorted and analyzed. The dimensions have been converted to aggregation blocks and relations between these aggregation blocks have been determined.

The list that's holding the aggregation blocks get iterated and linked with respect to their flags set from the previous stage of query transformation. Aggregation link is established with respect to the metadata of the index. That means, aggregation blocks will be linked to the others either as siblings or children according to their dimensions' "belongs" field.

III. Use of DQL During Query Generation

The Dimension Queries may be grouped into four main types corresponding to the types of the result sets they will generate upon the execution of the transformed Elasticsearch queries. This section will briefly explain these types. For a better comprehension, queries are represented in their textual notation during the following discussion, which are achieved automatically by using DQL and its graphical modeling environment.

In the first type, the query aims at getting direct results without making any grouping or filtering. If that's the case, there is only one constraint: As stated on the constraints section, the fields on the query must be on the same Elasticsearch document.

When the created query is in this type (see Fig. 4), dimensions must have three main fields. Function field of the dimensions must have a static value of "*include*" (shown in lines 6, 9, 12). Additionally, **orderby** field (line 5) may be added to one dimension to sort the results.

The corresponding translation (see Fig. 5) may seem simple because the translated query is obviously small and easy to write. However, the real power of the Dimension Queries, comes to stage when groupings and functions get involved with the query.

If the created query aims to fetch fields from different related documents (see "belongs" properties in Fig. 6 on lines 4, 7 and 10), the second type of query comes in. This type of query groups fields so different documents may be included in the result set. Again, on this type of query (Fig. 6), there is only one constraint: If a grouping, namely aggregation, starts with a dimension, all remaining dimensions must have a function applied on them.

```
01  { index:"indexName",
02    type: "doc",
03    dimensions: [
04     {property: "prop_1",
05      belongs: "doc", orderby: "asc",
06      function: "include" },
07     {property: "prop_2",
08      belongs: "doc",
09      function: "include" },
10     {property: "prop_3",
11      belongs: "doc",
12      function: "include"}
13    ],
14    from: 0,
15    size: 50 }
```

Fig. 4. Dimension Query without Groupings

```
01  { from: 0,  size: 50,
02    query: { bool: { disable_coord: false,
03        adjust_pure_negative: true,
04        boost: 1 }
05    },
06    _source: {
07      includes: ["prop_1","prop_2","prop_3"],
08      excludes: [ ]
09    },
10    sort: [{property_1.sort: {order: "asc"}}]
11  }
```

Fig. 5. Transformed Elasticsearch Query without Aggregations

```
01  {index: "indexName",
02    type: "topLevelDocName",
03    dimensions: [ {
04      property: "prop_1", belongs: "doc_1",
05      orderby: "asc", function: "agg"
06    }, {
07      property: "prop_2", belongs: "doc_2",
08      function: "sum"
09    }, {
10      property: "prop_3", belongs: "doc_2",
11      function: "sum"
12    } ],
13    from: 0, size: 50}
```

Fig 6. Dimension Query with Functions

While grouping the results, users may also want to do some calculations to see the summarized result. For example, let us consider a report to list the total invoice price for each customer. To do that, first an aggregation needs to be set up on customers ("function" field on line 5 in Fig. 6) then another summarization can be used on other dimensions. The reason for this aggregation requirement is, Elasticsearch needs to make some groupings to calculate summarized results such as sum and avg. Functions can only be applied when a grouping gets applied to the query.

Dimensions on this type of query must have same fields as the ones within the previous query, except the function field values may be "*agg*", "*count*", "*sum*", "*avg*" and "*percentage*" instead of "*include*" (see lines 5, 8 and 11 in Fig. 6). Fully generated Elasticsearch query for this DQ type can not be shown here due to space limitations. However, aggregations part of the generated query can be seen in Fig. 7.

```
01  ...
02  {aggregations: {
03   prop_1_doc_1_agg: {
04   terms: { field: "prop_1.keyword",
05    missing: "null", size: 2147483647,
06    min_doc_count: 1, shard_min_doc_count:0,
07    show_term_doc_count_error: false,
08    order: { _term: "asc" } },
09    aggregations: {
10     prop_2_doc_2_sum: {
11      children: { type: "documentName_2" },
12       aggregations: {
13        prop_2_doc_2_sum: {
14         sum: { field: "prop_2" } } } },
15        prop_name_3_doc_2_sum: {
16          children: { type: "doc_2" },
17          aggregations: {
18           prop_3_doc_2_sum: {
19             sum: { field: "prop_3" } } }
20  } } } } }
21  ...
```

Fig. 7. An excerpt from transformed Elasticsearch Query with Functions

If the users want to filter their data, they may create filter dimensions. Different usages of different filters are given in Fig 8. Applying a filter directly to the query is the case when the filters will be inserted within the *query* field of the transformed Elasticsearch query; thus affecting the whole result set as mentioned before. The filter (lines between 17 – 20 in Fig. 8) on this sample is for listing the results by the prop_3 field of the doc_2 with a value greater than 1000.

Applying filter to specific dimensions (lines between 13 – 14 in Fig. 8) will cause sub-query blocks to be created and inserted as filters to related aggregations. When this happens, the filters will be applied on only the related aggregation and, if there is any, to its sub aggregations.

Finally, the usage of "not have dimension" is shown between the lines 16 – 19 in Fig 8. In this example, a sample filter has been added to the "not have dimension". Before the "not have" filter gets applied, its inner filter will be applied first to narrow down the results.

An excerpt from the generic filter of the generated Elasticsearch query is given in Fig 9. On the transformed query, **prop_2** on line 4 is the name of the field to which the filter applies. It corresponds to the **property** field's value on the dimension query (see Fig. 8, line 14). The field called **from** (Fig 9, line 5) is the value of the **value** field previously indicated in Fig. 8, line 19. Finally, the **type** field in the line 14 (Fig. 9) is the name of the Elasticsearch document, containing the related fields. It corresponds to the **belongs** field (Fig. 8, line 18) in the Dimension Query filter.

```
01  {index: "indexName",
02    type: "topLevelDocName",
03    dimensions: [ {
04      property: "topLevelDocName",
05      belongs: "prop_1",
06      orderby: "asc", function: "agg" },
07     {property: "doc_1",
08      belongs: "prop_2", function: "avg" },
09     {property: "doc_1",
10      belongs: "prop_3", function: "avg",
11      filters: [ {
12        property: "doc_1", belongs: "prop_4",
13        function: "lt", values: [ 3000 ]}] },
14     {property: "prop_2", belongs: "doc_1",
15      function: "gt", values: [ 1000 ] },
16     {belongs: "doc_2", function: "nothave",
17      filters: [ {
18        property: "prop_4", belongs: "doc_2",
19        function: "gt", values: [ 10000 ]}]
20    } ],
21    from: 0, size: 50}
```

Fig. 8. Dimension Query with Dimension Filters

```
01  ...
02  {"has_child": { "query": { "bool": {
03    "must": [ { "range": {
04     "prop_2": {
05      "from": 1000, "to": null,
06      "include_lower": false,
07      "include_upper": true,
08      "boost": 1.0
09     } }
10    } ],
11    "disable_coord": false,
12    "adjust_pure_negative": true, "boost": 1.0
13   } },
14   "type": "doc_1",
15   "score_mode": "sum", "min_children": 0,
16   "max_children": 2147483647,
17   "ignore_unmapped": false, "boost": 1.0
18  } }
19  ...
```

Fig. 9. An excerpt from transformed Elasticsearch Query's Generic Filter

Depending on the function of the filter in the DQ, Elasticsearch query filter properties have different usages. On Fig 9. line 5, **to** field is used as the upper limit when the DQ function is either *less than or range.* On the given example **include_upper** field is *true* since it is a *greater than* filter and the upper value limit is infinity. **include_lower** field acts like the same when the filter function is *less than*. The same fields get used when the filter is *less than or equal to* and *greater than or equal to*.

```
01  ...
02  {"aggregations": {
03   "prop_1_topLevelDocName_agg": {
04    "terms": {
05     "field": "prop_1.keyword",
06     "missing": "null",
07     "size": 2147483647,
08     "min_doc_count": 1,
09     "shard_min_doc_count": 0,
10     "show_term_doc_count_error": false,
11     "order": { "_term": "asc" }
12    },
13    "aggregations": {
14     "prop_2_doc_1_avg": {
15      "children": { "type": "doc_1" },
16      "aggregations": { "prop_2_doc_1_avg": {
17        "avg": { "field": "prop_2" } },
18   "prop_3_doc_1_avg_filters_prop_4_lt_4000":
19        { "filters" },
20   "prop_3_doc_1_avg_filters_prop_4_lt_4000":
21        { "avg": { "field": "prop_3" } }
22  } } } } } }
23  ...
```

Fig. 10. An excerpt from transformed Elasticsearch Query's Aggregations

Part on the aggregations included in the same transformed query is given in Fig. 10. Aggregation names (bold texts on lines 3, 16 and 20 in Fig. 10) are generated by combining **property**, **belongs** and **function** fields on dimensions. The dimension specific inner filter (Fig. 11) is inserted in place of the bold *filters* text in line 19 of Fig. 10. The transformed "not have dimension" is given in Fig 12. Note that the inner filter is nearly the same as the one in Fig. 10. The *must not* keyword in line 2 determines the purpose of the filter.

```
01  {"filters": { "filters": [ {
02   "bool": { "filter": [ { "range": {
03    "prop_4": {
04     "from": null,
05     "to": 4000,
06     "include_lower": true,
07     "include_upper": false,
08     "boost": 1.0
09     } } } ],
10     "disable_coord": false,
11     "adjust_pure_negative": true,
12     "boost": 1.0
13    }
14   } ],
15   "other_bucket": false,
16   "other_bucket_key": "_other_"
17z } }
```

Fig 11. An excerpt from transformed Elasticsearch Query's Inner Filter

```
01  ...
02  {"bool": { "must_not": [ { "has_child": {
03   "query": { "bool": { "filter": [ {
04    "range": { "prop_4": {
05      "from": 10000, "to": null,
06      "include_lower": false,
07      "include_upper": true,
08      "boost": 1.0
09      } } } ],
10     "disable_coord": false,
11     "adjust_pure_negative": true,
12     "boost": 1.0 } },
13    "type": "doc_2",
14    "score_mode": "sum",
15    "min_children": 0,
16    "max_children": 2147483647,
17    "ignore_unmapped": false,
18    "boost": 1.0
19    } } ],
20   "disable_coord": false,
21   "adjust_pure_negative": true,
22   "boost": 1.0 } }
23  ...
```

Fig. 12. An excerpt from transformed Elasticsearch Query's "Not Have" Filter

IV. EVALUATION

An evaluation of using DQL has been performed at the industrial level with the participation of a group of developers from Galaksiya Information Technologies (http://galaksiya.com/). Galaksiya is a software company, located in Izmir, Turkey and its business domain mainly consists of Big Data and its applications. In some of their software solutions, the developers in the company recently started to work on Elasticsearch and related data storage.

At the beginning of the evaluation, we have determined the logistics as the target domain and created a logistics database for our case study. The main reason for choosing that domain is logistics datasets are very large in volume thus making them hard to query. Considering an end-user scenario to create a report over a logistics dataset to view the latest activities around the world, we have created a sample database by using the most active 51 ports on the world, 82 random selected shipping company names, distributed to 19 random countries. Each company on the dataset has random amount of ships with a total of 5000. The dataset has around 4750 auto-generated voyages with randomly selected goods. Total number of goods in the system is 50000, again all randomly generated.

Fig. 2 also shows a DQL instance model prepared for this evaluation. In the query panel residing at the upper middle of the IDE, there exist model items correspond to the required data dimensions in the query, namely Company, Market Value, Vessel, MMSI, IMO, Departure Port, Shipment Item and Amount. On the right panel under the filters, the Vessel and the Company are the defined filters which can be used in the query. Once dropped on a dimension or to the query, they will be transformed into filters within the related dimension or into a filter dimension depending where they are being applied.

As being an instance of DQL, the created query aims at listing the amount of the goods on each shipment with the information of departure ports, vessel details and company information. In addition, the same query model leads to prepare the query results inside a report grouping the data by the companies, vessels and departure ports.

For the qualitative assessment of DQL usage, five software developers became volunteer and agreed on being an evaluator. All of these evaluators has B.Sc. in computer science / software engineering and two of them are M.Sc. students in computer related fields at the time of this evaluation performed. Evaluators possessed the experience of developing software in industrial scale considering Big Data and/or Linked Open Data applications for different business domains (3 years on the average). Although they were skilled with creating database queries and working with data storages, they had no or very little knowledge on the query language required for Elasticsearch. After a brief introduction of DQL and its IDE, the evaluators were requested to create the same report given in Fig. 2. Upon completion their modeling session, a questionnaire including the following open-ended questions was given to the evaluators and their responses were gathered:

1. How does DQL and its IDE make writing Elasticsearch queries easier?
2. Did you encounter any difficulties while modeling queries and creating reports with using DQL? If any, please provide your suggestions to fix them.
3. Do you think DQL is easy to learn and use?

All the evaluators agreed on the biggest advantage of DQL that it eliminates the syntax errors which may be encountered while creating a query since there is no query writing process. They also agreed that the use of the DQL removed hardcoding the Elasticsearch queries hereafter. And most of them indicated that it is possible to create Elasticsearch queries without writing a single line of code. One of the evaluators stated that DQL's graphical syntax is comprehensive enough to cover all Elasticsearch domain and accompanying IDE helped them for determining and visualizing the details of queries from scratch. Some of the evaluators found the model panel residing on the left side of the DQL IDE (see Fig. 2) very helpful by means of dynamically showing the whole data model pertaining to the query under development.

For the second question, some of the evaluators stated that applying a filter to the report directly or using it separately on dimensions is a little bit confusing at the first time but after using the editor for a while, it gets simpler. Based on the feedbacks gained from the evaluators, visual concrete notations required for query modeling and organization of them inside the IDE were also re-arranged since some of the evaluators found the arrangement of these components a bit complicated.

Finally, for the last question, everyone agreed that even end-users with no knowledge on Elasticsearch would be able to use DQL for Elasticsearch query design and implementation after a small training. Most of the evaluators confirmed that there is no need to know any kind of syntax and programming (or querying) language for a person to use DQL and its IDE. They also added that little knowledge about basic query logic is enough for a user before using DQL. However, all of the evaluators also answered the third question by indicating there is still a learning curve to get used to the DQL editor, but with a short training session it becomes easy to learn and design reports.

In order to measure whether the use of DQL speeds up query creation, each evaluator's query design with using DQL has been recorded. The evaluators completed the generation of the Elasticsearch query required for the above logistics case study around 30 mins. on the average. The evaluators were also requested to create the same query again but this time without using DQL. They just used Elasticsearch query syntax and the result was amazing: It took around 6.5 hours to complete writing the same query on the average. Although that measurement was achieved from

only a single case study, we believe that the speedup gain obtained with using DQL in here is promising since the experts in the company confirmed that the query handled in here is complex enough comparing with the exact queries created in their commercial applications for datasets which are almost same size with the logistics dataset used in the case study.

V. Related Work

Like other domains, in order to master the problems of creation, management and evolution of databases and querying on these databases, the researchers investigate the ways of applying MDD principles and/or proposing the use of DSLs / DSMLs. For instance, MDSheet is proposed in [25] for model-driven engineering of spreadsheets. End-users can build sheets within MDSheet framework via its tool. The framework is enriched with a model-driven query language [20] which supports most of the SQL standards. The language is also structured as a DSL and the related MDE framework is integrated with Google Query function in [26]. Similarly, FDL [27] is a description language for spreadsheets, which is empowered with visualization and analysis tool for constructing the separation between the input of formulas and the output of calculation results. Although these studies provide a good MDD framework for spreadsheets, it gets very difficult to extract information on a single potentially large matrix in an effective way inside spreadsheets and this deficiency may cause spreadsheets a weak alternative for databases, especially the ones as being Elasticsearch storages.

Ristic et al. [21] define a model-driven database reverse engineering mechanism through a chain of model-to-model transformations. These transformations are applied between physical database schema and generic relational schema. A similar model-to-model transformation approach is followed in [22] for automatically achieving a form type data model again from a generic database schema. Hence, the form type specification represents a platform independent prescription model of both future screens and report forms which can be generated later for a complete application. Popovic et al. [23] propose a DSL, called IIS*CFuncLang, to specify application-specific functionalities of business applications for different domains at the platform-independent model level. The DSL enables modeling the system to be developed and generalization of the required executable codes is realized via some model transformations. Hence, specifications defined with the DSL can be converted into executable PL/SQL program codes. These studies bring valuable MDD solutions on database processing, query generation and reverse engineering of databases. However, the metamodels, the transformations and the DSLs defined in these studies do not consider the Elasticsearch engines and hence, deriving an MDD framework for generating queries, structured according to Elasticsearch specifications on various query types, is not covered in these studies.

Research on Elasticsearch has been recently emerged due to novelty brought into query structures. Kononenko et al. [3] discuss how Elasticsearch differs from the traditional relational databases and give some concrete applications of using Elasticsearch queries. In addition, they give their assessment on the strengths and the weaknesses of Elasticsearch for querying new software repositories. Elasticsearch's inverted index capabilities are used in [28] for implementing an optimized intelligent search algorithm. Query optimization with using this algorithm is employed in the retrieval of medicine data. Finally, a social media analysis system is introduced [29] in which features of Elasticsearch are used on analyzing Big Data. Two ways of giving Twitter data as input to Elasticsearch are defined and their performances are compared by means of consuming hardware resources and the capacity of processing tweets. Our work contributes to the research on Elasticsearch by introducing a DSML and its supporting IDE which can be used to facilitate and expedite the creation of Elasticsearch queries by following a MDD process.

VI. Conclusion

A DSML, called DQL, for supporting MDD of Elasticsearch queries has been introduced in this paper. Based on the derived metamodel of Elasticsearch queries, a graphical concrete syntax is provided for query modeling inside the IDE of the language. All required constraint checks and query validations are automatically performed on the models prepared inside this IDE and Elasticsearch queries are generated from these models. Furthermore, IDE is capable of executing these auto-generated Elasticsearch queries on remote repositories and creating reports covering the execution results. The conducted evaluation showed that the use of the language significantly decreases the development time required for creating Elasticsearch queries. Finally, qualitative assessment, based on the developers' feedback, exposed how DQL facilitates the development of Elasticsearch queries.

In the future work, our aim is to extend DQL's coverage on different types of Elasticsearch use cases. Also, we plan to enrich DQL's query modeling environment with improved visualization components especially for reporting Elasticsearch results. In order to determine how these new components enable more feasible query generation, the evaluation performed on DQL will be improved and experiment settings will be structured with including some sort of hypothesis testing as being considered in similar efforts like [30-32].

Acknowledgment

We would like to thank software developers from Galaksiya Information Technologies for their cooperation and valuable feedbacks. This work was supported by Yaşar Group (http://yasar.com.tr/en/).

REFERENCES

[1] A. Bialecki, R. Muir, G. Ingersoil. 2012. "Apache Lucene 4", in *Proc. SIGIR 2012 Workshop on Open Source Information Retrieval*, Portland, Oregon USA, pp. 17–24.

[2] Elasticsearch BV. 2014. "Elasticsearch - The Heart of the Elastic Stack", available at: https://www.elastic.co/products/elasticsearch (last access: July 2018)

[3] O. Kononenko, O. Baysal, R. Holmes, M. W. Godfrey. 2014. Mining modern repositories with elasticsearch. In *Proc. 11th Working Conference on Mining Software Repositories (MSR 2014)*, Hyderabad, India, pp. 328–331, DOI: 10.1145/2597073.2597091.

[4] Elasticsearch BV. 2015. "Kibana - Your Window into the Elastic Stack", available at: https://www.elastic.co/products/kibana (last access: July 2018)

[5] B. Selic. 2003. The pragmatics of model-driven development. *IEEE Software* 20: 19-25, DOI: 10.1109/MS.2003.1231146

[6] J. Poruban, M. Bacikova, S. Chodarev, M. Nosal. 2014. "Pragmatic Model-Driven Software Development from the Viewpoint of a Programmer: Teaching Experience", in *Proc. 3rd Workshop on Model Driven Approaches in System Development (MDASD@FedCSIS'14)*, Warsaw, Poland, pp. 1647–1656, DOI: 10.15439/2014F266.

[7] M. Brambilla, J. Cabot, M. Wimmer. 2017. *Model Driven Software Engineering in Practice, Second Edition*, Morgan & Claypool, DOI: 10.2200/S00751ED2V01Y201701SWE004

[8] J. Whittle, J. Hutchinson, M. Rouncefield. 2014. The state of practice in model-driven Engineering. *IEEE Software*, 31(3):79-85, DOI: 10.1109/MS.2013.65.

[9] G. Kardas. 2013. Model-driven development of multi-agent systems: a survey and evaluation. *The Knowledge Engineering Review*, 28(4): 479-503, DOI: 10.1017/S0269888913000088

[10] S. Mustafiz, X. Sun, J. Kienzle, H. Vangheluwe. 2008. Model-driven assessment of system dependability. *Software & Systems Modeling*, 7(4): 487-502, DOI: 10.1007/s10270-008-0084-1.

[11] H. B. Saritas, G. Kardas. 2014. A model driven architecture for the development of smart card software. *Computer Languages, Systems & Structures*, 40(2): 53-72, DOI: 10.1016/j.cl.2014.02.001.

[12] A. Harbouche, N. Djedi, M. Erradi, J. Ben-Othman, A. Kobbane. 2017. Model driven flexible design of a wireless body sensor network for health monitoring. *Computer Networks*, 129(2): 548-571, DOI: 10.1016/j.comnet.2017.06.014.

[13] F. Erata, C. Gardent, B. Gyawali, A. Shimorina, Y. Lussaud, B. Tekinerdogan, G. Kardas, A. Monceaux. 2017. "ModelWriter: Text & Model-Synchronized Document Engineering Platform", in *Proc 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE 2017)*, Urbana-Champaign, Illinois, USA, pp. 907-912.

[14] M. Mernik, J. Heering, A. Sloane. 2005. When and how to develop domain-specific languages. *ACM Computing Surveys*, 37(4): 316-344, DOI: 10.1145/1118890.1118892.

[15] M. J. Varanda Pereira, M. Mernik, D. da Cruz, P. Rangel Henriques. 2008. Program Comprehension for Domain-specific Languages. *Computer Science and Information Systems*, 5(2): 1-17, DOI: 10.2298/CSIS0802001P.

[16] I. Lukovic, M. J. Varanda Pereira, N. Oliveira, D. da Cruz, P. Rangel Henriques. 2011. A DSL for PIM specifications: Design and attribute grammar based implementation. *Computer Science and Information Systems*, 8(2): 379-403, DOI: 10.2298/CSIS101229018L.

[17] T. Kosar, M. Mernik, J. Gray, T. Kos. 2014. Debugging measurement systems using a domain-specific modeling language. Computers in Industry, 65(4): 622-635, DOI: 10.1016/j.compind.2014.01.013.

[18] B. Bryant, J-M. Jezequel, R. Lammel, M. Mernik, M. Schindler, F. Steinmann. 2015. "Globalized Domain Specific Language Engineering", in *Globalizing Domain-Specific Languages*. B. Combemale, B. Cheng, R. France, J-M. Jezequel, B. Rumpe (eds). *Lecture Notes in Computer Science*, 9400: 43-69, DOI: 10.1007/978-3-319-26172-0_4.

[19] G. Kardas, B. T. Tezel, M. Challenger. 2018. Domain-specific modelling language for belief-desire-intention software agents. *IET Software*, DOI: 10.1049/iet-sen.2017.0094.

[20] J. Cunha, J. P. Fernandes, J. Mendes, R. Pereira, J. Saraiva. 2013. "Querying model-driven spreadsheets", in *Proc. 2013 IEEE Symposium on Visual Languages and Human Centric Computing (VL/HCC 2013)*, San Jose, CA, USA, pp. 83-86, DOI: 10.1109/VLHCC.2013.6645247.

[21] S. Ristic, S. Aleksic, M. Celikovic, V. Dimitrieski, I. Lukovic. 2014. Database reverse engineering based on meta-models. *Central European Journal of Computer Science*, 4(3): 150-159, DOI: 10.2478/s13537-014-0218-1.

[22] S. Ristic, S. Kordic, M. Celikovic, V. Dimitrieski, I. Lukovic. 2016.. "A Model-to-Model Transformation of a Generic Relational Database Schema into a Form Type Data Model", in *Proc. 4rd Workshop on Model Driven Approaches in System Development (MDASD@FedCSIS'16)*, Gdansk, Poland, pp. 1577–1580, DOI: 10.15439/2016F408.

[23] A. Popovic, I. Lukovic, V. Dimitrieski, V. Djukic. 2015. A DSL for modeling application-specific functionalities of business applications. *Computer Languages, Systems & Structures*, 43: 69-95, DOI: 10.1016/j.cl.2015.03.003.

[24] Elasticsearch BV. 2015. "Elastic Stack and Product Documentation", available at: https://www.elastic.co/guide/index.html (last access: July 2018)

[25] J. Cunha, J. P. Fernandes, J. Mendes, J. Saraiva. 2012. "MDSheet: A framework for model-driven spreadsheet engineering"., in *Proc. 34th International Conference on Software Engineering (ICSE 2012)*, Zurich, Switzerland, pp. 1395-1398, DOI: 10.1109/ICSE.2012.6227239.

[26] J. Cunha, J. P. Fernandes, J. Mendes, R. Pereira, J. Saraiva. 2015. "Design and Implementation of Queries for Model-Driven Spreadsheets", in *Central European Functional Programming School*. V. Zsok, Z. Horvath, L., Csató (eds). *Lecture Notes in Computer Science*, 8606: 459-478, DOI: 10.1007/978-3-319-15940-9_13.

[27] Y. Horry. 2017. Financial information description language and visualization/analysis tools. *Computer Languages, Systems & Structures*, 50, 31-52, DOI: 10.1016/j.cl.2017.05.005.

[28] C. Bhadane, H. A. Mody, D. U. Shah, P. R. Sheth. 2014. Use of Elastic Search for Intelligent Algorithms to Ease the Healthcare Industry. *International Journal of Soft Computing and Engineering*, 3(6), 222-225.

[29] P. P. I. Langi, Widyawan, W. Najib, T. B. Aji. 2015. in Proc. 2015 International Conference on Information, Communication Technology and System (ICTS 2015), Surabaya, Indonesia, pp. 181-186, DOI: 10.1109/ICTS.2015.7379895.

[30] F. Haser, M. Felderer, R: Breu. 2016. Is business domain language support beneficial for creating test case specifications: A controlled experiment. *Information and Software Technology*, 79, 52-62, DOI: 10.1016/ j.infsof.2016.07.001.

[31] A. N. Johanson, W. Hasselbring. 2017. Effectiveness and efficiency of a domain-specific language for high-performance marine ecosystem simulation: a controlled experiment. *Empirical Software Engineering*, 22(4), 2206-2236, DOI: 10.1007/s1066.

[32] T. Kosar, S. Gaberc, J. C. Carver, M. Mernik. 2018. Program comprehension of domain-specific and general-purpose languages: replication of a family of experiments using integrated development environments. *Empirical Software Engineering*, DOI: 10.1007/s10664-017-9593-2.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 863–866
DOI: 10.15439/2018F313
ISSN 2300-5963 ACSIS, Vol. 15

Approaches to Semantic Mutation of Behavioral State Machines in Model-Driven Software Development

Anna Derezińska
Warsaw University of Technology
Institute of Computer Science
Nowowiejska 15/16, 00-665 Warsaw Poland
Email: A.Derezinska@ii.pw.edu.pl

Łukasz Zaremba
Warsaw University of Technology
Institute of Computer Science
Nowowiejska 15/16, 00-665 Warsaw Poland

***Abstract*—Behavior of UML state machines can be a source of interpretation problems in model to code transformation. Different solutions to the semantic variants could be defined as a special kind of mutations, similarly as in the mutation testing. State machines together with class models can be a source of an Model-Driven Software Development process aimed at building an executable application. We have compared several approaches to creating applications based on models in which semantic mutation operators of state machine behavior are used. The most promising approach has been utilized to extend the Framework for eXecutable UML (FXU) with semantic mutation facilities. The framework supports code generation from UML classes and their state machines as well as developing C# applications according to selected mutations of state machine behavior. The tool has been used in evaluation of a case study.**

I. Introduction

AUTOMATIC code generation in Model Driven Software Development (MDSD) can be based not only on structural models, like UML classes, but also on behavioral models, e.g. state machines [1]. One of the obstacles is not sufficient support for generation and verification of applications of this kind.

Mutation testing is used for evaluation of test suites and generation high-quality tests [2]. Syntactic changes injected into a source code are supposed to be detected by test cases. Modified programs, so-called ***mutants***, are run against tests. An abnormal program behavior confirms ability of the tests to detect the type of faults introduced by ***mutation operators*** during mutant generation.

Mutation testing approach has been extended for different software artefacts to be mutated and tested in a software development life cycle. A mutated source can be a model or specification [3], including state machines [4], [5]. A special kind of mutation testing is focused not on syntactical changes of an input, i.e. code, model, or another artefact, but on changes in its semantic or other implementation features [5]-[7].

In this paper we focus of different architectural approaches to combine semantic mutation of state machines into an MDSD process. It is assumed that an executable application is created based on UML classes and hierarchical state machine models [8]. The final code project is built with all necessary library notions, so the target application can be run as any other general–purpose application in a standard environment. The developed application should reflect system requirements that are specified in the input models, therefore the final testing is performed at application level, and not at model level.

Presentation of this idea and possibility of its realization is the main contribution of the paper. We compare different approaches to performing such semantic mutation in regard to their applicability in practice and complexity of realization (Sec. III). Complexity analysis of the approaches helps to select the best one which has been implemented in FXU – a tool that supports MDSD from UML classes and their behavioral state machines with the target to C# applications (Sec. IV). Therefore, we have also shown how the semantic mutation testing can be practically combined into an MDSD process. To the best of our knowledge it is the first implementation of such mutation approach.

II. Related Work

The main background of this work originate from areas of code generation from state machines, interpretation of state machine behavior, and mutation testing.

UML model to code transformation based on class models can be extended with state machine models [1]. Tools that support this usually respect only a subset of notions, omitting complex concurrency issues. Some solutions that apply comprehensive set of state machine concepts do not support C#, apart from FXU [8].

The UML specification has included some semantic variation points, in particular concerning behavior of state machines. They should be resolved in different ways while a model has to be interpreted or a model-based application executed. In most of implemented solutions, there are different resolutions of behavioral interpretation problems, often without precise statements about selections taken.

Mutation testing approach has been used to applications in different programing languages [2], including C# [9]. This idea was also used to mutate UML models, e.g. class

models [3], or automata-based models, mainly dealing with syntactical changes of diagrams [4].

Behavioral models, including state machines, have been also considered as an object of semantic mutation [6], in some variants called also an implementation mutation [4], [5]. In this kind of mutation there are no changes introduced into a model graph structure, but different semantic interpretations are considered [7].

III. Different Approaches to Combining Semantic Mutation of State Machines into MDSD

In a model-based process of software development mutation testing can be used at different levels, applied to various software artefacts, and with evaluation of effects in different process stages. Realizations of mutation of model semantics can be classified into three main types [6]:

1) Simulation/interpretation of a model with different parameters mimicking semantic variants.
2) Semantic expressed in a set of configurable rules that is combined in an executable model or a target application.
3) Imitation of semantic mutations using syntactic changes of elements in a model or in a code.

In this paper we focus on approaches that belong to the second realization type, discuss mutating of input model semantic, but testing the final code application.

A. Categories and Strategies of Mutation Operators

The following general mutation categories which refer to elements mutated in MDSD can be distinguished:

A) design or construction mutation
B) semantic mutation
C) semantic consequence-oriented mutation

The first category includes typical mutation testing defined for programming languages, as well as modifications of input models. However, in this paper we do not deal with this category.

Semantic faults can be imitated by semantic mutation, or semantic consequence-oriented mutation, or such structural mutations that reproduce semantic faults [5]. In comparison to design mutation, semantic mutations do not modify an intermediate source form of a model or code but apply another interpretation of it. Transformation rules from a source to an intermediate form are modified.

The third mutation category is associated with realization of a given meaning of modelled programming concepts. System realization consistent to a given semantic determines a final system behavior. However, according to a semantic, behavior of a system or its part can be nondeterministic. This mutation category is aimed at imitation of different behavioral combinations.

This kind of mutation was considered as implementation-oriented mutation [5] specified in the context of the Harel statecharts. However, an approach to realization of such mutations proposed in this paper is different to those from Trachtenbrot [5]. The details of the semantic operators are beyond the scope of this paper and will be published in [10].

A mutation operator could be applied in many places of a program, but in the ***first order mutation,*** code is changed in one place per one mutant [2]. In general, the number of generated mutants will be denoted as *MN*, and equal to:

$$MN = \sum_{i=1..N} OP_i \quad (1)$$

where N is a number of operators and OP_i is the number of program places in which the *i-th* operator can be applied.

Considering behavior variants, we generally assume that only one operator is applied. However, the application of such operators can be mutually dependent. It means, for example, that only after operator OP_x had been selected, a variant determined by another operator OP_y could be used. Taking into account such dependency of operators, the final mutant can still be counted as a first order mutant under application of a composite operator OP_xOP_y.

A model usually includes many state machines, thus the same operator could be used to one or many state machines at the same time. Hence two strategies could be considered:

1) ***all state machines***, i.e. the same mutation operator refers to all state machines in a model to be transformed,

2) ***one or selected state machines***, i.e. only selected state machines (usually one) have different behavior determined by the operator.

In the first strategy, the number of generated mutants depends linearly on the number of operators and is lower than in the second approach. The interpretation of the behavior is also simpler. The latter strategy could result in higher number of mutants, especially for complex systems with many state machines. The number of all possible mutants is of order of $N*K$ where N is a number of mutation operators and K is the number of state machines.

B. Approach I – Multiple Code Generation

This is a simple approach based on a straightforward creation of code in a MDSD process. For each mutant, i.e. for a pair <model, semantic>, a separate process towards a target application will be performed. A result of the transformation would be a code that implements model with the semantic. The application code might be slightly different for each mutant. Each mutant has its own code project and requires to be compiled.

The main process metrics are summarized in Table I. Approach I is simple and independent of a code generator, but have many disadvantages. The new code has to be written in each mutant while a method body is supplemented. Moreover, it could be repeated before adjusting a mutant to any test run. Therefore, the approach could be used for a quick verification of a semantic mutation operator, but it not convenient for the mutation testing in MDSD.

C. Approach II – Multiple Libraries

Drawbacks of the first approach imply that mutation testing process should be independent from model-to-code transformation and from supplementing of the generated code. This idea has already been partially supported if we

separate code generation and run-time libraries, where the library deliver the semantic of state machines.

The rules of a state machine behavior could be encapsulated in a library, therefore we could mutate the library and not a generated code that could be independent. Model transformation and code supplementing would belong to one process which is independent from the mutation testing process. The model to code transformation is performed only once, and we could also extend some code one time, if necessary. In dependence of a selected mutation, an appropriate library could be chosen and used to build the final target application.

The main drawback of this approach is necessity to maintain many versions of the library (Table I). The number of versions is equal to the product of supported mutation operators and their possible interpretations. Hence, the complexity rises up quite fast.

D. Approach III – Mutants with Common Base

This approach extends the library with different variants of classes that represent various concepts of state machine. The whole idea is similar to the *Strategy* design pattern. Each class implements a single interface that corresponds to one state machine notion. All classes are gathered into one library that could be added to a final application. Based on an input model and selected mutation operators, multiple classes are created for each state machine.

A base class specifying behavior is generated for each model class that has its state machine. This base class could include methods originated from the model class as well as additional methods for initialization of the state machine of a default semantic. Moreover, for each mutant a class derived from the base class is created. Each such inherited class redefines methods for initialization of the state machine corresponding to the semantic of a given mutant.

If a method body is supplemented in the generated code, it can be done once in the base class. Creation of an executable application requires one compilation of all mutants together. Tests should be defined for a base class in order to be run for all generated mutants.

An advantage of this approach is a single implementation of additional code. Selection of mutation operators in experiment iterations, i.e. updating a variant, could be realized by substitution of a constructor in the generated mutant code that would require changing in the code generator, which is not a flexible solution.

E. Approach IV – Configurable Library

In the fourth approach we combine advantages of approaches II and III. Considering a set of mutants, we can use a pair <model, a set of semantics> instead of a set of pairs <model, semantic>. Moreover, the source code generated from a model is explicitly separated from the library code. The generated code can use the library only by dedicated interfaces, placed in an additional intermediate layer. The generated code does not depend on the library classes that implement those interfaces.

In result, one version of code is generated for each state machine of a model. It would be used in an original application and a mutated one. Therefore, supplementing of a method code body is performed only once. We need also only one compilation of the application.

Furthermore, mutation testing process uses one consistent run-time library. Consequently, maintenance and extending of the semantic mutation operators would be uncomplicated. In a single mutant, various state machines can be executed according to different semantic variants, if desired.

This approach has many advantages, but creating of objects of state machines could take more time due to reflection mechanism used in the intermediate layer. On the other hand, this activity is performed only once during the live time of an object that includes a state machine.

F. Comparison of Approaches

The approaches are summarized in Table I. We have compared some relevant metrics of the process and product complexity. Only in the first primitive approach we have to build many code projects (row 1) and supplement the same code in many applications (row 5). In other cases one project is used for all mutants regardless of the number of operators and the number of their interpretations.

When multiple variants are introduced into libraries, only one class in the source code corresponds to one model class (row 2). The second approach requires many libraries (row 3), while the others can use a single one. An important time overhead is associated with multiple compilation, which is necessary for two first approaches only (row 4).

Summing up quantitative data in the upper part of the Table (rows 1-5), we can conclude that the fourth approach has the lowest complexity (1 in all metrics).

The bottom part of the Table assesses the mutation testing process flexibility and extensibility. Here, also the last approach would be the most beneficial. Iterative mutation testing can be easily performed, and new semantic mutation ideas could be easily introduced.

IV. Realization of Semantic Mutation with FXU

Framework for eXecutable UML (FXU) creates executable C# from UML models [8]. It was the first tool that supported transformation of state machines to C#, and still belongs to comprehensive tools that covers all notions of behavioral state machines, with complex, orthogonal states, different pseudostates, also history, etc. [1]. The FXU Generator transforms UML classes and their state machines into C# code. The FXU Library contains implementation of all state machine concepts. The final application is built as a project including the generated code and the library.

Basing on the analytic evaluation of the approaches, FXU has been extended to support semantic mutation of state machines using the fourth approach - configurable library. Both strategies, all-state machines and one selected state machine, have been implemented. The reconfigured FXU Library provides versatility of state machine semantic mutation opera-

TABLE I.
COMPARISON OF APPROACHES I-IV (MN – NUMBER OF MUTANTS)

Metric	I	II	III	IV
1 Number of generated projects	MN	1	1	1
2 Number of code classes originated from a model class which has its state machine	MN (one in a project)	1	MN +1	1
3 Number of run-time libraries	1	MN	1	1
4 Number of compilation runs	MN	MN	1	1
5 Number of spots where the same code is placed in project(s)	MN	1	1	1
6 Mutant creation process independent of code generation	No	Yes	No	Yes
7 Separate compilation needed to create any executable mutant	Yes	Yes	Yes	No
8 Easy extensibility with other semantic mutations	High	Med ium	Med ium	High
9 Difficulty in performing an iterative mutation testing	Low	High	Med ium	Low

tors, including semantic mutations and semantic consequence-oriented mutations (Sec III).

Evaluation of the mutation testing process with the extended FXU has been performed on a case study used in the previous MDD experiments [11]. It referred to modeling of a presence server in a social network. Here, we have focused on the application verification, showing different application alternatives reflecting activities consistent with various semantic variants of UML state machines.

A set of unit tests for the application was developed. The test project was supplemented with a configuration file of state machine semantic. The tests followed two schemata, in which (i) we checked a correctness of only one class and its behavior specified by its state machine, or (ii) a whole subsystem was verified. An example of the latter case could be servicing of a data publishing request. It was verified if a valid status was set in appropriate places. In case of tests that check one class and one state machine, semantic for the whole was mutated. In case of subsystem tests, two types of mutants were configured. (A) All state machines behaved according to the same semantic variant within the same test run. (B) Different state machines of the involved classes used various semantic variants within the same test run.

All tests were run against the created mutants and positively evaluated in the environment. The behavior of the mutants corresponded to expectations given in the input models and semantic variants.

V. CONCLUSION

Different approaches to introducing semantic mutation of state machines have been compared. The best solution in terms of complexity and flexibility has been implemented in the FXU, the framework transforming class and state machine models into C# applications. While using this tool support selected behavioral variants to state machines were accomplished and verified in mutation testing experiments.

REFERENCES

[1] E. Dominguez, B. Perez, A.L. Rubio, and M.A. Zapata, "A systematic review of code generation proposals from state machine specifications," *Information & Software Technology*, 54, no. 10, 2012, pp. 1045-1066. http://dx.doi.org/10.1016/j.infsof.2012.04.008

[2] M. Harman and Y. Jia, "An analysis and survey of the development of mutation testing," *IEEE Transactions Software Engineering*, vol. 37, no. 5, 2011, pp. 649-678, http://dx.doi.org/10.1109/TSE.2010.62

[3] A. Derezińska, "Object-oriented mutation to assess the quality of tests," in *Proc. of the 29th Euromicro Conf., IEEE Comp. Society*, Los Alamitos, California, 2003, pp. 417-420. htpp://dx.doi.org/10.1109/EURMIC.2003.1231626

[4] M. Trakhtenbrot, "New mutation for evaluation of specification and implementation levels of adequacy in testing of statecharts models," in *Proc. of the 3rd Workshop on Mutation Analysis (MUTATION'07)*, Windsor, 2007, pp. 151-160. http://dx.doi.org/10.1109/TAIC.PART.2007.23.

[5] M. Trakhtenbrot, "Implementation-oriented mutation testing of statechart models", *Proc. 3rd Int'l. Conf. on Software Testing, Verification, and Validation Workshops,* Paris, 6-9 April 2010, pp.120-125. http://dx.doi.org 10.1109/ICSTW.2010.55

[6] J.A. Clark, H. Dan, and R.M. Hierons, "Semantic mutation testing," in *Science of Computer Programming*, no. 78 pp. 345-363, 2013. http://dx.doi:10.1016/j.scico.2011.03.011

[7] M. Trakhtenbrot, "Mutation patterns for temporal requirements of reactive systems," in *Proc. of 10th IEEE Intern. Conf. on Software Testing, Verification and Validation Workshops*, 2017, pp. 116-121. http://dx.doi.org 10.1109/ICSTW.2017.27

[8] A. Derezińska and R. Pilitowski, "Realization of UML class and state machine models in the C# Code Generation and Execution Framework," *Informatica* vol. 33, no 4, pp. 431-440, Nov. 2009.

[9] A. Derezińska and A. Szustek, "Object-Oriented Testing Capabilities and Performance Evaluation of the C# Mutation System," in *Proc. 4th IFIP TC2 Central and Eastern European Conference on Software Engineering Techniques CEE-SET 2009*, LNCS, vol. 7054, pp. 229-242, Springer, 2012. http://dx.doi.org/:10.1007/978-3-642-28038-2_18

[10] A. Derezińska, "Mutating state machine behavior", unpublished.

[11] A. Derezińska, M. Szczykulski, "Towards C# application development using UML state machines – a case study," in *Emerging Trends in Computing, Informatics, System Sciences, and Engineering*, T. Sobh, K. Elleithy, Eds. LNEE vol. 151, Springer, 2013, pp. 793-803, htpp://dx.doi.org /10.1007/978-1-4614-3558-7_68

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 867–876
DOI: 10.15439/2018F64
ISSN 2300-5963 ACSIS, Vol. 15

Reverse Engineering of Legacy Software Interfaces to a Model-Based Approach

Mathijs Schuts*, Jozef Hooman†, Ivan Kurtev‡ and Dirk-Jan Swagerman§
*Philips, Best, The Netherlands
Email: mathijs.schuts@philips.com
†ESI (TNO), Eindhoven, The Netherlands, and
Radboud University, Nijmegen, The Netherlands
‡Altran, Eindhoven, The Netherlands
§Philips, Best, The Netherlands

***Abstract*—Cyber-physical systems consist of many hardware and software components. Over the life-cycle of these systems, components are replaced or updated. To avoid integration problems, good interface descriptions are crucial for component-based development of these systems. For new components, a Domain Specific Language (DSL) called Component Modeling & Analysis (ComMA) can be used to formally define the interface of such a component in terms of its signature, state and timing behavior. Having interfaces described in a model-based approach enables the generation of artifacts, for instance, to generate a monitor that can check interface conformance of components based on a trace of observed interface interactions during execution. The benefit of having formal interface descriptions also holds for legacy system components. Interfaces of legacy components can be reverse engineered manually. In order to reduce the manual effort, we present an automated learner. The learner can reverse engineer state and timing behavior of a legacy interface by examining event traces of the component in operation. The learner will then generate a ComMA model.**

I. Introduction

THE high-tech industry creates complex cyber-physical systems. The architectures for these systems consist of many hardware and software components. These components can be self-created or made by a third party supplier. Components interact with each other using software interfaces. Good interface descriptions are crucial for component-based development of cyber physical systems. Typically, however, software interfaces are only described in terms of their signature, i.e., the set of operations. Sometimes also the allowed sequence of operations is specified, for instance in terms of a state machine or a few example scenarios. The timing behavior of an interface is almost never described. For instance, the expected frequency of notifications and the allowed time between the call of an operation and the corresponding response. Violations of assumptions about timing behavior, however, are an important source of errors over the complete life cycle of these systems.

To overcome the drawbacks of current interface definitions, we have developed a Domain Specific Language (DSL), called ComMA as an abbreviation for Component Modeling and Analysis. ComMA [1] is currently used at the business unit Image Guided Therapy (IGT) of Philips for the formal definition of signature, state and timing behavior of software interfaces. ComMA specifies the signature of a server, i.e., the operations it offers to clients and the notifications it can send to clients. In addition, a ComMA interface definition includes a state machine which specifies the allowed interactions between client and server, timing constraints on sequences of operations, and data constraints on the parameters of operations.

Based on a ComMA specification, a large number of artifacts are generated automatically, for example:

- A visualization of state machine, timing and data constraints by means of plantUML[1].
- A Microsoft Word document according to the prescribed Philips template with the interface specification; this also uses comments in the ComMA specification including Doxygen-style comments[2].
- A simulator of the interface based on the state machine.
- Proxy source code in C++ and C# for the middleware technology SSCF of Philips IGT for transparent deployment of software components. SSCF is an abbreviation of Simple Service Communication Framework.
- A monitor which can be used to check whether an implementation of the interface conforms to the specification. This is done based on an execution trace that is recorded or sniffed during the usage of the implemented interface. The monitor checks conformance to the specified state machine and the timing and data constraints.

The monitor is very useful to check interface compliance after software updates or hardware upgrades. The monitor stores the timing information from the trace that is used to check the timing constraints. This information can be visualized to obtain insight in the timing characteristics. This is, for instance, useful when an updated hardware component is obtained from a supplier. Then the impact on the Philips part of the interface can be determined based on the differences between the characteristics of the old and the updated component.

Given the benefits of the ComMA approach, all new major system interfaces of Philips IGT are modeled and checked using ComMA. There are, however, hundreds of existing interfaces and it would be beneficial to apply the power of

[1]www.plantuml.com
[2]www.doxygen.org

the ComMA framework also to these interfaces. A manual transformation would require a large reverse engineering effort. Hence, the goal of the work described here is to support this transformation automatically such that the manual effort is reduced significantly.

Our approach is based on model learning techniques to obtain a first version of an interface state machine in ComMA. The main contribution is that we also learn the timing constraints. Since the learned interface may not be complete and states will not have meaningful names, manual changes will be needed. These changes are validated by the monitor generated by ComMA.

Concerning the model learning techniques, we have experimented earlier with active learning which stimulates the system under learning actively and infers an hypothesis based on the responses of the system [2]. Active learning requires the implementation of an adapter to connect the System Under Learning (SUL) with the learner. This adapter has to deal with behavior of the SUL that does not match the assumptions of the learning techniques, such as a SUL which is not input enabled or a SUL which sends no output or multiple outputs after a stimulus. This technique also requires frequent resets of the SUL which may be time consuming. Furthermore, nondeterminism of the SUL is a problem for active learning.

To avoid these issues, the approach described here is based on passive learning [3] where traces of SUL behavior are used to derive an hypothesis about the state behavior. Our algorithm is based on regular inference [4]. In particular, we use the algorithms described in [5], [6].

A disadvantage of passive learning is that only the behavior that is represented in the used traces will be in the resulting state machine. Hence, compared to the active learning approach, the model might be less complete. In our case, however, this is acceptable since the learned model is intended as a starting point for subsequent manual editing.

Related work

There are several model-based techniques to formally describe interfaces. Related to our approach is the Analytical Software Design (ASD) method [7] which includes formal interface specifications represented as state machines. An ASD interface model plays a similar role as a protocol state machine of UML [8]. An ASD interface not only describes the services offered by the server; it also specifies the operations allowed by the client. So it can be seen as a contract between client and server, similar to the Design by Contract approach [9]. Franca[3] is a related domain-specific language for the definition and transformation of interfaces.

All these approaches lack the ability to describe the timing aspects of the interface behavior and to check if an existing implementation conforms to an interface specification which includes timing constraints. Testing of real-time behavior by means of UPPAAL-TRON is described in [10]. In an industrial case, a timed automata model is obtained by first manually modeling the behavior of the system and next manually tightening the timing tolerances in an iterative ways using model-based testing.

An approach to obtain timing information of a component from execution traces is described in [11]. Models include worst case execution times of method calls. Downside of this approach is that the source code needs to be instrumented to acquire the execution traces and by doing so the timing behavior is influenced. In addition, only the time of a method call is captured, not the timing between events. In our approach the code does not need to be instrumented and timing between all event types is captured.

Structure of this paper

The paper is organized as follows. Section II provides a brief overview of the definition of interfaces in ComMA. Next we describe in Section III how an interface model can be obtained for an existing interface by manual editing. Automated support for reverse engineering of state behavior is presented in Section IV. Next, Sections V & VI, addresses the reverse engineering of state and timing behavior respectively. Results of experiments with our approach are presented in Section VII. Section VIII concludes the paper.

II. Model-based Definition of Interfaces

In this section, we introduce ComMA as far as needed to understand the remainder of this paper. The ComMA framework consists of the following four main languages:

- A language to describe the signature of an interface, see Section II-A.
- A language to capture observed interface interactions in the form of timed traces, see Section II-B.
- A language to describe the behavior of an interface, see Section II-C.
- A language to specify the generators to be used, see Section II-D.

The languages are illustrated by a test interface, called ITest, of a power control unit, see Figure 1. For the sake of explanation we made few modifications to the language instances. A predecessor of this unit has been introduced in [12].

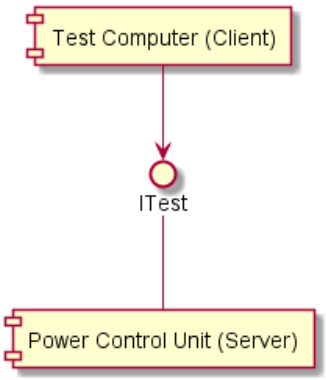


Fig. 1: Interface ITest of a power control unit

[3]franca.github.io/franca/

A. ComMA Signature

In a ComMA interface three types of operations are distinguished:

- *Commands* are synchronous operations from client to server. The client receives a *reply* from the server.
- *Signals* are asynchronous operations from client to server. Signals do not have a reply.
- *Notifications* are asynchronous operations from server to client. Notifications do not have a reply.

Listing 1 shows the signature of the ITest. First it defines two enumeration types, Stimulus and State. Next two commands are defined: 1) operation InjectStimulus with one parameter of type Stimulus; it replies a boolean value 2) operation GetState which replies a value of type State. Finally, the notification StateUpdate with one parameter State is defined. Observe that this example does not include any signal.

```
signature ITest {
  types
  enum Stimulus { VideoOnButton SystemOffButton ..}
  enum State { VideoOn VideoOnTransitioning SystemOff ..}

  commands
  bool InjectStimulus(Stimulus s)
  State GetState

  notifications
  StateUpdate(State state)
}
```

Listing 1: Example of a signature

B. ComMA Trace

The trace language is used to represent observed interface interactions. They can be, for instance, the captured network traffic or events written to a log file. An event is the occurrence of an operation. The language is independent of the technology used to record interactions; converters transform a technology-specific sequence of observed events to an instance of the ComMA trace language. An example of a ComMA trace is given in Listing 2. This example is based on an interface with the signature described in Listing 1. The listing shows two events, a command and its reply. Note that the time delta (in microseconds) between this event and its predecessor is denoted by "Timestamp" and the keyword "OK" indicates that this is a reply of the preceding command.

C. ComMA Interface

The behavior of an interface in terms of the allowed sequences of operations can be expressed in ComMA by the combination of a state machine and a number of constraints. The state machine describes the allowed order of the events between server and client. As an example, Listing 3 presents interface "ITest" which imports the signature of Listing 1.

Listing 3 shows the following:

- A variable "systemStateNotificationPending" is defined and initialized.
- The initial state is "SystemOff".

```
Timing: 1464181458.066471
Timestamp: 0.000000
src address: 192.168.32.1
dest address: 192.168.32.2
Interface: ITest
Command: InjectStimulus
Parameter: ITest::Stimulus : ITest::Stimulus::VideoOnButton

Timing: 1464181458.072651
Timestamp: 0.006180
src address: 192.168.32.2
dest address: 192.168.32.1
Interface: ITest
Command: InjectStimulus OK
Parameter: bool : true
```

Listing 2: Fragment of a ComMA trace

```
interface ITest{
  variables
  bool systemStateNotificationPending

  init
  systemStateNotificationPending := false

  initial state SystemOff {
    transition trigger: ITest::GetState do:
      reply(ITest::State::SystemOff)
      next state: SystemOff

    transition trigger: InjectStimulus(ITest::Stimulus s)
      guard: (s == ITest::Stimulus::VideoOnButton) do:
        systemStateNotificationPending := true
        reply(true)
      next state: VideoOnTransitioning

    ..
  }

  state VideoOnTransitioning {
    transition trigger: ITest::GetState do:
      reply(ITest::State::VideoOnTransitioning)
      next state: VideoOnTransitioning
    OR
      do: reply(ITest::State::VideoOn)
      next state: VideoOn

    transition guard: systemStateNotificationPending do:
      systemStateNotificationPending := false
      StateUpdate(ITest::State::VideoOnTransitioning)
      next state: VideoOnTransitioning

    ..
  }

  state VideoOn { .. }
}
```

Listing 3: Example of a ComMA state machine

- The first transition is triggered by the "GetState" operation. The replied state value is "SystemOff". This is a self-transition.
- The second transition is triggered by "InjectStimulus" with parameter "VideoOnButton". After replying value "true", the state machine transitions to state "VideoOnTransitioning".
- The second state is "VideoOnTransitioning".
- The first transition of this state is triggered by the "GetState" operation. The replied state value can be either "VideoOnTransitioning" or "VideoOn". This non-determinism is indicated with the "OR" keyword.

- In the second transition there is “StateUpdate” notification with parameter “VideoOnTransitioning”. Observe that this notification happens only once in the “VideoOnTransitioning” state which is coded by the “systemStateNotificationPending” variable.

Note that implicitly any behavior that is not defined in the state machine is not allowed.

In addition, a ComMA interface definition allows the specification of the timing behavior as a set of timing constraints. Listing 4 shows two examples of timing constraints:

- TimingRule0 describes the allowed time between an occurrence of command “GetState” and its reply. The Lower Specification Limit (LSL) is 2.4 ms and the Upper Specification Limit (USL) is 3.8 ms.
- TimingRule1 shows how constraints on more than two events can be grouped. It describes the allowed time between an “InjectStimulus” event and its reply, and the allowed timing between the reply and an occurrence of the “StateUpdate” notification.

```
timing constraints

TimingRule0
command ITest::GetState
  and reply(ITest::State::SystemOff)
  -> [ 2.4 ms ..  3.8 ms ] between events

group TimingRule1
command ITest::InjectStimulus(
    ITest::Stimulus::VideoOnButton)
  and reply(true)
  -> [ 5.9 ms ..  7.3 ms ] between events
  - [ 76.7 ms ..  165.3 ms ] -> notification
    ITest::StateUpdate(ITest::State::VideoOnButton)
end group
```

Listing 4: Example of a few timing constraints

Note that the ComMA trace of Listing 2 satisfies constraint TimingRule1, since the observed time delta between command and reply in this trace is approximately 6.2 ms which is between 5.9 ms and 7.3 ms.

D. ComMA Generator Specification

ComMA contains a separate language to specify which artifacts should be generated and it also allows the definition of parameters for these generators. An example is given in Listing 5 for a project called “Test” which imports the ITest interface. The project includes multiple generators:

- A “Monitor” to check if a ComMA trace conforms to the ComMA interface; in this case it takes file “Test.traces” as input.
- “SscfHeader”, is a generator that is explained in Section IV of this paper. The generator takes a “ITest.sscfheader” file as input.
- “Minedmodel”, is a generator that is explained in Sections V & VI. The generator takes a “Test.traces” file as input, excludes some parameters and filters out some unsolicited events as explained later.

```
Project Test {
  Compound Interface ITest {
    version
    ``1.0"

    description
    ``Demo project with Test component."
  }

  Generate Monitor {
    trace files
    ``Test.traces"
  }

  Generate SscfHeader {
    header files
    ``ITest.sscfheader"
  }

  Generate Minedmodel {
    trace files
    ``Test.traces"

    exclude parameters int string

    unsolicited events
    "dummyCMD2M"
    "dummyM2CMD"
  }
}
```

Listing 5: Example of a generator specification

III. Manual Reverse Engineering

Existing interfaces can be modeled manually in ComMA. This manual approach is depicted in Figure 2 and consists of the following steps:

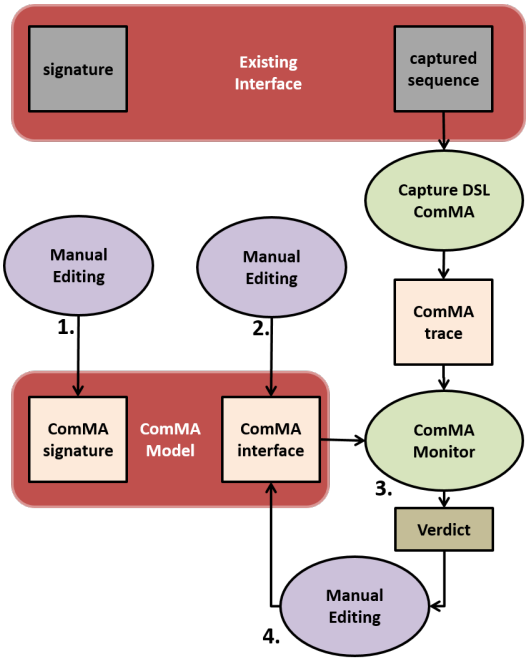


Fig. 2: Manual approach

1) The signature of the interface is defined manually.
2) A first version of the behavior of the interface is defined manually.

3) a) An interaction sequence of the existing interface is captured during execution or testing, for instance by sniffed network traffic or logging of events. From this recorded sequence, a ComMA trace has to be created. Typically this is done by a dedicated DSL.
   b) From the ComMA trace and the manually defined ComMA interface, a monitor is generated using the existing ComMA generator. With this monitor we check if the captured trace conforms to the defined ComMA model.

   In Figure 2, "Verdict" is the outcome of the interface conformance check.
4) The verdict of monitoring leads to three possibilities, assuming the used trace is correct:
   - Fail and the ComMA generator lists the issues; fix the issues in the model.
   - Pass; there are two options:
     - Done, the model captures all required behavior; the engineer has to decide this based on domain knowledge or, for instance, design documents.
     - Not done, extend the model with new behavior.

IV. Automated Reverse Engineering Support

In this section, we describe our reverse engineering approach. It can be seen as an extension of the manual approach presented in Section III, where we automate steps 1 and 2 of Figure 2. The automated approach is depicted in Figure 3.

The automated approach consists of the following steps:

1) We assume the signature of an existing interface is available in some representation. This can, for instance, be an IDL file in case of a COM interface or a header file using macros in C++ for another technology. The aim is to generate a ComMA signature from this representation. This requires a parser that accepts instances of an interface representation. Next a generator to generate a ComMA signature file has to be constructed.

   At Philips IGT, most signatures are available in the SSCF format. Hence, we created a DSL for the translation of a C++ header file with SSCF macros to a ComMA signature file. Listing 6 depicts an example of the SSCF interface description. From this example, the generator will automatically generate Listing 1. We do not discuss this DSL in more detail since the transformation is trivial for the Philips specific SSCF technology. The generator is called "SccfHeader" and requires an SSCF header file as input. Listing 5 show how this generator can be used.
2) Similar to step 3 of the manual approach, the behavior of a legacy interface is manifested by some sequence of events which are translated into a ComMA trace. In this case, the so-called ComMA Learner is used to construct a state machine and timing constraints based on one or more ComMA traces. Hence, we assume that the traces used in the learning are correct.

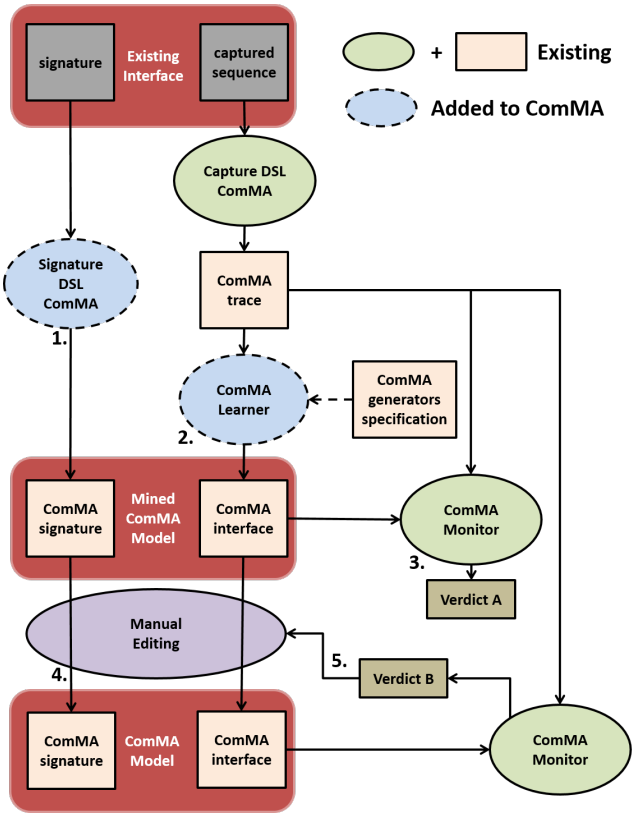


Fig. 3: Interface mining approach

```
SSCFTS1_BEGIN_INTERFACE(ITest)
  SSCFTS1_BEGIN_METHODS(ITest)
    SSCFTS1_INTERFACE_METHOD_1(bool, ITest,
      InjectStimulus, in(Stimulus))
    SSCFTS1_INTERFACE_METHOD_0(State, ITest, GetState)
  SSCFTS1_END_METHODS
  SSCFTS1_BEGIN_EVENTS(ITest)
    SSCFTS1_INTERFACE_EVENT_1(ITest, StateUpdate, State)
  SSCFTS1_END_EVENTS
SSCFTS1_END_INTERFACE
```

Listing 6: Fragment of an sscfHeader file

   a) The generation of a state machine by the ComMA Learner is described in Section V.
   b) The generation of timing constraints by the ComMA Learner is described in Section VI.

   Listing 5 shows how the ComMA Learner is called; the exclusion of parameters is explained in Section V.
3) Next, the existing generator of ComMA is used to generate a monitor and to check if the trace which is the starting point of step 2 indeed conforms to the learned interface. If the learner works correctly, the result should be a pass, so this is mainly a consistency check before continuing with the next steps.

The next two steps should be executed incrementally such that the changes on the model are small and can be easily reverted when they make the monitoring fail.

4) To create a more readable, complete and maintainable

version of the learned ComMA model, it is edited manually. For instance to add meaningful state names, reorder states, or to merge states and transitions.

5) As before, we use the generated monitor to check if the trace of step 2 still conforms to the edited ComMA interface. If not, the error has to be corrected, otherwise more changes can be made.

V. Learning State Behavior

In this section, we describe the learning of state machines. Figure 4 depicts the internal components of the Leaner. The "Serialize" component is used to format ComMA traces into a format which can serve as input for the "Algorithm" component. The "Deserialize" component converts the output of the "Algorithm" component into a ComMA interface.

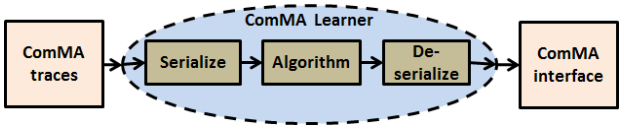


Fig. 4: Components of the learner

The general assumption is that the ComMA traces are correct, i.e., they represent valid behavior of the component.

A. Serialize

The "Serialize" component takes ComMA traces as input. It converts these traces into event strings. An event string starts with an interface name, followed by an event name, all parameter values, and finally the event type (command, reply, signal, or notification). Note that the conversion ignores all timing and address information in a trace.

B. Algorithm

The "Algorithm" component constructs a state machine based on the work described in [5], [6]. It uses a set of triggers, in our case *Commands* and *Signals*, and a set of listed actions, in our case *Replies* and *Notifications*. Triggers lead to transitions and action lists to states, following the pattern of a Moore machine where the output depends on the state only [13]. Based on one or more sequences of event strings, as a result of the previous component, the algorithm will construct a minimal (non-deterministic) finite state machine consistent with all input sequences. States with the same list of actions are merged, uniting the sets of their incoming and outgoing transitions. Note that this is different from (evidence-based) state merging [14] because the algorithm we use is linear and the resulting state machines might be non-deterministic.

C. Deserialize

The "Deserialize" component represents the output as a ComMA interface state machine. This means that the resulting Moore machine of the algorithm has to be transformed into a Mealy state machine where output depends on the state and the input trigger [15]. Moreover, a few restrictions on ComMA state machines have to be taken into account, such as limitations on the number of notifications on a transition. These restrictions are needed to enable the generation of monitors.

Listing 7 contains an example of a learned state machine for the "ITest" interface. Since the traces do not contain state information, the learned states are numbered. Notifications take place on transitions from a separate state with an underscore "_" in the state name. These states are added by the "Deserialize" component to meet the ComMA constraints mentioned in the previous paragraph. Observe the "OR" keyword which indicates that a reverse engineered state machine can be non-deterministic.

```
interface ITest{
  initial
  state s0 {
    transition trigger: ITest::InjectStimulus(
        ITest::Stimulus arg0)
      guard: (arg0 == ITest::Stimulus::VideoOnButton) do:
      reply(true)
      next state: s0_0_0
  }

  state s0_0_0 {
    transition do:
      ITest::StateUpdate(ITest::State::VideoOnTransitioning)
      next state: s1
  }

  state s1 {
    transition trigger: ITest::GetState do:
      reply(ITest::State::VideoOnTransitioning)
      next state: s1_0_0
        OR do:
      reply(ITest::State::VideoOnTransitioning)
      next state: s12
  }

  state s12 {
    transition trigger: ITest::GetState do:
      reply(ITest::State::VideoOn)
      next state: s13
  }
}
```

Listing 7: Example of a learned ComMA state machine

D. Tuning the Learner

The ComMA Learner can be tuned to ignore certain parameter values of events. For instance, an *int* parameter that acts like a cookie and is increased every transition might be excluded from the learning process. If we would not ignore the cookie, then the resulting state machine would become very large and restrictive. Hence, a new trace with different cookie values would not be accepted by the monitor. In such cases parameter values can to be ignored. Listing 5, shows how the parameters values for *int* and *string* are excluded from the learning algorithm. This means that the "Serialize" component does not include the *int* and *string* parameter values in the generated string.

VI. Learning Timing Constraints

In this section, we describe how we learn the timing constraints introduced in Section II. The timing constraints are

created during step 2b of our automated reverse engineering approach.

Our algorithm assumes that the client initiates the observed interface communication. Hence, observed events from the server are assumed to be triggered as a consequence of a *command* or a *signal* sent by the client. Clearly this holds for a *reply* event since it is caused by a *command* from the client. Our assumption means that a *notification* event is triggered by a *command* or *signal* from the client.

This pattern is used to avoid race-conditions by design. The latter is consistent with the solicit communication scheme in other approaches like ASD [16]. To avoid that *notification* events are triggered by unsolicited events, e.g. periodic alive events, unsolicited events can be filtered out of a trace by instrumenting the ComMA Learner as has been done for the "dummyCMD2M" and "dummyM2CM" events in Listing 5.

As shown in step 2 of Figure 3, the algorithm is fed with a trace of event observations. To learn timing characteristics, it is useful if the trace is long and contains many instances of events occurring in timing constraints. The algorithm performs the following steps on this trace:

1) Step 1 of the algorithm groups events according to the occurrence of trigger events of the client. Hence every event group starts with either a *command* or a *signal*. When in the trace the next event is a *command* or a *signal*, a new event group is created. Otherwise, the event is either a *reply* or a *notification* and it is added to the current event group. *Replies* and *notifications* have two attributes that represent minimum and maximum time differences with the previous event. These attributes are called LSL (Lower Specification Limit) and USL (Upper Specification Limit). In step 1 they are equal and initialized to the value of "Timestamp" of the event (note that this represents the delta time with the previous event). Figure 5 illustrates the event grouping. The output of this step is a list of event groups.

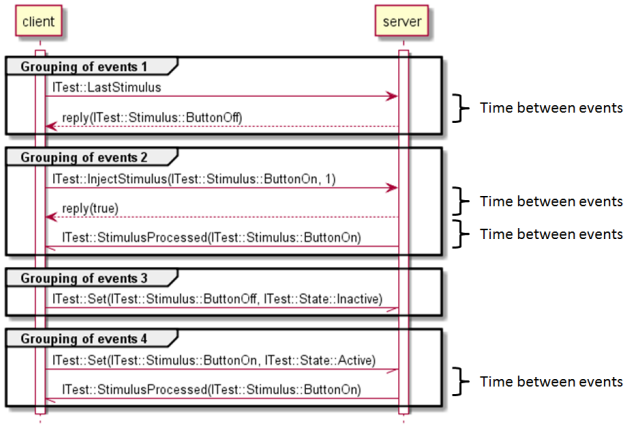


Fig. 5: Example trace timing

2) The list of the previous step will typically contain many groups that have the same events. For instance, many groups consisting of signal S and notification N. Only the time difference between S and N might be different. In step 2 the first occurrence of such groups is placed in a new list. Every event group in the new list will become a timing constraint.
3) Next the algorithm iterates over the event groups list of step 1 and matches every event group in it to a unique event group in the list of step 2. When a match is found, the LSL value of an event of the unique group is compared with the matched group. If the LSL value of an event of the matched group is smaller than the LSL of the unique group, then the unique group LSL value is updated with the value of the matched group. Likewise, the USL value of an event of the unique group is compared with the matched group. If the USL value of an event of the matched group is larger than the USL of the unique group, then the unique group USL value is updated with the value of the matched group.
4) Finally, the LSL and USL values of the unique event groups are used to create the timing constraints. The resulting constraints can be reviewed and the LSL and USL values can be relaxed manually in step 4 of our automated reverse engineering approach of Figure 3.

A more detailed formulation of this algorithm is given by the following methods. Method *CreateEventGroups* implements step 1 of the algorithm. It creates the event groups.

CreateEventGroups$(events)$ ::=
$group \leftarrow \emptyset$
$groups \leftarrow \emptyset$
FORALL $evt \in events$ **DO**
  IF $evt.type$ is Command **THEN**
    IF $group \neq \emptyset$ **THEN**
      $groups \leftarrow add(groups, group)$
      $group \leftarrow \emptyset$
    FI
    $group.trigger \leftarrow$ *Command*(evt)
    $previousEvt \leftarrow group.trigger$
  FI
  IF $evt.type$ is Signal **THEN**
    IF $group \neq \emptyset$ **THEN**
      $groups \leftarrow add(groups, group)$
      $group \leftarrow \emptyset$
    FI
    $group.trigger \leftarrow$ *Signal*(evt)
  FI
  IF $group \neq \emptyset$ **THEN**
    IF $evt.type$ is Reply **THEN**
      $action.string \leftarrow$ *Reply*$(evt, previousEvt)$
      $action.LSL \leftarrow evt.timestamp$
      $action.USL \leftarrow evt.timestamp$
      $group.actions \leftarrow add(group.actions, action)$
    FI
    IF $evt.type$ is Notification **THEN**
      $action.string \leftarrow$ *Notification*$(evt, previousEvt)$
      $action.LSL \leftarrow evt.timestamp$

$action.USL \leftarrow evt.timestamp$
$group.actions \leftarrow add(group.actions, action)$
FI
ELSE
Trace does not start with Signal or Command.
FI
OD
RETURN $groups$

As an example, consider the trace of Listing 2. Observe that the time difference between the command and its reply is described in the value after the Timestamp keyword of the reply. This time stamp is stored into the LSL and USL attributes of the reply event.

Next we present a helper method that is used in subsequent methods. Method *AreTheSameGroup* determines if two event groups are the same, that is, they have the same trigger and actions.

AreTheSameGroup($group0, group1$) ::=
$areTheSameGroup \leftarrow true$
IF $group0.trigger = group1.trigger$ **AND**
$group0.actions.size = group1.actions.size$ **THEN**
FORALL $i \leftarrow 0; i < group0.actions.size; i := i + 1$ **DO**
IF $group0.actions[i].name \neq$
$group1.actions[i].name$ **THEN**
$areTheSameGroup \leftarrow false$
FI
OD
ELSE
$areTheSameGroup \leftarrow$ *false*
FI
RETURN $areTheSameGroup$

The method *FindUniqueEventGroups* implements step 2 of the algorithm and returns a new list of unique event groups.

FindUniqueEventGroups($groups$) ::=
$uniqueGroups \leftarrow \emptyset$
FORALL $group \in groups$ **DO**
isUnique $\leftarrow true$
FORALL $group' \in uniqueGroups$ **DO**
IF *AreTheSameGroup*($group, group'$) **THEN**
isUnique $\leftarrow$ *false*
FI
OD
IF $isUnique$ **THEN**
$uniqueGroups \leftarrow uniqueGroups \cup group$
FI
OD
RETURN $uniqueGroups$

The method *DetermineTiming* implements step 3 of the algorithm. It takes the output of steps 1 and 2 as input and returns an updated unique groups list.

DetermineTiming($uniqueGroups, groups$) ::=
FORALL $group \in uniqueGroups$ **DO**
FORALL $group' \in groups$ **DO**
IF *AreTheSameGroup*($group, group'$) **THEN**
$action.LSL \leftarrow min(action.LSL, action'.LSL)$
$action.USL \leftarrow max(action.USL, action'.USL)$
FI
OD
OD
RETURN $uniqueGroups$

Using these methods, we create an algorithm to acquire timing constraints in the following way:
$groups =$*CreateEventGroups*($events$)
$groups_{uniq} =$ *FindUniqueEventGroups*($groups$)
$timingRules =$ *DetermineTiming*($groups_{uniq}, groups$)

As a last step, the timing rules are added to the interface file after the state behavior.

VII. Results

In this section, we present the results of our experiments and an analysis of the results.

A. Experiments

To validate the ComMA Learner we used two cases for which we already constructed an interface manually earlier: the power control unit and a third-party operating table [17]. For the power control case we use a trace called "Trace 1". For the operation table, two traces were used, called "Trace 2" and "Trace 3". The latter two traces are recordings of two different scenarios. Table I shows the characteristics of these traces by listing the number of commands, replies, signals and notifications, together with the types of the parameters.

We experimented with the ComMA Learner on the three traces and the exclusion of certain parameter types. The experimentation results are shown in Table II. In the last column, "Verified" refers to step 3 of the approach described in Figure 3, i.e., the monitoring; "yes" means that we could create a monitor and the verdict was that the trace is accepted by our generated monitor while "no" denotes that we could not generate a monitor because of the size of the state machine.

As explained in Section V-B, the algorithm can take more than one trace as input. Observe that learning based on Trace 2 and Trace 3 separately leads to 33 and 32 unique event groups, respectively, when excluding string and int. Using both traces leads to 55 groups, hence 10 groups are part of both traces. In the "Verified" column, "yes & yes" means that the monitor accepts both traces.

```
interface ITest {
  in all states {
    transition trigger: ITest::dummyCMD2M
    transition do: ITest::dummyM2CMD
  }

  initial
  state s0 { .. }
}
```

Listing 8: Example of a generated ComMA state machine with unsolicited operations

TABLE I: Characteristics of traces

	Command		Reply		Signal		Notification			
Trace	Nr.	Arg. Types	Nr.	Arg. Types	Nr.	Arg. Types	Nr.	Arg. Types	Events Total	Transitions Total
1	39	enum	39	enum bool	0	-	11	enum	89	39
2	0	-	0	-	2964	enum bool string int	2125	enum bool string int	5089	2963
3	0	-	0	-	915	enum bool string int	600	enum bool string int	1515	914

TABLE II: Results of learning experiments

Experiment			Learner Output				
Trace	Excl.	Unique Groups Nr.	Timing Rules Nr.	States	Transitions	Time (in ms)	Verified
1	-	14	10	21	30	17	yes
1	bool	14	10	21	30	6	yes
1	enum	4	0	9	14	3	yes
1	all	4	0	9	14	2	yes
2	-	689	19	2636	3324	342	no
2	string int	33	1	92	124	140	yes
3	-	202	30	615	816	2	yes
3	string int	32	2	88	119	3	yes
3	all	29	0	82	110	3	yes
2 & 3	string int	55	0	163	217	11	yes & yes
2 & 3	all	49	0	146	194	17	yes & yes

TABLE III: Results of second learning experiment with filtering of periodic events

Experiment			Learner Output				
Trace	Excl.	Unique Groups Nr.	Timing Rules Nr.	States	Transitions	Time (in ms)	Verified
2	-	676	19	1971	2637	7	no
2	string int	25	1	74	93	1	yes
3	-	191	30	468	649	1	yes
3	string int	26	2	76	96	2	yes
3	all	25	0	75	96	2	yes

The model of the third party operating table is very large and unreadable. The main reason is the number of operations and the fact that the system components periodically exchange keep-alive events. These periodic events become part of the action lists which increases the number of possible states significantly. Because of this we have improved the instrumentation of the ComMA Learner by filtering out periodic events from a trace.

As an example, Listing 5 specifies that the unsolicited events "dummyCMD2M" and "dummyM2CMD" have to be removed from the input trace. Then the generated state machine contains a part that allows the corresponding operations in all states. Listing 8 provides an example where "dummyCMD2M" is a *Signal* and "dummyM2CMD" a *Notification*.

Table III is an update of Table II where these two events are filtered from Traces 2 and 3. Observe that filtering reduces the number of states and transitions of the resulting model.

B. Analysis

When inspecting the learned models, we observed that the state machine for the power control case is quite readable. For this case, Listing 3 presents a fragment of the manually crafted model and Listing 7 presents a fragment of the generated model. Next we compare both state machines:

- States "s0" and "SystemOff" map because the VideoOnButton can be injected in this state. The "GetState" was not present in the observed trace and therefore not in the learned state machine
- The "VideoOnTransitioning" state in the manually crafted model is presented by the "s0_0_0" and "s1" states of the learned model. The learned state machine does not use state variables, but encodes this behavior in separate state. Observe that the learned model is more restrictive because "StateUpdate" needs to come before "GetState" while this is not required for the manual crafted model.
- States "s12" and "VideoOn" map because the GetState operation replies "VideoOn".

VIII. Concluding Remarks

We presented a manual and automated approach to reverse engineer existing legacy software interfaces. The benefit of the automated approach compared to the manual approach is that less manual labor is required for the creation of a ComMA model. Based on sequences of observed operations, a ComMA model is automatically generated that describes the external visible behavior of a software component in terms of its state and timing behavior.

We applied our approach on two cases for which we had

manually crafted ComMA models and traces available. In our experiments, the ComMA monitor generated from a learned model accepts all traces that were used to learn the model.

We observed that the learned state machines can become very large and restrictive. For example, when an operation has an integer as a parameter and the trace has many occurrences of this operation with many different values for the integer, then the Learner will create a transition for every different value. However, this parameter value might be irrelevant for the state behavior of the learned component. In such situations, it is desirable to exclude integer values from the state machine learner and we instrumented the learner to allow this.

A general strategy could be to first learn a state machine without excluding any parameters and then incrementally exclude parameter types until the resulting state machine is manageable. The final step then would be a manual editing of the state machine.

With our approach the quality of the traces is very important. All behavior that is not in the input traces will not be in the resulting model.

In the future, we will apply our approach on legacy interfaces for which we do not have a manually crafted model.

Acknowledgments

We are grateful to the anonymous reviewers for useful comments and suggestions.

REFERENCES

[1] I. Kurtev, M. Schuts, J. Hooman, and D.-J. Swagerman, "Integrating interface modeling and analysis in an industrial setting," in *MODELSWARD*, 2017. doi: http://dx.doi.org/10.5220/0006133103450352 pp. 345–352.

[2] F. Vaandrager, "Model learning," *Commun. ACM*, vol. 60, no. 2, pp. 86–95, 2017. doi: http://dx.doi.org/10.1145/2967606

[3] W. van der Aalst, *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[4] T. Berg, B. Jonsson, and H. Raffelt, "Regular inference for state machines with parameters," in *International Conference on Fundamental Approaches to Software Engineering*. Springer, 2006. doi: http://dx.doi.org/10.1007/11693017_10 pp. 107–121.

[5] I. Buzhinsky and V. Vyatkin, "Modular plant model synthesis from behavior traces and temporal properties," in *22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2017. doi: http://dx.doi.org/10.1109/ETFA.2017.8247578 pp. 1–7.

[6] ——, "Automatic inference of finite-state plant models from traces and temporal properties," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1521–1530, 2017. doi: http://dx.doi.org/10.1109/TII.2017.2670146

[7] G. H. Broadfoot, "ASD case notes: Costs and benefits of applying formal methods to industrial control software," in *Formal Methods (FM 2005)*, J. Fitzgerald, I. J. Hayes, and A. Tarlecki, Eds. Springer, 2005. doi: http://doi.org/10.1007/11526841_39 pp. 548–551.

[8] G. Booch, J. E. Rumbaugh, and I. Jacobson, *The Unified Modeling Language user guide*, ser. Addison-Wesley object technology series. Addison-Wesley-Longman, 1999.

[9] B. Meyer, "Applying "design by contract"," *IEEE Computer*, vol. 25, no. 10, pp. 40–51, 1992. doi: http://dx.doi.org/10.1109/2.161279

[10] K. G. Larsen, M. Mikucionis, B. Nielsen, and A. Skou, "Testing real-time embedded software using UPPAAL-TRON: An industrial case study," in *Proceedings of the 5th ACM International Conference on Embedded Software*, ser. EMSOFT '05. ACM, 2005. doi: http://dx.doi.org/10.1145/1086228.1086283 pp. 299–306.

[11] Z. Gu, S. Wang, S. Kodase, and K. G. Shin, "An end-to-end tool chain for multi-view modeling and analysis of avionics mission computing software," in *24th IEEE Real-Time Systems Symposium, 2003 (RTSS 2003)*, 2003. doi: http://dx.doi.org/10.1109/REAL.2003.1253256 pp. 78–81.

[12] M. Schuts and J. Hooman, "Using domain specific languages to improve the development of a power control unit," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015. doi: http://dx.doi.org/10.15439/2015F46 pp. 781–788.

[13] E. F. Moore, "Gedanken-experiments on sequential machines," *Automata Studies, Annals of Math. Studies*, 1956.

[14] M. J. Heule and S. Verwer, "Software model synthesis using satisfiability solvers," *Empirical Software Engineering*, vol. 18, no. 4, pp. 825–856, 2013.

[15] G. H. Mealy, "A method for synthesizing sequential circuits," *Bell System Technical Journal*, no. 34, p. 1045–1079, 1955. doi: http://dx.doi.org/10.1002/j.1538-7305.1955.tb03788.x

[16] J. Hooman, R. Huis in 't Veld, and M. Schuts, "Experiences with a compositional model checker in the healthcare domain," in *Foundations of Health Information Engineering and Systems (FHIES 2011)*, LNCS 7151. Springer, 2012. doi: http://dx.doi.org/10.1007/978-3-642-32355-3_6 pp. 93–110.

[17] I. Kurtev, J. Hooman, and M. Schuts, "Runtime monitoring based on interface specifications," in *ModelEd, TestEd, TrustEd*. Springer, 2017. doi: http://dx.doi.org/10.1007/978-3-319-68270-9_17 pp. 335–356.

6th Conference on Multimedia, Interaction, Design and Innovation

MIDI Conference provides an interdisciplinary forum for academics, designers and practitioners to discuss the challenges and opportunities for enriching human interaction with digital products and services.

The main focus of MIDI Conference is exploring design methods for creating novel human-system interaction, developing user interfaces and implementing innovations in user-centred development of advanced IT systems and on-line services.

Topics

Topics of interest include (but are not limited to) the following areas:

- interactive multimedia and multimodal interaction design
- novel interaction techniques, voice interfaces, interactive multimedia
- ubiquitous, multimodal, pervasive and mobile interaction, wearable computing
- novel information visualization and presentation techniques, Augmented/Virtual Reality
- design methods for usability, accessibility and outstanding user experience
- prototyping of user interfaces and interactive services
- human-centred design practices, methods and tools, user interface design
- unfolding trends in HCI research and practice, customer experience, Service Design
- advances in user-centred interaction design
- understanding people and interactions: theory, concepts, models and methods
- understanding people and interactions: contextual, ethnographical and field studies
- critique and evolution of methods, processes, theories and tools for human-computer interaction
- novel methodologies for conceptualization, design and evaluation of interactive products and services

Event Chairs

- **Marasek, Krzysztof,** Polish-Japanese Academy of Information Technology, Poland
- **Romanowski, Andrzej,** Lodz University of Technology, Poland
- **Sikorski, Marcin,** Polish-Japanese Academy of Information Technology, and Gdansk University of Technology, Poland

Program Committee

- **Biele, Cezary,** Information Processing Institute, Poland
- **Brocki, Łukasz,** Polish-Japanese Academy of Information Technology
- **Forbrig, Peter,** University of Rostock
- **Grudzień, Krzysztof**
- **Guttormsen, Sissel,** University of Bern, Institute of Medical Education, Switzerland
- **Kaptelinin, Victor,** Umea University
- **Koržinek, Danijel,** Polish-Japanese Academy of Information Technology, Poland
- **Kołakowska, Agata,** Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Landowska, Agnieszka,** Gdansk University of Technology, Poland
- **Marti, Patrizia,** University of Siena, Italy
- **Masoodian, Masood,** Aalto University
- **Miler, Jakub,** Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Pribeanu, Costin,** National Institute for Research and Development in Informatics - ICI Bucuresti
- **Satalecka, Ewa,** Polish-Japanese Academy of Information Technology
- **Slavik, Pavel,** Czech Technical University
- **Wichrowski, Marcin,** Polish-Japanese Academy of Information Technology, Poland
- **Wieczorkowska, Alicja,** Polish-Japanese Academy of Information Technology, Poland
- **Winkler, Marco,** University Paul Sabatier
- **Wojciechowski, Adam,** Institute of Inf. Techn., Lodz Univ. of Techn.
- **Woźniak, Paweł W.,** University of Stuttgart, Germany
- **Wołk, Krzysztof,** Polish-Japanese Academy of Information Technology, Poland
- **Ziegler, Juergen,** University of Duisburg-Essen

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 879–886
DOI: 10.15439/2018F298
ISSN 2300-5963 ACSIS, Vol. 15

Optimizing the Number of Bluetooth Beacons with Proximity Approach at Decision Points for Intermodal Navigation of Blind Pedestrians

Jakub Berka, Jan Balata, Zdenek Mikovec
Faculty of Electrical Engineering,
Czech Technical University in Prague
Prague, Czech Republic
berkajak@fel.cvut.cz, balatjan@fel.cvut.cz, xmikovec@fel.cvut.cz

***Abstract*—Navigation in urban environments is very challenging for blind pedestrians. Although many navigation approaches using various principles or sensors to help visually impaired people exists nowadays they still have problems to navigate in complex buildings, find entrances to buildings or to navigate to correct public transport stops. Current solutions use a large number of sensor that needs to be installed in the environment needed to track every single move of the user. We present a solution to reduce the number of installed sensors by using previously developed set of landmark-enhanced navigation instructions allowing us to lower the necessary number of Bluetooth beacons by using them only for proximity notification at indoor decision points, indicating public transport station and entrances. The evaluation in the field study (N = 8) suggests a good potential of the approach, especially in terms of usability, recovery from going astray and beacon deployment cost. Further, we provide guidance on beacons placement in the environment.**

***Index Terms*—BLE, beacons, intermodal, navigation, blind pedestrians**

I. Introduction

FOR the visually impaired people, it is vitally important to be able to travel independently and freely. The limitation in travel related activities has a negative impact on their quality of life and can result in worsening psychical condition and low self-esteem [1].

Many electronic devices, navigation aids, and navigation systems are now widely available for blind pedestrians. They are based on various principles of positioning such as Global Satellite Navigation System (e.g. GPS[1], GLONASS[2]) based systems for outdoor navigation (Blind Square, Ariadne GPS, Kapten Mobility); Bluetooth Low Energy beacons (NavCog), RFID[3] readers [2], or cameras [3] for indoor navigation. However, positional error for GPS in a city is about 28 meters for 95% of the time [4]. Similarly, indoor positioning systems often require high deployment costs and are not suitable to be used in other than indoor environments moreover, there is no global standard for indoor navigation systems yet.

In this paper, we propose a method for decreasing the number of Bluetooth beacons used for indoor positioning. We have identified various places on typical routes where the beacons are necessary because of complicated navigation and orientation, e.g. entrances to buildings, public transport stations, open spaces/areas (hallways, courtyards) and indoor decision points (location, where navigation instruction is needed because of multiple routes, can be taken such as corridor crossing). In the rest of the routes, we replace the beacons with landmark-enhanced navigation instructions, which provide necessary guidance when traveling. Instead of continuous location tracking, we use Bluetooth beacons for proximity estimation.

[1]Global Positioning System
[2]Global Navigation Satellite System
[3]Radio-frequency identification

II. Related Work

In this section, we describe approaches related to indoor and outdoor navigation of blind pedestrians. Furthermore, we focus on methods that use Bluetooth Low Energy beacons for navigation or proximity estimation.

A. Pedestrian navigation of blind

Successful navigation depends on the spatial knowledge about the environment. There are three levels of environment knowledge applied for navigation in cities: knowledge of landmarks, knowledge of route and overview knowledge [5]. Landmarks can be defined in various ways as says Golledge in [6], a landmark is something capable of attracting attention, i.e. it has dominance visible form and stands out from the surrounding environment. For navigating visually impaired people outdoors the suitable landmarks can be e.g. street corners and their different shape, pedestrian crossings, steps, etc. For indoor navigation visually impaired people still use the landmarks because the three levels of environment knowledge are valid indoors as well, only the characteristic of landmarks change (e.g. door, stairs, type of rooms).

B. Bluetooth Localization

Work of Gorovyi et al. [7] shows the application of beacons for real-time users positioning based on trilateration calculation using Received Signal Strength Indicator (RSSI) values from three or more beacons. Performed accuracy test showed

that beacon calibration improves system efficiency (1-2 meters in their case).

The Bluetooth technology used for indoor localization is often combined with utilization of other sensors to improve the accuracy of the navigation in an indoor environment (Accelerometer, Barometric sensor). Czogalla and Naumann [8] developed indoor navigation for 8000 m^2 public indoor environment with 35 beacons installed. The route is presented to users by visual map and directions by vocal instructions.

Commercial solutions for indoor location or positioning based on the beacons often use a triangulation/trilateration approach such as Indoo.rs[4]. The difficulty of these solutions is mainly in a large number of beacons required. To achieve their proposed accuracy 1-3 meters it is necessary to install beacon every 7-10 meters, e.g., with Infsoft indoor navigation[5].

C. Indoor Navigation for Blind

Work of Ahmetovic et al. [9] resulted in *NavCog* system, which relies on Bluetooth beacons installed in an environment and provides sub-meter precise localization with a minimum of 1 beacon every 6 meters and navigation assistance for people with visual impairments. They achieve it by representing the environment in the one-dimensional graph which results in lower number of beacons to be used and further, they use multi-modal probabilistic state estimation algorithm and Particle Filtering framework to more precisely estimate user's position. *NavCog* system gives to user the "turn-by-turn" metric navigation instructions, distance announcements inform the user about the distance to next action (e.g. "18 meters"), action instructions give information about turning direction or transit information (e.g. move between floors). It also provides accessibility instruction (e.g. if there is a curb that is easy to trail with a cane) or surrounding information (e.g. building description) on request.

To help visually impaired children in school to move and play independently Freeman et al. [10] used Audible Beacons as wearable bracelets that support wireless communication and provide audio output. Beacons are also placed in the school environment. They presented various scenarios based on their solution. Beacon bracelets can inform children about their nearby points of interest, by playing a specific sound of this POI[6]. As the children get closer to the POI the sound is played louder. It can help to find sighted friends, who are wearing bracelets as well. Bracelets and beacons placed in the environment can help to learn the layout of the school including entrances.

Guo et al. [11] developed Landmark-based Mapless Indoor Navigation called FreeNavi that requires only WiFi fingerprints collected on the device. This system applies knowledge of humans being able to navigate through and identify the environment by landmarks. FreeNavi constructs a virtual map only by landmarks descriptions and their connectivity relations. Virtual map construction algorithm is based on WiFi signal strength data and also landmark fingerprints and the user traces, this data is crowdsourced and then the map created. The generated map does not contain information of turning directions (left or right), i.e. users have to find out by themselves at the junctions in which direction they have to continue to next landmark.

Finding entrances to the desired building was subject of researchers focused on crowdsourcing. The study says that almost 65% of blind and visually impaired people suffered the mobility hindrance of hard to find entrances in the international survey by Zeng and Weber [12]. To solve this issue authors used collaborative method for collecting information about entrances and buildings and they also created reference point for each entrance, their concept does not use GPS data but it is expected to use some GPS-base tool to navigate to reference point, when user approaches the entrance s/he gets structured and also unstructured collected information by other users. [12].

Our proposed solution is based on automatically generated landmark-enhanced navigation instructions for outdoor [13], indoor [14] environments and their combination for different environments transitions and usage of public transportation [15]. For automatic generation of the navigation instructions it is necessary to use specially modified GIS which contains information about sidewalk network and special features of the urban environment (slope, surface quality, railings, corners) as it is presented in [13]. In our work, we aim to address the issue of finding the entrances [12], as well as a large number of installed Bluetooth beacons [9] by using them only for notification of progress at decision points without trilateration.

III. Navigation Application Prototype

We developed a prototype of the navigation application which provides navigation instructions for outdoor, public transport and indoor navigation for blind pedestrians. The transition between different modes are seamless – at the transition from one mode/environment to the other, the application provides summarized description of the mode/environment such as the description of the public transport lines and stations or description of the building. In selected locations, the application provides notification about user's progress on the route triggered by Bluetooth beacons.

A. Route itinerary

The route itinerary contains detailed navigation instructions for each segment of the route, which consists of a description of the surrounding environment and action to be performed by the user. Construction of navigation instructions is based on the sidewalk-based GIS[7] for outdoor environments [13], template system for public transport stops and stations and environment transitions (e.g., from outdoor to indoor) [15] and landmark-enhanced navigation instructions for indoor [14]. Navigation instructions are stored in the navigation application in the graph data structure (see Fig. 1 and Tables I, II).

[4]Indoo.rs – https://indoo.rs/

[5]Infsoft – https://www.infsoft.com/

[6]Point of Interest

[7]Geographic Information System

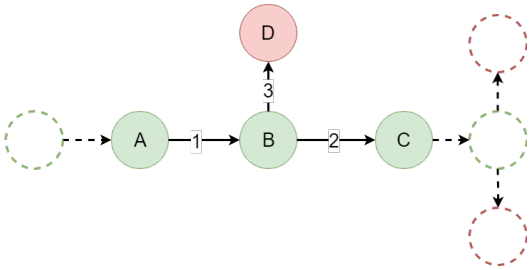


Fig. 1. Example of the generated graph structure. The detailed information about the content of the nodes and edges are presented in Tables I, II.

TABLE I
EXAMPLE OF THE NODES DATA IN GRAPH DATA STRUCTURE.

Node	isOnRoute	BeaconID
A	true	undefined
B	true	ODUM
C	true	vf3i
D	false	4vXL

B. Beacons Installation

The navigation application uses RSSI from beacons to estimate its proximity. There are several factors influencing the signal transmission: the signal can be absorbed by other persons moving nearby beacons; the signal can be interfered by other devices working on the same frequency spectrum as beacons (2.4 GHz, same spectrum used by Wi-Fi 802.11); many materials can act as barriers for Bluetooth signal such as plaster, concrete, bulletproof glass, or metal materials[8]. These limitations have to be taken into account during beacon installation.

In the indoor environment, the beacons are installed only at the decision points – corridor junctions, corridor bents, floor mezzanines and complicated stair system. We placed the beacons mainly on the walls 2.5 – 3.5 meters above ground (see Fig. 2).

At public transport stations, the suggested place for installing the beacon is the info-table. Same as in indoor environment the beacon should be placed above the people's heads. In our case, we placed it at the tram station oriented towards the sidewalk.

To help users with environment transitions we placed the beacon near the building entrances. The beacon should be placed with respect to possible directions of the user's approaches to the entrance. If possible, as much as possible in the sidewalk level. When placing the beacon at the entrance our intention was first to slow down or stop the user. Second to let the user read the detailed description of the entrance from route instructions and find the correct door.

Moreover, we placed the beacon at the decision point in the semi-outdoor environment (university campus courtyard). This environment is composed of the roadways rather than sidewalks, and the users have to navigate through open spaces.

[8]Apple – https://support.apple.com/en-us/HT201542

Fig. 2. Beacons installation at the suitable places.

To help them find the decision point at the roadway turns, we placed the beacon on the building near it at 3,5 - 4 m high.

After the beacons installation, we configured beacons to avoid interferences of two or more beacons in mind and adjusted advertising interval and transmission power to ensure sufficient signal coverage. Finally, we collected RSSI to determine the thresholds to trigger location about the progress on the route.

C. Route Progress Notifications

The route is represented as an oriented graph.

The graph has two types of the nodes, *correct* and *error*. The correct nodes represent the decision points on the generated route. The *error* nodes currently used only for indoor segments represent the wrong turns off the route. E.g. if the route contains junction of corridors with one correct turn-off and two wrong turn-offs the graph will have one *correct* node and two *error* nodes. If there is a beacon installed at a particular decision point, its ID is stored in the node. For outdoor segments of the route, we store IDs only for nodes at entrances and public transport stops.

The edge represents a route segment. Edges on the correct route hold data about the segment number and the navigation instruction. Edges off the correct route lead to an error node. The part of the route graph we created for the evaluation purposes can be seen in Fig. 3.

We use the proximity-based approach – beacons serve as proximity traps at decision points, therefore the number of installed beacons is highly reduced. Route graph representation enables us to provide users with error prevention and also error recovery at the more complicated decision points – when the user goes astray s/he is notified and can use backtracking to get back on the route. Beacons ID are tied to navigation instruction displayed on a screen of the phone.

D. User Interaction

The navigation application was implemented in multiplatform Ionic framework version 3. The user interface provides controls for navigating between individual navigation instruction and for manual location verification. The user can

TABLE II
EXAMPLE OF THE EDGES DATA IN GENERATED GRAPH DATA STRUCTURE.

Edge	isOnRoute	Segment Nr.	Segment Description	Segment Action
1	true	1	You are at the turning of the road, the building E is in front of you.	Turn left and go approximately 30 meters through open space to pyramidal stairs, by your right hand. There is an entrance to the building E.
2	true	2	You are by the entrance to the building E. Above the pyramidal stairs there is big wooden door and glass swing door right behind, leading inside the building.	Go up the stairs and through the doors inside the building.
3	false	undefined	undefined	You are on the wrong way, return back to the beginning of the segment.

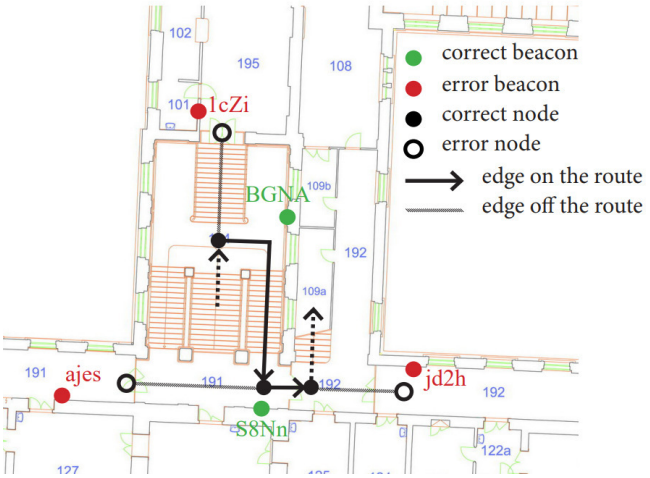


Fig. 3. Part of the experimental route represented as an oriented graph. With positions of installed beacons.

navigate between individual navigation instruction by pressing the buttons "Next Segment" and "Previous Segment".

When the user approaches the correct beacon, the application automatically notifies the user by vibration and text *"You are near the end of the segment."* Similarly, when the user turns off the route and approaches in the nearby of the error beacon from an actual segment, the application will four times short vibrate and announce to user *"You are on the wrong way, return to the start of this segment."*

These two notifications from application happen automatically when the user is in the vicinity of the corresponding beacon. However, the application also provides the possibility to manually verify user's position. When the user presses "Verify location" button the 5 seconds timeout starts and application is ranging for correct or error beacons. When the 5 seconds limit expires and the user is not in the nearby of any beacon application will announce *"Verification of your location was not successful."*

IV. EVALUATION

The navigation application supported one experimental route with an outdoor-indoor combination and usage of public transportation. Navigation instructions were automatically generated for outdoor (using Naviterier[9]) and manually prepared for the indoor, public transport and environment transitions.

A. Participants

Eight visually impaired participants were recruited via email leaflet (in the leaflet we mentioned that touchscreen smartphone will be used for the experiment). They were aged from 33 to 53 years ($mean = 40.25$, $SD = 7.27$). Five participants were congenitally blind and 3 were late blind. Three participants had Category 4 visual impairment (light perception) and 5 participants had Category 5 (no light perception) [16]. One participant had a guiding dog. All of the participants were native Czech speakers.

B. Apparatus

Route. The experimental route was located in city centre of Prague, Czech Republic with the combination of outdoor, semi-outdoor and indoor environments and use of public transportation (ride with tram).

It was approximately 700 m long (excluding tram ride), consisted of 36 segments. Twenty six Beacons Pro and 1 Tough Beacon[10] were installed on the route in total, 16 beacons were placed on the route decision points and 11 beacons were placed on the turnoffs from the route. The beacons for public transport station and at the first entrance were held by 2^{nd} experimenter glued on the paper folder above the head (as we did not get permission for long term placement).

Equipment. The participants were equipped with HTC One 801n smart-phone – Android 5.0.2, with running Talk Back screen reader set to Czech language and with the installed navigation application. The smart-phone had a lanyard that participant could hang on his/her neck, to have free hands when necessary. We also gave participants chip card that is needed to open a door inside the university campus buildings.

Data Collection. During each route walk-through, we were shadowing participants and recording third-person video. The navigation application records a log file containing data with timestamps about the interaction with the application, i.e. button presses, location notifications (triggered automatically)

[9]Naviterier – https://www.naviterier.cz/en
[10]Kontakt.io – https://kontakt.io/

Fig. 4. Photos from tram station and entrance to first building, show how the second experimenter hold the beacon.

and location verification requests (triggered by the user), and data about ranged beacons during the walk-through, with information about beacon ID and current RSSI.

C. Procedure

The experiment lasted around 1 hour. At first, we explained to the participant the purpose of the experiment, collected the demographic information and explained the operation and control of the application.

Then we accompanied the participant to the start of the route and we explained the task: *"Imagine, you were invited to participate in a study at the University Campus in a building E, room 317. To reach the destination use the navigation application, which will give you navigation instructions and it will notify you whether you are going the correct way at public transport stops and at indoor locations where you will also be notified if you go astray."*. They were instructed to proceed as if they were alone. However, whenever they felt unsafe, they should ask for our assistance.

After the participant finished the route, s/he was debriefed and asked about subjective judgement about the level of safety (*"I felt safe during the route walkthrough."*), efficiency (*"I think that thanks to application and navigation instructions I proceeded efficiently."*), information sufficiency (*"The Information I received from application was sufficient."*), comprehension (*"The instructions in decision points of the route were comprehensive."*) and confidence when finding entrances (*"I felt confident thanks to application when finding entrances to buildings."*) and tram station (*"I felt confident thanks to application when finding tram station."*) on a 5 point Likert scale as a level of agreeing with presented statements.

V. Results

All participants successfully completed the route. The average completion time was 44.2 minutes (SD = 8.7 minutes).

A. Tram station

Participants P1, P6, P8 found the tram station successfully, the beacon triggered the automatic notification.

Further, participants P2 and P3 misheard the notification from the application when finding the tram station, but the manual verification of location afterward was successful and helped then find the tram station. P7 also misheard the notification from the application, missed the info table at tram station, and continued further on the sidewalk (went astray) until we stopped him and returned him to the route.

Participants P5 and P8 confused the telephone booth, which was about 20 meters before the tram station with the info table of the tram station. They both tried to verify the location there and the application correctly responded. Afterward, P5 was able to find the tram station without the location verification as he switched to next segment early and P8 found it successfully with help from automatic notification from the navigation application.

Participant P6 was impressed by the automatic notification at the tram station, he said: *"As a blind when I am finding the tram station I have to walk near the building due to public notice so it is hard to find the station at the sidewalk, the automatic notification is very helpful."* P8 had the same opinion about stations: *"It is not necessary to receive the notifications everywhere but at tram station or near the entrances it is very good."*

B. Semi-outdoor.

At the courtyard, P1 did not follow the curb curved to the right and he continued in a straight direction. He complained about the missing information about curved curb. P3 and P4 skipped to the next segment before they reached the turn, but they continued without problems. Participants P6 and P7 had trouble with the manual location verification as they tried it few meters after they passed the beacon. P7 did not receive the automatic notification when he was near the beacon on the road turn however he continued without problems. Rest of the participants did not have problems when walking through the university courtyard and received the location notification.

After the experiment P7 explained that he lacked information about the distance and direction to the beacon: *"I would welcome the information about the distance to the decision point, now I am not sure if it is in front or behind me."*

C. Entrances

Entrance to building A (in the second niche of the protruding facade): Participants P1, P3, P4, P5 and P6 found the correct entrance to the first building. P3 found the entrance without the automatic notification but afterward verified the location manually twice and twice it was successful. P2 missed the entrance to the first building, the beacon triggered late, after a while we stopped him and returned him back to the route. P7 and P8 missed the entrance to the first building, the beacon did not trigger. P7 tried to verify location manually in the first niche, but the beacon did not trigger.

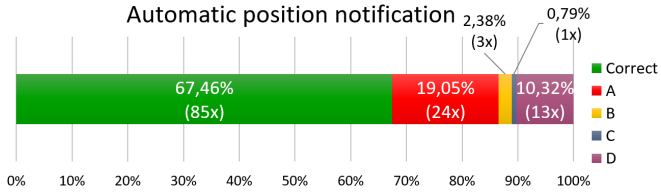


Fig. 5. Results of automatic notifications at decision points (total=126).

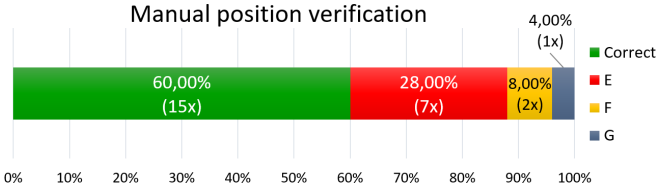


Fig. 6. Results of the manual position verification (total=25).

Entrance to building E (with pyramidal stairs): Participants P1, P2, P5 and P7 found the entrance with help of automatic notification from the application. P3 found the entrance but skipped to the next segment too early before the beacon can trigger. P4 found the entrance but the beacon did not trigger. P6 was proceeding to the entrance along the building and so the signal from beacon was shielded by the building, but participant found the entrance correctly. P8 stopped near the entrance when he got the notification at the right moment, but then continued straight and missed the stairs on the right, because he was proceeding on the very left side of the road during the segment leading to entrance stairs.

D. Indoor

Only P2 (once) and P4 (twice) went astray during the walk-through indoors. The error beacons triggered correctly and participants were notified that they should return to the start of the segment. P2 recovered from the error successfully alone. P4 at first recovered with an assistance and on the second error turn recovered alone successfully. P1, P3, P5, and P8 needed an assistance when finding the chip card reader next to the door. All participants but P5 found the correct door in the final segment.

Participant P6 mentioned that *"The vibrations and notifications were very accurate, it vibrated where it should for example at the last step of the stairs."* P3 noted that it would be beneficial at public transport stations and building entrances and in large buildings *"It would be great to have it for example in big hospital compounds."* Participant P2 was not satisfied with automatic notification at several indoor parts of the route: *"... in atrium it gave me the notification too early."*

E. Beacons

The automatic notification about the user position near the beacon was expected to happen 126 times in total during the whole experiment with 8 participants. Not always was the notification triggered as expected. In Figure 5 we present the results of the automatic triggering. The reasons why the automatic notification failed are various, e.g. signal interference, bad configuration of a beacon. Bellow, we present the detailed list of the different types of failures.

A: Notification did not trigger, near the beacon. Due to signal distortion or interference, bad configuration of the beacon or the signal covered by participant's body when turned away from the beacon. In two cases the users were behind the thick wall which absorbed the beacon signal.

B: Notification triggered too early, on the beginning of the actual segment. It happened during the short segments indoors. The distance it should trigger was to approximately 4 meters, but it triggered 8-10 meters away. The signal could be mirrored by a metal surface.

C: Notification triggered too late, when the participants just passed the end of the segment, therefore, they missed the decision point.

D: The application is ranging only for beacons in particular segment displayed on the screen. For more information see Section III-D. Some of the participants were reading the segments in advance and did not return to the actual segment in the application. We do not evaluate this situation as a failure.

The manual verification was used by participants 25 times in total during the experiment. Some participants tried to verify their location after they received the automatic notification and continued few steps ahead. The manual verification then could fail, because they stepped out of the beacon's range, or they covered the signal with their body. In Figure 6 we present the results of the manual position verification. Bellow is the list of various types of the manual verification failures.

E: Manual verification did not trigger the notification, near the beacon. Due to the signal distortion or interference, bad configuration of the beacon or the signal covered by participant's body when turned away from the beacon.

F: Manual verification failed, after the correct automatic notification. Participants stepped out of the range of the beacon, or the signal was covered by their own body.

G: Manual verification was successful on the second attempt.

F. Subjective judgement

Fig. 7 shows that 50% of the participants strongly agreed on information sufficiency they obtain from the application, one participant disagreed. 63% of the participants strongly agreed on comprehension at the synchronization points, one participant disagreed. 38% of the participants agreed about confidence when finding entrances, 2 participants disagreed. 63% of the participants agreed about the confidence when finding the tram station, 2 participants disagreed. 25% strongly agreed on safety and efficiency, one participant disagreed on safety and one disagreed on efficiency.

G. Discussion

The results clearly show that correct setting of the beacons is crucial for the proper functioning of this navigation system.

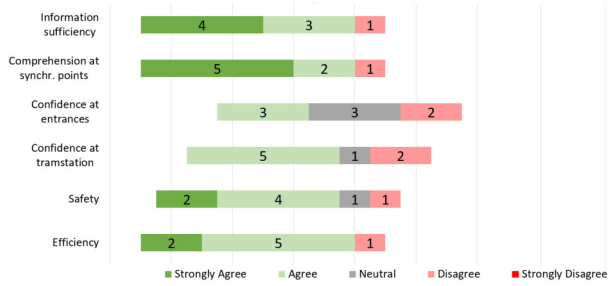


Fig. 7. Subjective judgements about level of information sufficiency, comprehension, confidence, safety and efficiency (N=8).

When placing a beacon at the entrance to a building, it should be remembered that users can come from different directions and that all of them should be covered by the signal. If it is a proportionally complicated entrance, it can be considerate to place a larger number of beacons to cover user's arrivals from all directions. It is not necessary to have beacons installed on each segment, sometimes it is not even physically possible, specifically on very short segments there may be a problem that the correct and error beacon will be very close to each other, which may result in an error location notification to the user even he is near the correct beacon or vice versa.

We experienced that there can be spaces where it is really difficult to correctly configure the beacons. In our case, it was the atrium with metal glass construction connecting two buildings. We tried to place there two beacons on the opposite sides of the atrium. But the signal distortion was too high, we decided to install there only one beacon to prevent the signal interference with the second beacon.

During the evaluation, some of the beacons triggered the position notification too early or too late. This issue can be solved either by the reconfiguration of the beacon or by moving the beacon 1 meter forward/backward.

The automatic notification often did not trigger because the user switched to the next segment. In our current solution the application is ranging the beacons only for the currently displayed segment. In future, this should be improved and the user notified about the surrounding information when arrives near any of the beacons.

Two of the participants took the wrong turn during the evaluation. They were able to recover from this error. We can say that even without the distance and direction information our proposed combination of navigation instructions with beacons only at decision points can solve the user's walkthrough errors using backtracking.

The manual location verification feature did not result in higher reliability. The participants tend to stop using this feature during the walkthrough. It happened that the automatic verification of the location worked correctly, but the manual verification afterward did not. The main reason was that the participant crossed the decision point by few steps and tried to verify the location beyond the beacon range. This could have influenced the user's confidence level at the decision point. We think that this feature can be omitted from the application.

If we compare our approach to NavCog [9] in terms of the number of beacons necessary, for the indoor and semi-indoor (courtyard) parts of the route (250 meters and 26 segments) we needed 26 beacons in total including error beacons. If we placed the beacons every 6 meters we would need approximately 42 beacons to cover only the route excluding error beacons.

VI. Conclusion

We designed a prototype of navigation application that has two main building blocks, the landmark-enhanced navigation instructions and the location synchronization system that uses the minimum number of beacons possible and that is also capable of error prevention and error recovery.

We conducted a qualitative study of high-fidelity prototype of this system with 8 visually impaired participants. As previous studies shown [13] the landmark-enhanced navigation instructions are suitable navigation for blind pedestrians, still there are many difficulties which we wanted to solve with utilizing beacons as synchronization points, i.e. finding entrances, public transport stations, help when identifying the landmarks and give users more confidence during the walk-through.

As we found out, the synchronization points successfully complement the navigation system using only navigation instructions. The main benefit of our solution lies in the use of a low number of beacons. But maintaining the effectiveness of navigation thanks to detailed and landmark-enhanced navigation instructions.

For the future, we see a potential in installing the beacons only on the most used and hard-to-find decision points namely at public transport station and entrances to public buildings.

Acknowledgment

The research has been supported by projects: Navigation of handicapped people funded by grant no. SGS16/236/OHK3/3T/13; Research Center for Informatics (OP VVV CZ.02.1.01/0.0/0.0/16_019/0000765); Automated mapping of routes and barriers for pedestrians and disabled people funded by grant no. TH02010839 of the Technology Agency of the Czech Republic realized by Central European Data Agency, a.s.

References

[1] R. G. Golledge, "Geography and the disabled: a survey with special reference to vision impaired and blind populations," *Tran. of the Inst. of British Geographers*, pp. 63–85, 1993. [Online]. Available: https://www.jstor.org/stable/623069

[2] J. Faria, S. Lopes, H. Fernandes, P. Martins, and J. Barroso, "Electronic white cane for blind people navigation assistance," in *WAC 2010*. IEEE, 2010, pp. 1–7.

[3] M. Bujacz, P. Baranski, M. Moranski, P. Strumillo, and A. Materka, "Remote guidance for the blind – A proposed teleassistance system and navigation trials," in *HSI 2008*. IEEE, 2008, pp. 888–892. [Online]. Available: https://doi.org/10.1109/HSI.2008.4581561

[4] M. Modsching, R. Kramer, and K. ten Hagen, "Field trial on GPS Accuracy in a medium size city: The influence of built-up," in *3rd workshop on positioning, navigation and communication*, 2006, pp. 209–218.

[5] A. W. Siegel and S. H. White, "The development of spatial representations of large-scale environments." *Adv. in child development and behavior*, vol. 10, p. 9, 1975. [Online]. Available: https://doi.org/10.1016/S0065-2407(08)60007-5

[6] R. G. Golledge, "Human wayfinding and cognitive maps," *Wayfinding behavior: Cognitive mapping and other spatial processes*, pp. 5–45, 1999.

[7] I. Gorovyi, A. Roenko, A. Pitertsev, I. Chervonyak, and V. Vovk, "Real-time system for indoor user localization and navigation using bluetooth beacons," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, May 2017, pp. 1025–1030. [Online]. Available: https://doi.org/10.1109/UKRCON.2017.8100406

[8] O. Czogalla and S. Naumann, "Pedestrian indoor navigation for complex public facilities," in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Oct 2016, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IPIN.2016.7743672

[9] D. Ahmetovic, M. Murata, C. Gleason, E. Brady, H. Takagi, K. Kitani, and C. Asakawa, "Achieving Practical and Accurate Indoor Navigation for People with Visual Impairments," in *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, ser. W4A '17. New York, NY, USA: ACM, 2017, pp. 31:1–31:10. [Online]. Available: http://doi.acm.org/10.1145/3058555.3058560

[10] E. Freeman, G. Wilson, S. Brewster, G. Baud-Bovy, C. Magnusson, and H. Caltenco, "Audible Beacons and Wearables in Schools: Helping Young Visually Impaired Children Play and Move Independently," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 4146–4157. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025518

[11] Y. Guo, W. Wang, and X. Chen, "FreeNavi: Landmark-Based Mapless Indoor Navigation Based on WiFi Fingerprints," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–5. [Online]. Available: https://doi.org/10.1109/VTCSpring.2017.8108350

[12] L. Zeng and G. Weber, "A pilot study of collaborative accessibility: How blind people find an entrance," in *MobileHCI '15*. ACM, 2015, pp. 347–356. [Online]. Available: https://doi.org/10.1145/2785830.2785875

[13] J. Balata, Z. Mikovec, and P. Slavik, "Landmark-enhanced route itineraries for navigation of blind pedestrians in urban environment," *Journal on Multimodal User Interfaces*, Apr 2018. [Online]. Available: https://doi.org/10.1007/s12193-018-0263-5

[14] J. Vystrčil, Z. Míkovec, and P. Slavík, "Naviterier - indoor navigation system for visually impaired," *Smart Homes 2012*, 2012.

[15] J. Balata, J. Berka, and Z. Mikovec, "Indoor-Outdoor Intermodal Sidewalk-Based Navigation Instructions for Pedestrians with Visual Impairments," in *International Conference on Computers Helping People with Special Needs*. Springer, 2018, pp. 292–301. [Online]. Available: https://doi.org/10.1007/978-3-319-94274-2_41

[16] WHO, "ICD update and revision platform: change the definition of blindness," 2009. [Online]. Available: http://www.who.int/blindness/Change the Definition of Blindness.pdf

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 887–891
DOI: 10.15439/2018F287
ISSN 2300-5963 ACSIS, Vol. 15

A mixed reality application for sketching in prototyping workshops

Katia Cirillo
Daimler AG
Mercedesstrasse 137
70327 Stuttgart, Germany
Email: katia.cirillo@daimler.com

Sascha Herr, Nico Koprowski, Omar Sanchez
Daimler AG
Mercedesstrasse 137
70327 Stuttgart, Germany
Email: sascha.herr@daimler.com
nico.koprowski@daimler.com
omar.w.sanchez@daimler.com

***Abstract*—Sketching is a method used in user-centred design to visualize first drafts of a product. In corporate environments, sketching is often employed in ideation workshops with participants of various disciplines including end users. The aim of sketching is to promote communication and create a better understanding between stakeholders. However, participants are sometimes reluctant to engage in the activity for fear of inferior drawing skills. In order to counteract this phenomenon, we designed a mixed reality application that supports users in sketching, particularly in workshop settings. Two independent user studies showed conflicting results depending on the assumed perspective. First hand users find that the application effectively supports them in creating high quality sketches with high enjoyment in the process, although they do not see creativity enhancements or higher time efficiency. In contrast, third parties that rate the produced artifacts could not distinguish between application supported and free hand sketches in terms of uniformity, comprehensibility and quality.**

I. Introduction

PROTOTYPING is a method for early evaluation of product designs in the user-centred design process. In prototyping workshops, hand-painted drawings called sketches serve to visualize initial design ideas quickly and easily. This promotes communication between participants which iteratively contributes to a better result [1]. Sketches vary according to the draftsman's knowledge and skills. This happens, for example, because users do not recognize sketched elements, such as buttons or icons.

Our contribution is to support users in sketching and to improve the quality of sketches. We aim to improve understanding of sketches. In this context, we evaluate if a mixed reality application for the Microsoft HoloLens meets the requirements as a supporting application for sketching in prototyping workshops.

A. User-centred design process

User-centred design is an iterative process of defining and specifying user requirements [2]. A prototype serves as a "representation of all or part of an interactive system, that, although limited in some way, can be used for analysis, design and evaluation" [2]. In the creation of products, multidisciplinary teams support creativity and create better solutions. Workshops thus provide a suitable framework for the creation of prototypes. Buxton defines sketching as the process of quickly creating handmade drawings to visualize first drafts [1]. According to the author, sketches mainly serve as means of communicating design ideas. They can be distributed to potential users for feedback. Criticism and suggestions for improvement are most valuable in early stages of development. Sketches also serve as documentation for the design process in order to keep early design decisions comprehensible later on. In sum, sketching promotes communication in design creation, enables experimentation and supports the creative process.

This work was supported by Daimler AG

B. Requirements analysis

We determine the context of use on participants of prototyping workshops. One target group we address consists of persons that are well versed in the development of digital applications but cannot draw well. The other target group consists of end users of a planned product who have no experience in sketching user interfaces.

Based on the context of use, we establish the following hardware requirements for a tool that supports sketching:

1) Displaying drawing templates: The hardware should provide a possibility of displaying drawing templates that users can trace. Templates help the target group to draw professionally looking sketches and support them in sketching.

2) Use of pen and paper: In the design process, designers prefer the use of pen and paper rather than computer devices [3], [4]. That's why we determine that users do not have to sketch on computer devices. This implicates that users need their hands free for holding a pen and interact with physical objects.

3) Fast and easy set-up: As workshops take place in different premises, we set a mobile solution that includes a fast and easy set-up as a further requirement.

We determine further requirements for the software. These are derived from the definition we gave about sketching in section I-A.

4) Improvement of the quality of sketches: We aim to improve sketches by offering support to users for the drawing process. We try to establish conventions for elements in sketches. Consequently, sketches will look consistent no matter

which user has drawn it. By this, we try to achieve a better understanding of the sketches.

5) Not limiting creativity or discussion: Our solution supports sketching and should not interfere with the goals of the method. By definition, sketching aims to support creativity. Furthermore, sketches motivate workshop participants to communicate and discuss. A requirement that we define is thus not to limit creativity or discussion.

6) Joy of use: Positive feelings while using a product supports creativity [5]. Thus we aim to develop a solution that users like to use.

C. Related Work

Studies that support paper prototyping with overhead projector and camera have been rated positively by participants and experts [6], [7]. Such a hardware setup is time-consuming. A study that solves this problem concerns the use of mobile devices [8]. In this approach, the drawing is not done on paper but on the devices. However, it has already been proven that in early stages of design creation, designers prefer the use of pen and paper rather than computer devices [3], [4].

Another study supports drawing with pen and paper by applying the onion skinning technique [9]. The draftsman holds a mobile device with one hand and draws underneath the device with the other hand. According to the authors, feedback was mostly positive but users considered the device too heavy and discomfortable to hold. For further work, the authors propose a solution for see-through displays.

SketchAR [10] is an application with drawing templates for Augmented-Reality. The device projects drawing templates onto a surface in the room. Users then trace the projected lines with a pen. The application includes drawing templates of everyday objects, such as animals, plants, and humans. However, elements that are necessary for use in the context of sketching digital products are missing.

An application that meets the requirements that we defined in section I-B has not been found during research. However, in the presented work, users accept applications supporting prototyping positively.

II. Sketching Tool

Microsofts' HoloLens [11] meets the hardware requirements of a supporting tool for sketching defined in section I-B. The head-mounted device makes it possible to project digital content onto a surface with fixed world anchors. The advantage over smartphones is that the device is head-worn, leaving hands free for drawing. There is no need of further technical equipment and cables. This is suitable for workshops, as these are often held in external premises where technical equipment varies. As hardware requirements are met with the HoloLens, the following section examines how to meet software requirements defined in section I-B.

A. Concept

The tool offers graphics of UI elements for sketching, which a user may then trace. For a better overview of the available UI widgets, they are classified into categories. Users access the sketching mode by choosing an UI widget from the menu. A graphical visualization appears in front of a whiteboard. As support in sketching, we employ the onion skinning technique as it has been positively received in previous studies. Users step in front of the whiteboard and trace the lines of the projected element with a pen. Users can re-position, scale or rotate the element. If users have painted the element, they remove the digital projection. The hand-painted element remains on the whiteboard.

User interact with the tool by hand gestures. Audio output and speech commands are not used. We assume that audio output and speech commands distract users and create distance between user and workshop participants.

B. HoloLens prototype

The implemented prototype is shown in Fig. 1. We captured the prototype in the Unity Editor. The first image shows the sketch elements menu. The second image includes an UI widget in sketch mode and the main menu below the element. The third image displays the bounding box that allows users to scale and rotate the sketch element.

III. User study I: User tests with the prototype

We evaluated the suitability of the HoloLens prototype in sketching workshops. We asked users to create sketches with and without the application. Afterwards, they gave feedback and rated their experiences in a questionnaire.

A. Participants

We selected test users without extensive experiences with sketching to prevent influence through routine. The sample represented users that are insecure with the sketching process. We carried out user tests with five female and four male participants. The age range was between 21 and 33 years. The mean values were 4.89 for the previous experience with smartphone apps, 1.89 for head-mounted displays and 2.89 for sketching.

B. Design

In order to evaluate the effect of the application, we conducted a within subjects design. Each test person performed one task per condition: once only with pen and paper and once with the aid of the HoloLens application.

In each condition the task for the participants was to sketch an application that was defined beforehand. It was assumed that participants were familiar with smartphone apps. We selected apps based on a survey on the use of smartphone functions [12]: chat, music player, picture gallery, news articles, and public transport. We also added a task planner app. A textual specification of the use cases ensured that participants concentrated on creating sketches instead of spending effort to understand the task. We aimed to define an environment similar to workshop situations and offered test users to address the moderator for questions regarding the content of the use case.

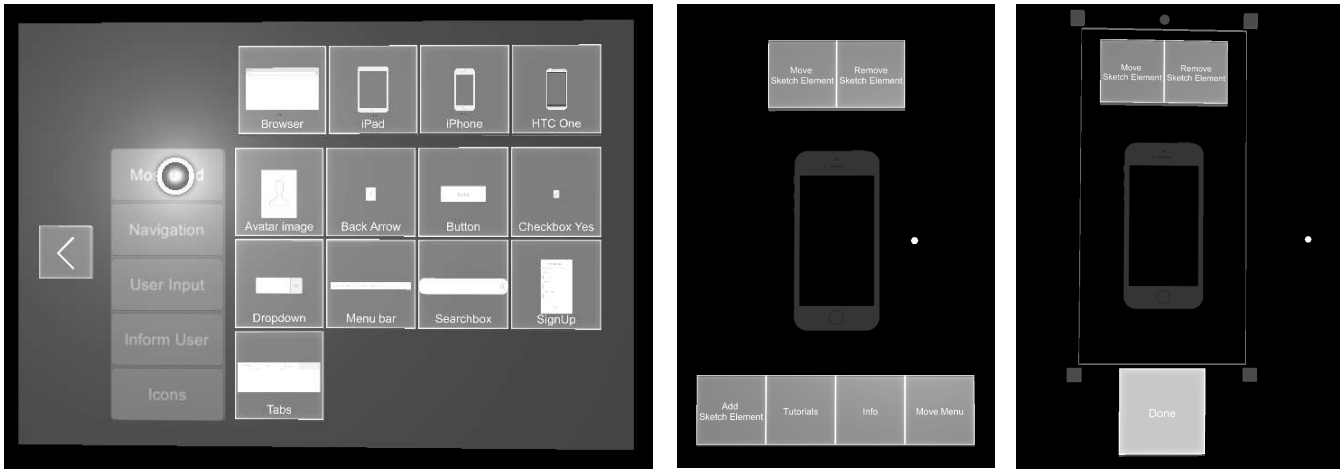


Fig. 1. Images of the HoloLens prototype from the Unity Editor. The first image (from left to right) shows the sketch elements menu, the second image shows a UI widget in sketch mode and the main menu. The third image displays the bounding box that allows users to scale the sketch element.

For each participant, we randomized the use cases and the order of conditions to avoid carry over effects. Furthermore, an equal distribution of the factors gender and prior experience was considered.

C. Questionnaire criteria

Participants rate criteria on a questionnaire after having performed the two tasks. The criteria included requirements from section I-B and were derived from similar work [7], [8]. In the questionnaire, users rate their extent of agreement on a Likert scale. The criteria were as follows.

1) Effectiveness: In order to answer whether the developed application is suitable for creating sketches, the effectiveness is evaluated. It is defined that the application is effective if the content needed to achieve the task is available.

2) Creativity: By definition, the method sketching supports the creative process [1]. Therefore, we evaluate whether the test users think that the application supports them in creativity.

3) Time-efficiency: Experts determine rapid execution as an important property of prototyping tools [7]. Since time of execution is an uncontrollable constant that varies between people, it is not feasible to measure it. Instead, we query an assessment of the test users.

4) Enjoyment: Positive feelings support creativity [5]. An expert also classifies enjoyment as important for prototyping tools [7]. We ask users whether the solution of the task was more fun with the support of the application.

5) Quality of results: We assume that the application is more likely to be used if users consider the results to be more professional than without the application.

D. Procedure

We carry out the user tests with each test person individually. After an introduction, test subjects read through the first task and start with the head-worn HoloLens. All test users operate the HoloLens by hand. Test users solve the tasks uprightly in order to simulate the workshop scenario.

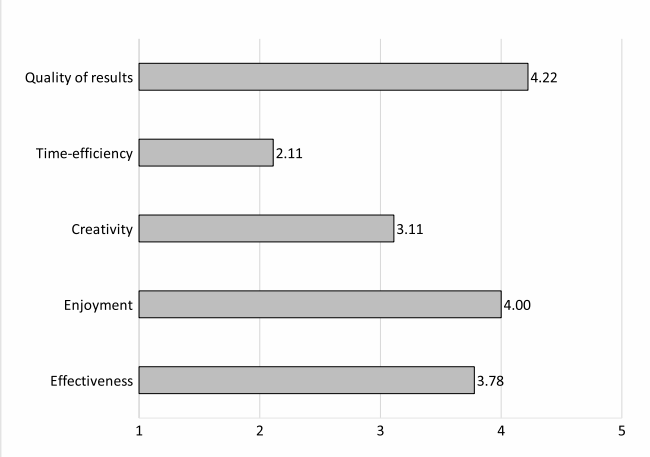


Fig. 2. Mean values of the criteria quality of results, time-efficiency, creativity, enjoyment, and effectiveness. The scale ranges from one (This does not apply) to five (This applies)

They start to sketch with a ballpoint pen on paper sheets. After completing the first task, the test users read and solve the second task. The moderator removes the sheets of the first round so that the user can no longer look at them. The test users fill out the questionnaire after completing both assignments.

E. Results

The questionnaire evaluates the criteria quality of results, time-efficiency, creativity, enjoyment, and effectiveness. The results are shown in Fig. 2. The graphic includes the mean value of the ratings from the test users per criterion.

The mean value for the quality of results represents the highest value (M=4.22, SD=0.83). The test users agreed that the solution of the task was more fun with the system (M=4.00, SD=0.87). They slightly agree that the application offered the contents they needed (M=3.78, SD=0.83). They neither agree

nor disagree in average that the application supports their creativity (M=3.11, SD=1.36). The only statement that test users do not agree with is the time-efficiency of the application (M=2.11, SD=1.27).

In a free text field for further notes, test users referred to the HoloLens application and noted more consistency in sketches; usage became easier after adaption; more enjoyment; added value of support with drawing templates; suitability for large elements instead of small elements; usage for small elements is time-consuming and impractical; no additional support of creativity; support for visualization of symbols; usage better for copying the element instead of tracing lines directly from drawing template.

IV. User study II: Evaluation of the sketches

We decided to conduct a subsequent study in order to evaluate the sketches from study I. Another group of participants evaluates the sketches which were created with and without the help of the application.

A. Participants

We select participants without any specific prior knowledge for the assessment of sketches. As this work is supported by Daimler AG, we determine German employees of the company as population and select a random sample. The number of employees of Daimler AG on 31.12.2017 in Germany was 172,089 [13]. In the online survey, 70 male and 25 female persons participate. The age is given in ranges of 20-29 years (31 participants), 30-39 (34 participants), 40-49 (16 participants), 50-59 (13 participants) and 60-69 (one participant). No person is younger than 20 years or older than 69.

B. Design

We plan the study with a within subjects design. We use the resulting sketches from the first study as stimulus. Each participant evaluates sketches with both features, i.e. with and without HoloLens application. Participants give their rating on a Likert scale in an online questionnaire.

Each participant evaluates three out of six use cases of each stimulus. To ensure that the same prerequisites apply to both stimuli, the tool stores each use case once per stimulus. We randomize the selected use cases and order of their occurrence. We give no information about the context in which the sketches were created. Users evaluate each use case independently of the other.

C. Questionnaire criteria

With the criteria we selected, we aim to evaluate if the application supported sketches look better.

1) Comprehensibility: Sketches are used to communicate design ideas in an understandable way [1]. By this criterion we determine whether participants recognize concepts and elements, such as buttons, arrows or icons.

2) Basis of discussion: Sketches serve as a basis of discussion and means of getting user feedback [1]. The survey participants evaluate if the sketches serve as a basis of discussion.

3) Uniformity: We assume that more uniformity in the presentation of the sketches leads to a better understanding. We define uniformity in sketches if users have chosen the same representation of elements.

4) Quality of results: The representation of sketches varies depending on the draftsmans' skills or the degree of maturity of the sketches. To counteract this effect, the application aims to improve it. Participants evaluate whether the sketches appear to be made in a professional context, for example by an experienced team or with the help of a tool.

D. Procedure

Participants evaluate the sketches in an online survey. Each questionnaire starts with an introduction to sketching and the evaluation criteria. Each of the following six pages shows a use case consisting of three sketches and a Likert scale.

E. Results

For the results we only consider questionnaires where the participant reached the last page. Table I shows mean values and standard deviations for the evaluated criteria.

TABLE I
Mean values and standard deviations for the evaluated criteria in user study II.

Criterion	Conventional sketches		HoloLens sketches	
	Mean	SD	Mean	SD
Uniformity	3.14	1.35	3.18	1.35
Comprehensibility	3.25	1.7	3.06	1.47
Basis of discussion	3.26	1.74	3.19	1.37
Quality of results	2.36	1.11	2.66	1.34

V. Discussion

The suitability of the application for sketching is derived from the criteria time-efficiency and effectiveness, which were evaluated in the user tests. The results show that sketching takes longer. Users needed time to familiarize themselves with the device and the application. Test users searched for suitable elements. Additionally, they needed more time to draw elements. Interactions are unfamiliar with first-time users. However, these limitations can be improved through training. Not all test users stated that the required content was provided. This can be improved easily by adding additional elements.

Based on user study II, we do not see any clear effects in the application supported sketches. We could not prove that sketches get more uniform with our tool. It follows that we cannot assess whether more uniformity in the presentation of sketches has a positive effect on comprehensibility.

Nevertheless, we recognize that test users were positive about our application. Test users believe that they have achieved more professional results. In addition, they enjoy solving the task with the application. We assume that an improvement in hardware and the application will lead to better results in time-efficiency.

In summary, the evaluation revealed that test users rate the application positively in terms of quality of results and enjoyment. Accordingly, the added value of using HoloLens

for sketching lies more in the use of the application than in resulting sketches. The application seems suitable to solve inhibitions of participants of prototyping workshops towards the method sketching. Thus, the application appears to be useful to familiarize participants with sketching. As test users say they have had more fun and think that the resulting sketches are more professional, it is conceivable to involve participants of workshops in the sketching process. Through the increased involvement of participants, we could encourage communication and discussion. Since this is one of the main reasons to use sketching, we suggest further investigation about the benefits of the application.

VI. Conclusion and Future Work

We investigated how a mixed reality application supports the method sketching in prototyping workshops. We described the suitability of the Microsoft HoloLens as a device for an application in sketching. The target group for the application are participants of prototyping workshops. We proposed drawing templates to support the sketching process and familiarize participants with the method.

We implemented a prototype on HoloLens. The device projects drawing templates onto a wall where users then trace the element with pen on a paper. In user tests, we evaluated the added value of the tool. As a result, the application has been well received by users. We therefore propose to carry out further studies. It seems interesting to put the focus of evaluation on user experience to determine the involvement of participants in the sketching process.

For better results in an evaluation of usability and user experience, we suggest to continue development of the prototype by expanding the range of UI widgets and improving the interaction concept. We suggest minimizing direct interaction by the user within the application.

Another relevant criterion is technology acceptance. If users do not want to use the device in workshops because they feel restricted or not taken seriously, the application will not be used. In addition, we suggest the evaluation of the suitability in collaborative group work. The application can furthermore be extended so that several HoloLens devices can be connected so that users can work together in remote work scenarios.

References

[1] W. Buxton, *Sketching user experiences : getting the design right and the right design*. Elsevier/Morgan Kaufmann, 2007. ISBN 0123740371

[2] International Organization for Standardization, "Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems," 2010. [Online]. Available: https://www.iso.org/obp/ui/{\#}iso:std:52075:en

[3] D. J. Cook and B. P. Bailey, "Designers' use of paper and the impliations for informal tools," *Ozchi*, pp. 1–10, 2005.

[4] M. W. Newman and J. A. Landay, "Sitemaps, storyboards, and specifications," in *Proceedings of the conference on Designing interactive systems processes, practices, methods, and techniques - DIS '00*. New York, New York, USA: ACM Press, 2000. doi: 10.1145/347642.347758. ISBN 1581132190 pp. 263–274.

[5] A. Filipowicz, "From positive affect to creativity: The surprising role of surprise," *Creativity Research Journal*, vol. 18, no. 2, pp. 141–152, 2006. doi: 10.1207/s15326934crj1802_2

[6] J. Laviole and M. Hachet, "PapARt: Interactive 3D graphics and multi-touch augmented paper for artistic creation," *IEEE Symposium on 3D User Interfaces 2012, 3DUI 2012 - Proceedings*, pp. 3–6, 2012. doi: 10.1109/3DUI.2012.6184167

[7] B. Bähr, *Prototyping of user interfaces for mobile applications*, ser. T-Labs Series in Telecommunication Services. Cham: Springer International Publishing, 2017. ISBN 978-3-319-53209-7

[8] Q. Su, W. H. A. Li, J. Wang, and H. Fu, "EZ-sketching: Three-level optimization for error-tolerant image tracing," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 54:1–54:9, 2014. doi: 10.1145/2601097.2601202

[9] C. Lins, E. Arruda, E. Neto, R. Roberto, V. Teichrieb, D. Freitas, and J. M. Teixeira, "Animar: Augmenting the reality of storyboards and animations," *Proceedings - 2014 16th Symposium on Virtual and Augmented Reality, SVR 2014*, pp. 106–109, 2014. doi: 10.1109/SVR.2014.40

[10] Sketchar.tech, "SketchAR." [Online]. Available: http://sketchar.tech/

[11] Microsoft, "Buy Microsoft HoloLens for Developers and Business | HoloLens." [Online]. Available: https://www.microsoft.com/de-de/hololens/buy

[12] Statista, "Smartphone features: usage in Germany 2017 | Survey." [Online]. Available: https://www.statista.com/statistics/436515/smartphone-features-usage-germany/

[13] Daimler AG, "Daimler Unternehmensprofil." [Online]. Available: https://www.daimler.com/dokumente/investoren/berichte/geschaeftsberichte/daimler/daimler-ir-daimler-im-ueberblick-2017.pdf

Stereoscopy in Graphics APIs for CAVE Applications

Jerzy Redlarski
Dept. of Intell. Inter. Systems,
Faculty of Electronics, Telecommunication and Informatics,
Gdańsk University of Technology,
G. Narutowicza 11/12,
80-233 Gdańsk, Poland
Email: jerredla@pg.edu.pl

Robert Trzosowski
Dept. of Intell. Inter. Systems,
Faculty of Electronics, Telecommunication and Informatics,
Gdańsk University of Technology,
G. Narutowicza 11/12,
80-233 Gdańsk, Poland
Email: robtrzos@pg.edu.pl

Mateusz Kowalski
Dept. of Intell. Inter. Systems,
Faculty of Electronics, Telecommunication and Informatics,
Gdańsk University of Technology,
G. Narutowicza 11/12,
80-233 Gdańsk, Poland
Email: matkowa1@pg.edu.pl

Błażej Kowalski
Dept. of Intelligent Interactive Systems,
Faculty of Electronics, Telecomm. and Informatics,
Gdańsk University of Technology,
G. Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: blakowal@pg.edu.pl

Jacek Lebiedź
Dept. of Intelligent Interactive Systems,
Faculty of Electronics, Telecomm. and Informatics,
Gdańsk University of Technology,
G. Narutowicza 11/12, 80-233 Gdańsk, Poland
Email: jacekl@eti.pg.edu.pl

***Abstract*—The paper compares the advantages and disadvantages of a variety of Graphics Application Programming Interfaces (APIs) from the perspective of obtaining stereoscopy in applications written for a CAVE virtual reality environment. A number of problems have been diagnosed and an attempt has been made to solve them using OpenGL, DirectX 11 and 12, Vulkan, as well as the Unity Engine which can internally use DirectX, OpenGL and Vulkan, but has problems and limitations of its own.**

I. Introduction

STEREOSCOPY (stereo imaging) is a technique for producing an illusion of depth by delivering two different 2D images (generated from two different points of view) to each of the viewer's eyes. It is based on predator (and human) way of 3D perception by means of binocular vision (stereopsis). Stereoscopy can rely on direct delivery of two images on separate screens mounted in front of each eye (stereoscope, HMD – Head-Mounted Display) or displaying two images on a common screen visible for both eyes and using special filtering glasses separating images (3D cinema, CAVE – CAVE Automatic Virtual Environment). Filtering glasses can use various technologies: active (separation in time by shutter glasses) and passive (spectrum separation or polarization separation) [1, 2, 4].

The Immersive 3D Visualization Lab (I3DVL) located at the Faculty of Electronics, Telecommunication and Informatics of the Gdańsk University of Technology contains three CAVEs of various sizes: closed BigCAVE with six screen-walls, open MidiCAVE with four screen-walls and MiniCAVE based on four 3D monitors [3, 5, 7]. Immersive 3D visualization in these CAVEs requires stereoscopy and its synchronization on all screen-walls. The popular game engine Unity serves as the basic development tool in I3DVL. Unfortunately, it offers limited support for the creation of software for CAVEs and needs some adaptation.

CAVE systems put a number of unique requirements on applications, such as the need to synchronize a large number of screens or projectors, or the need to support active and passive stereoscopy, preferably at the same time. This in turn may necessitate the use of a quadbuffer (a screen buffer with room for four screen-sized pictures) or similar solution, if graphical artifacts are to be avoided. Rendering has to happen at the correct frequency (120 Hz for our BigCAVE system), or else frame skipping may happen – or worse, stereoscopy is disrupted if frames intended for one eye are displayed to the other eye. Latency should also be minimized, as delays in rendering are known to induce discomfort and dizziness in viewers, sometimes preventing them from extended use of virtual reality.

Most of the applications run on our CAVE system have been developed in the Unity engine, using a third party, proprietary library to achieve synchronization and stereoscopy. However the library had major limitations, including unsatisfactory performance and inability to work with Unity versions past 5.0.2. We also wanted a solution where we'd have access to the source code, so that students and researchers could implement not only applications, but also make changes to or expand the library itself if necessary. This forced us to start working on a new library, since the existing ones were either outdated or had proprietary licenses.

For a long time, the Unity Engine itself had no native support for stereoscopy, while also limiting direct access to the underlying graphics APIs – particularly during the initialization stage. Thus, the only way to achieve synchronization and stereoscopy was to render the scene to a texture, then use a separate rendering context – created outside of Unity – to display the resulting image. Since the contexts

This work was supported in parts by Entity Grant to Finance the Maintenance of a Special Research Device (SPUB) from the Ministry of Science and Higher Education (Poland) and DS Funds of the Faculty of ETI at the Gdańsk University of Technology

were separate, it was possible to use different graphics APIs for each task, so our testing included using DirectX 11 and OpenGL in Unity to render images for both eyes, while using OpenGL, DirectX 12 or Vulkan in another window to display those images at the correct frequency and synchronized between projectors and with active glasses.

II. Single camera stereoscopy

One of our attempts involved the simplest way to achieve stereoscopy in Unity – by alternating the placement of the camera every other frame. To achieve fluid stereoscopic animation, the camera is synchronously (on all client computers) moved back and forth between points A and B, representing the position of left and right eye respectively. This needs to happen exactly 120 times per second (for our setup), each frame shown for the exact same time. It also needs to be synchronized with the active 3D glasses, which alternate between darkening left and right LCD shutters at the same rate.

While alternating the camera's position between frames is trivial, implementing the synchronization with glasses in Unity is problematic. Even when configured to keep a fixed frame rate of 120 frames per second, it's common for frames to vary in length, depending on unpredictable changes in rendering times. This results in the viewer often seeing either two overlapping pictures, or reversed pictures, where the left eye sees the picture intended for the right eye and vice versa – a (dizzying for the viewer) occurrence referred to as a desynchronization.

Attempts to reduce the issue by increasing the frame rate to 240 Hz didn't solve the problem, even when the number of frames between switching camera positions was adaptively adjusted based on the amount of time a frame actually took to render. Increasing the frame rate further (360 Hz) was barely possible, since even an empty scene in Unity, with no running scripts, physics or camera movement, typically oscillates around 300 to 330 frames per second.

In best cases, using this method resulted in stereoscopy that would switch between correct and reversed pictures every few seconds, and after less than a minute resulted in complete desynchronization. Most likely the problem lies in the fact that the Unity Engine is not designed with constant frame rates in mind. Additionally, this method is useless in applications that – due to their complexity – suffer from occasional or frequent drops in frame rates. Such applications would need to be better optimized and tested – which most applications don't do sufficiently for this method to work. This is especially true if they don't normally (for use without stereoscopy) need a constant frame rate of 120 frames per second – such as when the movement is very slow.

III. Stereoscopy support in Unity

To properly display stereoscopic graphics in a CAVE installation, a number of tasks needs to be done. Each screen requires two images rendered from two points – representing the left and right eyes. Each of the viewer's eyes can view any part of any screen at any time, in contrast to HMD (head mounted display) devices which split the display into parts visible by one eye only. In a CAVE, the left and right eye images need to use the same area of the display screens. To obtain correct perspective, the position of the eyes in regard to the display surface is crucial – it is calculated using the head tracking system. The interpupillary distance is then used to position both cameras correctly in eye positions. The images on screens that share an edge need to be consistent. This is achieved by correctly setting the projection matrices and view matrices, achieving an asymmetric frustum with edges passing through the corners of the display. The display surface is where the images for both eyes converge. Unlike in HMD devices, the frustum does not follow the rotation of the viewer's head, but rather is always facing the same direction. If the technology used (such as active shutter glasses) requires the left and right eye images to alternate, it is crucial to synchronize this between all screens, so that they all display images intended for the same eye at the same time – especially if each screen is governed by a separate instance of the application. To achieve this, the usual solution is to prepare two images every application update, then pass the task of displaying the correct one to synchronized graphics cards such as NVIDIA Quadro Sync (used in I3DVL).

When OpenGL is used to write software, one can use quadbuffers to achieve stereoscopy [9]. Firstly, one needs to set the `PFD_STEREO` flag in the `PIXELFORMATDESCRIPTOR` structure using the `SetPixelFormat` function [11]. This can only be done during context initialization. This in turn makes it possible to render to the left and right backbuffers, switching between them using `glDrawBuffer(GL_BACK_LEFT)` and `glDrawBuffer(GL_BACK_RIGHT)` calls.

In the Unity3D engine, rendering is done with the `Camera` component. To configure it for each CAVE screen the `projectionMatrix` and `worldToCameraMatrix` fields have to be set with appropriate matrices. We tried to use quadbuffering in Unity, but it proved impossible due to the lack of low-level access to the graphics device during the initialization step. It was not possible to set the aforementioned flag. We found two workarounds to this problem.

The first one was to inject a modified DLL library during application launch. The library would intercept the `SetPixelFormat` function call and set the `PFD_STEREO` flag, resuming with normal initialization afterwards. A program launched this way could use two cameras, switching between the target back-buffers in the `OnPreCull` camera event from the native plugin level. A native plugin in Unity is one that is compiled from non-virtual machine languages such as C and C++. The disadvantage of this solution is that it's hard to implement and requires a separate piece of software for injecting the DLL at application launch.

The other workaround is to create a separate OpenGL window (from a native plugin) which is only tasked with displaying the images, while the Unity application renders those images to a virtual texture and passes them to the OpenGL window. When launched, the Unity application would create the window (which needs to belong to the same process) with a new graphics context initialized with

the PFD_STEREO flag. Using the wglShareLists function we can enable the Unity context and the OpenGL window context to share a single display-list space. On the Unity side the two cameras create RenderTexture objects, the virtual textures, which are set as targetTexture of the cameras. This results in Unity rendering the image for left and right eye to those textures. The GetNativeTexturePtr function gives us pointers to these textures which can be passed to the OpenGL window. Thanks to resource sharing, the window can then bind those textures with the glBindTexture function. All that is left to do is for the window to draw those textures to the correct backbuffers.

This workaround has a number of disadvantages. One is the slightly lowered performance because the main unity application is now considered by the operating system to be running in the background, while the OpenGL window which does almost no work is in the foreground. This can result in suboptimal assignment of processing power to the threads, since modern operating systems tend to prioritize foreground windows. Another problem is that input from devices such as keyboard and mouse will be directed to the foreground window, which may require redirecting the input events to the Unity application.

Unity3D has been offering support for VR (virtual reality) applications for a while, but their focus is on HMD devices [8]. Since 5.4 version, they extended support to include stereoscopy-capable flat panel displays. The newly added "Stereo Display (non head-mounted)" option makes it possible to obtain stereoscopic images which can be rendered using a variety of graphic APIs – OpenGL, Vulkan, Direct3D 11 and Metal. At the time of this writing, it doesn't work with Direct3D 12 yet. The Camera object can be set to render images for both eyes or we can use two Camera objects each responsible for one eye. The matrices for these cameras are set using the SetStereoProjectionMatrix and SetStereoViewMatrix functions. We can also set in the project options whether rendering should alternate between left and right images or if we want to render both images simultaneously (known as the Single-Pass Stereo rendering optimization [10]).

Using Unity's engine for stereoscopy has a number of advantages. The engine unifies the implementation of stereoscopy for the different graphics APIs and operating systems. The engine itself also performs a multitude of optimizations (such as the aforementioned Single-Pass Stereo rendering) [8]. Creating and launching the application is easier – no need for injecting libraries or intercepting control devices input from the second window. However, there are also drawbacks. The main one is that (non-HMD) stereoscopy support in Unity is fairly new and a niche need, and thus software bugs and instability are common. Common problems include the random freezing of rendering for one of the eyes, especially with HDR (High Dynamic Range) or MSAA (Multi-Sample Anti-Aliasing) on. Other problems were encountered with Reflection Probes and

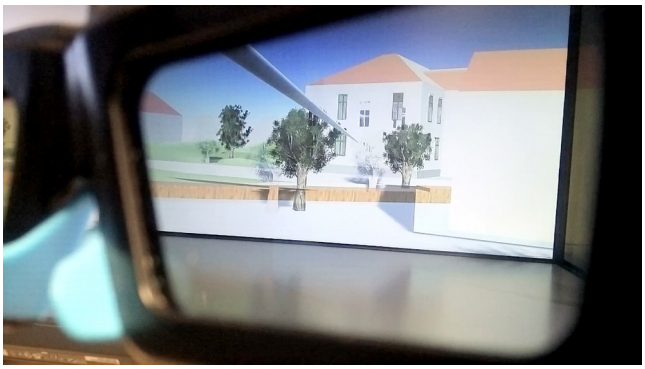


Fig 1. Improperly displayed tree shadows in Unity, as seen through active glasses.

Terrain objects (such as trees) improperly displaying shadows, which were rendered to the wrong eye or not at all, as shown in (Fig. 1).

IV. DirectX 12 and Vulkan in separate context

Vulkan is a modern 3D graphics API, released in 2016. Its main goal is to provide a low-level alternative to OpenGL as a multiplatform API, and thus offer better efficiency and more direct control over computations [12]. Compared to older APIs it excels at using multiple CPU cores in parallel (through support for multithreading), and the CPU load at runtime is reduced greatly by precompiling shaders to SPIR-V, an intermediate form that allows drivers to be much simpler [6].

Despite its advantages, Vulkan has a few drawbacks too. It is more verbose, requiring more code, and thus work, to get it to do things that are easier done in higher-level languages. This also means more room for bugs and harder maintenance. Another problem is the novelty, which means the API might see frequent changes as it matures. People looking to learn the API may have fewer options compared to older APIs that have hundreds of books and tutorials available. People who run into problems are less likely to find solutions on the internet. Few development tools, third-party libraries and game engines have support for Vulkan.

For our purposes, Vulkan would offer a few significant advantages – the ability to run linux-based applications on the CAVE being the primary one. This would be a benefit for applications that intend to use the Triton supercomputer which is connected to our CAVE and runs Linux (having both the front-end and back-end run on Linux isn't strictly necessary, but might be beneficial for some applications, especially if using the high-speed Infiniband connection and/or distributing computations between Triton and the client computers in our CAVE). However, at the moment most applications developed for our CAVE use the Unity engine and thus run on both Windows and Linux, with Windows having better support. Additionally, not all of our utility software (for managing projectors, tracking, and management of virtual reality applications themselves) are available for Linux. Other advantages of Vulkan, such as lower latency and better performance, are shared with DirectX 12. Thus, for applications that want to push graphics

quality as far as possible, or use very complex scenes, while still maintaining the high frame rates necessary for comfortable VR experience and stereoscopy, either of those two APIs can be used.

Microsoft DirectX is a collection of APIs for multimedia functionality in applications, such as games and video, on Microsoft platforms: the Windows operating system and Xbox gaming console. Direct3D – the component dealing specifically with the rendering of 3D scenes, is the flagship part of DirectX and thus names DirectX and Direct3D are often used interchangeably. The DirectX software development kit (SDK), which is now a part of the Windows SDK, contains binary libraries, header source files and documentation. The latest edition of Direct3D is version 12, sharing many similarities with Vulkan – such as the low level approach, lowered CPU utilization and better support of multithreading [13].

Unity Engine allows rendering to texture, a mechanism which can be used to display a stereoscopic image with a separate application – e.g. using DirectX 12 or Vulkan. It is crucial to keep the latency low, because otherwise virtual reality can cause discomfort for users. Thus, to speed up communication between the Unity application and the window used for display, shared memory can be used. The Unity Engine would write rendered images to the shared memory, while the window would read and display them on the screens, properly synchronized with the active glasses. The engine offers the `Camera` object, which gives us control over rendered images. From the `Camera`, we can retrieve a `TargetTexture` and assign it to a `RenderTexture` object. This creates a new `Texture2D` object with the `Camera`'s texture. Calling the `GetNativeTexturePtr` method results in a resource address, which we can use to retrieve all the necessary data that needs to be shared with the DirectX 12 or Vulkan process. The process takes the raw data and needs to recreate the resource objects. Once the resource objects have been created, they need to be updated every frame. The resource objects need to be linked with the previously written `Shader`, which is tasked with separating the left and right eye pictures into appropriate 'backbuffers'. Using a `Quadbuffer` allows for fluent rendering of stereoscopic images, since the engine can render two images (back left and back right) at the same time, while two other (front left and front right) are beings displayed – either one after the other (in case of active stereoscopy) or simultaneously if passive stereoscopy is used.

V. Conclusions

We found out that the Vulkan and DirectX 12 APIs offered enough low-level control to serve for our purposes, while also offering possible performance benefits. Vulkan is also supported on more platforms, which might make it the best choice for applications that require the computing power of a supercomputer – such as the Tryton cluster connected to our CAVE, which runs on Linux. However, these APIs have disadvantages as well – due to their low level design and novelty, maintaining a library based on them would require extra work.

The Unity Engine, over time, began to support different stereoscopy technologies and other virtual reality technologies. Their main focus were HMD (Head-Mounted Display) devices, but it was also possible to use these new features for CAVE systems. However, unlike HMDs which are now mass-produced to a few specifications, CAVE systems vary a lot, each being a unique installation, with different resolutions, sizes, projector positioning, tracking systems, etc. which means many parameters have to be set by hand (such as projection matrices, viewports). Nevertheless, using Unity's native support mechanisms has many benefits – the engine unifies the implementation of stereoscopy for the different graphics APIs (DirectX 11, DirectX 12, Vulkan, OpenGL Core, Metal) and operating systems. There's also no need to share textures between two rendering contexts – thus speeding up and simplifying the application. The engine itself also performs a multitude of optimizations (e.g. Single-Pass Stereo rendering). Support of control devices is also easier – especially now that there was no need for a second rendering window.

In conclusion, using the Unity Engine as basis for our library proved to be the easiest solution, offering many advantages in terms of simplicity and ease of use, both from library and application developer point of view. The main drawback is that it forces applications to be developed in Unity, which may not be the preferred development environment for all developers. It also has other limitations, and we may have to eventually expand our library to directly use graphics APIs, likely Vulkan or Direct3D 12, when the need to develop applications using other engines (such as the Unreal Engine or VBS engine) arises in our CAVE.

References

[1] S. Gateau, D. Filion, "Stereoscopic 3D Demystified: From Theory to Implementation in Starcraft 2," Game Developers Conference GDC 2011, http://www.nvidia.com/content/PDF/GDC2011/Stereoscopy.pdf .

[2] S. Gateau, S. Nash, "Implementing Stereoscopic 3D in Your Applications," GPU Technology Conference GTC 2010, https://www.nvidia.com/content/GTC-2010/pdfs/2010_GTC2010.pdf .

[3] I3DVL, "Immersive 3D Visualization Lab," https://eti.pg.edu.pl/i3dvl .

[4] J. Lebiedź, "3D visualization," Proceedings of the Polish Conference on Computer Games Development WGK 2013 (in Polish), vol. 3, Gdańsk 2013, pp. 105-115.

[5] J. Lebiedź, J. Redlarski, "Applications of Immersive 3D Visualization Lab," 24th International Conference on Computer Graphics, Visualization and Computer Vision WSCG 2016 – Poster Papers Proceedings, Plzeň 2016, pp. 69-74.

[6] P. Łapiński, Vulkan Cookbook, Packt Publishing 2017.

[7] A. Mazikowski, J. Lebiedź, "Image projection in Immersive 3D Visualization Laboratory,".18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems KES 2014, Procedia Computer Science 35, 2014, pp. 842-850, http://dx.doi.org/10.1016/j.procs.2014.08.251

[8] Unity Documentation, "How to do Stereoscopic Rendering," 2018, https://docs.unity3d.com/Manual/StereoscopicRendering.html .

[9] "NVIDIA 3D Vision Pro And Stereoscopic 3D," 2010, http://www.nvidia.com/docs/IO/40505/WP-05482-001_v01-final.pdf

[10] Unity Documentation, "Single-Pass Stereo rendering," 2018, https://docs.unity3d.com/Manual/SinglePassStereoRendering.html

[11] Microsoft Developer Network, "OpenGL on Windows", 2018 https://msdn.microsoft.com/en-us/library/dd374293(v=vs.85).aspx

[12] The Khronos Group, Inc. 2018, https://www.khronos.org/vulkan/

[13] Microsoft Developer Network, "Direct3D 12 Programming Guide", 2018, https://msdn.microsoft.com/en-us/library/windows/desktop/dn899121(v=vs.85).aspx

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 897–906
DOI: 10.15439/2018F174
ISSN 2300-5963 ACSIS, Vol. 15

Assessing the Communicability of Human-Data Interaction Mechanisms in Transparency Enhancing Tools

Patrick Barreto
Institute of Computing
Fluminense Federal University
Niterói, RJ, Brazil
Email: patrickbarreto@id.uff.br

Luciana Salgado
Institute of Computing
Fluminense Federal University
Niterói, RJ, Brazil
Email: luciana@ic.uff.br

José Viterbo
Institute of Computing
Fluminense Federal University
Niterói, RJ, Brazil
Email: viterbo@ic.uff.br

***Abstract*—The growing practice of accumulating personal data to generate predictions about users, leverages the need for mechanisms that allow people a more effective control of their data. An emerging field of studies called Human-Data Interaction (HDI), proposes the inclusion of human at the center of the data flow, providing mechanisms for citizens to interact explicitly with the collected data. Researches in HDI have discussed ways to offer Transparency Enhancing Tools (TETs), i.e., tools that support people on HDI issues related to privacy and personal data protection. Many works conducted about TETs focuses on usability issues, exploring aspects such as efficiency, user satisfaction and ease of learning. In this work, on the other hand, we aim to assess the communicability of HDI mechanisms in TETs. Hence, we applied the Semiotic Inspection Method (SIM) to investigate if and how HDI concepts are applied in two different TETs used for personal data management. We triangulated results from the study with findings from another investigation about communicability issues carried out in the same domain, but by observing and interviewing users.**

I. Introduction

The evolution of mobile devices, such as smartphones, tablets and sensors, influenced the society lifestyle as a whole by making more flexible the access to various services on the internet, and, eventually, by bringing advances in processing capacity and agility in mobile communication. As a result, people have begun to consume and share a significant and ever-increasing amount of data on their daily lives, which encompasses, for example, social information, events, health, lifestyle and consumption habits [1].

In ubiquitous computing scenarios, data collected by monitoring the user's activities can be used in analyzes and inferences to extract information about the behavior of individuals [2], [3]. In this scenario, data is used to make predictions related, for example, to the health status of people [4] or consumption trends [5]. Thus, the growing practice of accumulating personal data and generating inferences or predictions from them, leverages the need for research and creation of mechanisms that allow people a more effective interaction in this process of data manipulation [6].

Given this scenario, an emerging field of studies called Human-Data Interaction (HDI) proposes the inclusion of human at the center of the data flow, providing mechanisms to citizens to interact explicitly with these systems and the associated data [7]. The purpose is to enable users to understand by whom and in what form their respective data is used, and how to promote desirable effects and avoid undesired consequences [1].

Research in HDI is still incipient and has been gaining strength in the last five years, although some works in the Information Systems area reflect similar questions, among which are: transparency through open data, storage and use of personal data on the daily life of individuals [8] and privacy of information [9], [10]. Other works [11], [12], however, discuss ways to offer tools that support people on HDI issues related to privacy and personal data protection, and propose Transparency Enhancing Tools (TETs).

In [11], for example, the authors consider aspects of usability for TETs, proposing an interface prototype that seeks to offer the user a comprehensive view of their data stored and made available in different online services. This prototype is based on visualization techniques, seeking to associate the personal data of a user to the service for which this data was shared. The purpose is to provide transparency to users about their personal data collected by online services.

In [12], the authors presented PrivacyInsight, a software that allows the user to access their personal data, as well as the flow of data between interested entities involved in the storage and processing of this data. In addition, the tool enables the user to perform actions on their data in an indirect manner, i.e., through requests for correction or removal, thus providing means to exercise their right granted by law. This tool is based on the European Data Protection Directive (Council of European Union. 2016. Council regulation (EU) no 679/2016 - General Data Protection Regulation 95/45/EC[1]) and on usability requirements identified in the design of this solution.

Finally, in [13] the authors propose two Privacy Design Standards to facilitate the development of applications through solutions to recurrent privacy problems, focusing on trans-

[1]https://eur-lex.europa.eu/eli/reg/2016/679/oj

parency as a measure to establish broad control for users about their personal data. These standards are based on a set of factors, such as objective, user profile, usage context, problem, solution and consequences.

Such works, therefore, focus on usability issues in TETs, exploring aspects such as efficiency, user satisfaction and ease of learning. Our work, on the other hand, deals with the communicability [14] of mechanisms that enable Human-Data Interaction. For this purpose, we adopt a theory that considers the interaction of human beings with the interfaces of computational systems as a particular case of metacommunication, called Semiotic Engineering [15]. Some research questions are raised in the light of this theory, among which we highlight: How concepts of HDI can be communicated to users through TETs?; How interfaces can create means for the users to act in the data manipulation process?; and Which interactive elements can be used to communicate to the user about opportunities to use the data and its associated value?

In this context, this paper aims to investigate if and how HDI concepts are efficiently communicated (communicability) in two different TETs used for personal data management. For this, we apply the Semiotic Inspection Method (SIM) for scientific purposes [15]. We triangulated results from the study with findings from another investigation about communicability issues carried out in the same domain, but by observing and interviewing 5 (five) users. This work is organized as follows. In Section 2, we set the theoretical framework for the addressed theme. In Section 3, we present the methodology used in this work. In Section 4, we discuss the results obtained. Finally, in Section 5 and 6 we present the triangulation and final considerations, respectively.

II. Theoretical Framework

In this section, we present the theoretical basis necessary for the generation of this work.

A. Human-Data Interaction

The HDI area is an emerging field of interdisciplinary studies, which aggregates elements not only from the various branches of Computer Science, but also from areas such as Law, Psychology, Behavioral Economics and Sociology [1].

The literature presents some papers that address HDI with a focus on data analysis based on aspects of embedded interaction. In this approach, HDI is related to "Human manipulation and making sense of large complex and unstructured datasets". In [16], HDI is defined as "the customized delivery problem, creating the context of data understanding from large datasets". Thus, those proposals are based on aspects of HDI for the design of visualizations that allow to generate insights on large volumes of analyzed data.

There is a second approach, based on the proposals of Mortier *et al.* [7], which considers broader and more complex aspects, seeking mainly to address the problem of the management and use of personal data in society in general [17]. In other words, HDI is related to the manipulation of data, mainly personal, based on human factors [7].

The former approach is the one adopted in this work. It is related to the scenario in which the development of technologies and services for data generating, sharing and manipulating have, in general, allowed people to be in contact with digital tools and artifacts for consumption or production of information. Thus, people can produce data both Consciously (profile data in social networks, use of physical activity tools), and Unconsciously (robots monitoring our search history, cookies recording our browsing history, inferences of interest created from our purchase or search history) [7], [1].

Such data can be accumulated by different organizations that can perform inferences about sensitive issues related to our lives (health or emotional state, consumption habits or political preferences, for example) [7], so that these different analyzes make it possible to influence the user's behavior in a variety of ways [1]. Based on this perception, research in HDI seeks to address the new issues arising from the use of this ecosystem of personal data between different interested entities and their impacts on the actions of individuals and in society.

In [7], the authors establish three fundamental principles which address the challenges tackled in Human-Data Interaction, such as: Legibility, Agency and Negotiability. Legibility is concerned with making data acquisition and analytic algorithms more transparent and understandable to users, since, in general, interactions with data flows and processes are often obscure to people. Agency aims to provide individuals with the means to manage their data and their access by third parties, as well as to seek effective ways of acting in these systems, to the extent that individuals find it appropriate. This includes not only the ability to opt in or out of data collection and processing, but also the broader ability to engage with data collection, storage and use, and to understand and modify data and inferences. Finally, Negotiability is concerned with the various dynamic relationships that arise from data processing. This topic covers, for example, how understanding and individual attitudes change over time.

B. Semiotic Inspection Method

In this research, we used the Semiotic Inspection Method (SIM) for scientific purposes [15], a method of qualitative evaluation in Human-Computer Interaction (HCI), based on Semiotic Engineering [14]. With this method, the evaluators can analyze the communicability of the interactive artifacts [18]. The focus is to inspect the metacommunication from designer to user with the objective of identifying possible breakdowns in communication. First, in the preparation stage, the evaluation focus, the user profile and the inspection scenario are defined. In the evaluation stage, the evaluator examines the interface and classifies the signs as metalinguistic, static or dynamic.

Metalinguistic signs are the first to be analyzed, since they explicitly express and explain other parts of the metacommunication of the designer. This class of signs is usually found throughout the interface, either in instructions, explanations, warnings and error messages, with a focus on online help and user's manuals [19].

Static signs are those that communicate their meaning regardless of cause and effect relationships and can be interpreted from instant canvas pictures. Thus, they express the state of the system at a given time. They are represented by the elements present in the interface screens (or equivalents in non-visual interfaces), such as labels, images, text boxes, buttons, menus, etc., as well as layout, size, color, font and other characteristics. Its analysis should consider only the interface elements presented in each screen at an instant of time, without examining neither the behavior of the system, nor the temporal and causal relationships between interface elements [19].

The inspection of dynamic signs requires that in the analysis, the evaluator inspects the interaction process that the user can experience through the interface. These signs are perceived through changes in the interface that communicate to the user the behavior of the system as a result of user actions (clicking the mouse, pressing enter, changing the focus from one form field to another, etc.), by external events (receiving email, Internet connection failure, etc.) or over time. Dynamic signs are usually represented by animations, opening and closing dialogues, transitions between screens, or modifications to the elements of a screen (for example, activating a button, updating a text or image, modifying the layout of some interface elements, etc) [18].

To inspect the interface, SIM proposes 5 steps to be followed by the evaluator [20]. In the first three steps, the main goal is to reconstruct the metacommunication of the designer for each category of signs (metalinguistic, static and dynamic), using the following meta-model of the designer [20]: *"Here is my understanding , who you are, what I learned that you want or need to do, in what preferred ways and why. This is the system that I have therefore created for you, and this is how you can or should use it for meet a variety of purposes that fall within this version"*. The steps are:

- Step 1: Inspection of metalinguistic signs. In this step, the evaluator explores the documentation and help system.
- Step 2: Inspection of static signs. In this step, the evaluator inspects the static signs of the interface.
- Step 3: Inspection of dynamic signs. In this step, the evaluator inspects the signs that emerge from the interaction.
- Step 4: In this step, the evaluator contrasts and compares the metacommunication messages from steps 1, 2, and 3 and records possible problematic interpretations that may occur in user interaction time.
- Step 5: Appreciating the quality of metacommunication. In this step, the evaluator produces a report containing the communicability problems encountered, which may frustrate or prevent the user from understanding the message intended by the designer, affecting his productivity. In this method, the evaluator is the advocate of the user.

SIM can be used in scientific contexts and generate valid knowledge in HCI [15]. To do so, two other steps must be considered when applying the method. During the preparation phase, it is necessary to define the research question that researchers are looking for an answer. Also, after the application, a Triangulation step is added to the analysis. Triangulation involves the generation of other results (e.g. by other specialists or through compatible methods) which valides scientifically the results obtained through SIM.

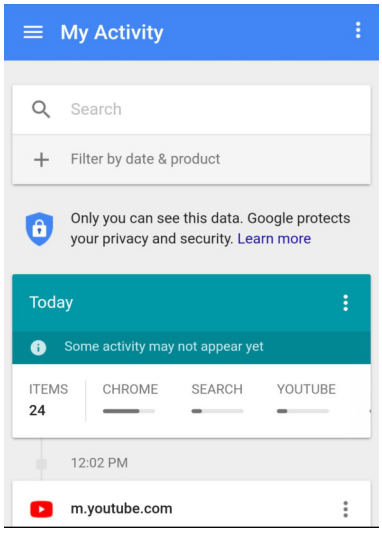


Fig. 1. MyActivity home screen.

III. Methodology

The methodology used in this work consists, in large part, in the application of the Semiotic Inspection Method (SIM), described in the previous section, with the focus on the evaluation of the communicability considering the concepts presented by the HDI theory. We use, therefore, the predictive paradigm, making use of an interpretative and qualitative method [21]. The inspections were carried out by two evaluators together, being one a junior level evaluator and the other a senior level evaluator (specialist). Based on the application of the SIM steps, we sought to answer two research questions (RQs): (i) Which are the communication strategies that potentially enable Human-Data Interaction?; (ii) What is the relationship among the elements found in the first question and the main concepts of HDI proposed by Mortier *et al.* [7], i.e., Legibility, Agency and Negotiability?

There are few TETs tools for data management that provide ways for controlling data. In addition, there are no records that these tools were designed based on the concepts that we adopt as the foundation of HDI. Thus, we performed two studies (S1 and S2) to identify traces of the application of the fundamentals of HDI in TETs. To do this, we have selected two tools: Google MyActivity [22] and Privacy Badger [23].

MyActivity allows users to exercise greater control over the data generated by the monitoring of their activities in Google's services and products. Privacy Badger is a add-on that aims to restrain the action of third-party domains that seek to collect data through unauthorized monitoring of user activity while he is browsing the web. Figures 1 and 2 show the interface of the home screen of both tools.

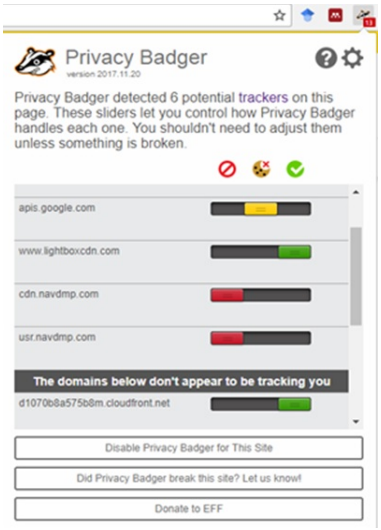


Fig. 2. Privacy Badger Home Screen.

These tools were chosen because they seek to offer the user an understanding of how their personal data are being collected or used by interested entities. Moreover, both tools provide the user with forms of control over the access and use of their data. The evaluation focus, the user profile and the reference inspection scenario in both tools were defined in the preparation stage of S1 and S2, as described below:

A. Preparation Stage for MyActivity

In what follows, we describe the evaluation focus, the user profile and the inspection scenario defined for the evaluation of MyActivity.

- User profile: The user uses mobile devices or computer to browse the internet, searching for leisure and entertainment options, and to make online purchases. He gets surprised to have the possibility of exercising control over his personal data.
- Inspection scenario: Ane performs various tasks through her smartphone. Every day she accesses social networks, conducts research of professional and personal interest and seeks to find the best path in traffic. Ane has always opted to make her personal data available to her applications and services from large companies like Google. With this, Ane seeks to experience a navigation based on their tastes, interests and types of content consumed, thus avoiding receiving unnecessary notifications or content. Recently, Ane was notified by Google about MyActivity, their Data Management tool. According to MyActivity's proposal, Ane has found that she can manage the Google's services to made them more useful. Ane was surprised to know that she can exercise control over her data or activities carried out through her smartphone. Hence, when Ane accesses Google's MyActivity, she wants to perform the following tasks: (a) To find out what types of data are being stored or monitored by Google; (b) To exercise some intervention in the availability and access to her data.

B. Preparation Stage for Privacy Badger

In what follows, we describe the evaluation focus, the user profile and the inspection scenario defined for the evaluation of Privacy Badger.

- User profile: The user uses mobile devices or a computer to browse the internet, searching for leisure and entertainment options, or shopping online. However, he is concerned about his privacy, i.e., preserving data about his browsing history against third parties.
- Inspection scenario: Bob uses his personal computer often to read news, emails, search products, shop online, access Internet banking, interact in social networks and search about leisure options. Bob manipulates his personal information to perform a good part of these actions. Thus, concerned about the risk of invasion of privacy by tracking robots (trackers), he resorted to some tools, among them, the Privacy Badger. Hence Bob wants to accomplish the following tasks when using this tool: (a) Identify all possible trackers that can monitor his activity when using the internet; (b) Block monitoring trackers.

IV. Results

In this section, first, we present the classes of signs found with the SIM inspection. Then, for both studies (S1 and S2), we tried to answer research questions (i) and (ii). For this, we analyze the main communication strategies found during the inspection, which involve the communication of mechanisms that enable Human-Data Interaction. These communication strategies were identified from traces of the application of the three main concepts of HDI theory: Legibility, Agency and Negotiability. Finally, we performed a comparative analysis between studies S1 and S2.

A. Classes of Signs

In this section, we present some visual design options that were identified through SIM. These options represent visual cues used to interact with systems and have been adopted by the designers of these applications. The MyActivity tool uses all of the identified types, while the Privacy Badger tool does not offer the 'Cards' and 'Modal' classes, as follows:

- **Cards** are the registered activities of the user. Each activity is represented by a title (showing the service used), a link for the activity performed preceded by a keyword that characterizes the type of activity recorded, such as 'Watched' (Videos), 'Visited' (Web Pages), 'Searched' (Google Search) or 'Viewed Area' (Use of Maps), for example. In addition, 'details' or 'delete' options are displayed in each activity log. Generally, a figure can be associated with an activity, and activity groupings can be done automatically to summarize the display of records. Thus, an option is displayed in the card footer if the user wants to view items that have been deleted. It is interesting to note that all activities compose a

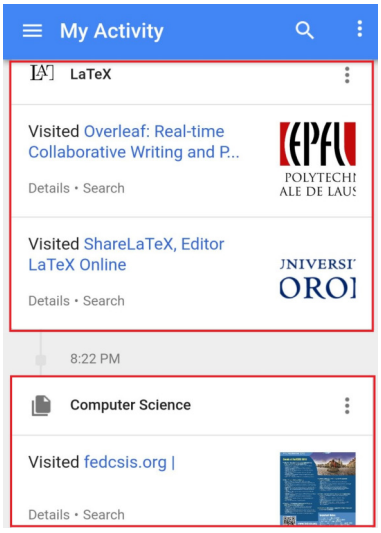


Fig. 3. Example of MyActivity 'Cards'.

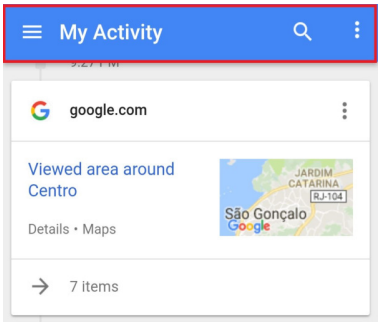


Fig. 4. Example of MyActivity Persistent Menu.

record history, so that these records are distributed in their respective days. Each card, which may be an activity or a grouping of records, is linked by a timeline with an associated timestamp. Figure 3 shows an example of cards.

- **Persistent Menu** allows the user to access a set of options at any time during their interaction with the tool. See the example in Figure 4.
- **Modals** are alerts usually displayed to confirm user actions, with texts that briefly describe their consequences. Figure 6 shows an example of a Modal in MyActivity.
- **Search Filters** allow the user to refine the view of records by context. For example, in MyActivity, you can select service types, date range, and more. In Privacy Badger, you can choose the status associated with the domain, for example. Figure 5 shows the example of a filter used by the Privacy Badger.
- **Sliding Buttons** represent user preferences regarding his privacy, i.e., what can be seen about him by entities interested in his personal data. Figure 7 shows examples of Sliding Buttons in Privacy Badger.

B. Communication Strategies in MyActivity

Based on the mapping of metalinguistic, static and dynamic signs, and their respective metamessages, the communication strategies identified in the MyActivity tool were:

- **CS1:** Provide different categorizations of the collected data;
- **CS2:** Show monitored activities and the level of use of products/services;
- **CS3:** Offer forms of intervention on data collection;
- **CS4:** Provide alerts of actions performed by the user;
- **CS5:** Provide means to report problems or collaborate with ideas.

About Communication Strategy (CS1), MyActivity, linked to MyAccount, gives the user a list of categories of data that can be collected about him, such as 'Location History' or 'Device Information'. For each category, a brief description of the purpose of the collection is presented. If he would like more information on this, the 'learn more' link will direct him to the 'Help' page. From this categorization, MyActivity also provides more specific categories of user activity on Google services, such as "Feedback no interest in YouTube" and "Location responses". In doing so, we believe that design intent was to provide a means for the user to have a comprehensive view on what types of data can be collected by Google, as well as to understand the company's objectives in acquiring the data of its users. In our interpretation we evaluated that this strategy points to some aspects of the concept of Legibility in HDI, since the tool makes available to the user information about who is monitoring him, the means used to perform the data collection, what types of data will be collected, and the intended interest in that process. However, we did not find explanations about the algorithms and methods used to generate inferences from the user data.

About Communication Strategy (CS2), MyActivity allows the user to access and review the history of activities he performed. By default, the most recent activities are displayed at the beginning of the history, i.e., in reverse chronological order. In addition, the user has access to the 'details' option that provides explanations on how monitoring is performed. The search engine helps the user find a specific set or activity in their activity history, giving the user better navigability, since a large number of records are expected. MyActivity also offers two types of activity view: packet-based (records are listed individually), or product/service-based (groups a sequence of records by their respective product/service). He can also check how often he uses the services each day. The activity history query should be based on its category. However, initially this may not be communicated appropriately to the user, since, by default, when accessing MyActivity, activities related to a category, usually 'Web and Apps Activity', are listed. Hence, we can observe indications of the concept of Legibility in this strategy, since the tool provides resources that inform the user about the capture of his activities in Google services and products.

Fig. 5. Menu filter in Privacy Badger.

About Communication Strategy (CS3), MyActivity, linked to MyAccount, offers the user ways to intervene on the data collected about him. For each data collection category, he can define whether monitoring is enabled or not. By default, some monitoring types are already enabled and others are disabled (paused). However, the user has the right to intervene, at any time, on the type of monitoring that he wants to pause or enable. The tool uses color to distinguish monitoring status, applying grayscale to 'paused' monitoring, and color to 'enabled' monitoring. In this case, we understand a possible intention to use colors is to communicate to the user about possible benefits from collecting their data. MyActivity also allows the user to delete activity logs. In this case, excluding records implies disregarding such data in the aggregation and processing made by Google. Therefore, we can perceive in this strategy the perspective addressed by the concept of Agency, so that the user has mechanisms to determine what types of data can be accessed and collected, as well as how his records are generated and deleted.

About Communication Strategy (CS4), MyActivity provides alerts when the user wants to take actions on his data, such as pausing/enabling a type of monitoring or deleting an activity, for example. These notifications tell the user about the implications of the intended action, generally seeking to encourage him to make his data available to Google, showing how this reflects benefits in delivering services or products. In addition, clarifications on how to collect and store such data are also communicated, thus proposing clarity and transparency, aiming at user confidence. From this, we can notice a link with the concept of Legibility, because this strategy seeks to deal with the user's concerns about his data and the processing that is performed from them.

About Communication Strategy (CS5), MyActivity presents a space that offers the user a feature to report usage problems or errors in the activity log. In this way, however, it is necessary to wait for the analysis of the request before it can be acknowledged. This mechanism also acts as a channel for sharing ideas or suggestions. Thus, we can establish a link with the concept of Agency, because this strategy allows the user means to inform and correct the data provided.

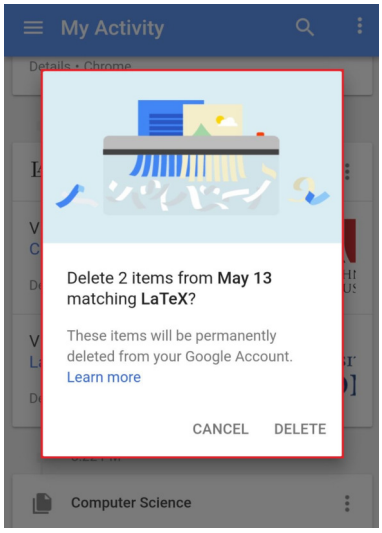


Fig. 6. Modal Example in MyActivity.

C. Communication Strategies in Privacy Badger

Based on the mapping of metalinguistic, static and dynamic signs, and their respective metamessages, the communication strategies identified in the Privacy Badger tool were:

- **CS6:** Offer forms of intervention regarding the use of data;
- **CS7:** Show the third-party domains identified.

About Communication Strategy (CS6): In general, the content of a web page can come from a number of different sources, i.e., third-party domains. On an e-commerce page, for example, the display of products will be carried out by a virtual store, the search engine can be from a company hired to make this service available and the ads will be from an advertising company. In this way, Privacy Badger analyzes the behavior of all identified third-party servers as the user navigates through different domains, applying one of the following statuses: 'Blocked', 'Partially Blocked' and 'Allowed'. If any domain is attempting to monitor the user's browsing record via cookies without permission, then PB will automatically block the content coming from that server. If this domain is providing an important content type for the page to work, then Privacy Badger will allow connections to this server, but will block its cookies injection in the browser. Finally, Privacy Badger will allow third-party content to be injected if no monitoring activity is detected.

The tool also allows the user to modify the applied status, thus offering the power to intervene in the result of the classification algorithm. In this way, this domain will be classified according to the decision of the end user. In addition, PB offers a local white-list where the user can add domains he trusts, so that they are out of the Privacy Badger analysis. Thus, we can observe a correlation with the concept of Agency in supporting the user to exercise control over access to their browsing data. However, the tool does not allow the user to

Fig. 7. Sliding Button in Privacy Badger.

define which types of data he wants to make available, for example.

About Communication Strategy (CS7): The Privacy Badger provides a list with the address and status of each identified third party domain, allowing the user to be aware of which third-party domains are active, as well as those who have attempted to perform some kind of hidden monitoring. However, the tool failed to communicate information about the identified domains, such as the type of content or what functionality a particular server was attempting to enter, which could assist user understanding of how third-party domains act. Thus, we can point out, in this strategy, a reference to the concept of Legibility based on the identification of which domains have attempted to collect the user data. However, the user will not be informed about some important aspects of the data collection, storage and processing, such as the purpose of the data collection or what types of data, in fact, can be collected about him.

D. Comparative Analysis between S1 and S2

One of the main challenges observed during the inspections was the partial application of the aspects advocated by the concepts of HDI, pointing to possible barriers or limitations in the adoption of human factors in relation to the use and storage of personal data by third parties. On the other hand, we have identified the possibility of enabling Human-Data Interaction without all traces of its concepts being present.

Regarding the concept of Legibility in HDI, we identified some of its aspects present in Communication Strategies CS1, CS2 and CS4 (Study S1) and in Communication Strategy CS6 (Study S2). We understand that this concept represents the first step in enabling Human-Data Interaction, setting guidelines that support people's understanding of the actions of third-party that are interested in their personal data. The communication problems encountered may point to resistance in making more transparent the algorithms used to infer new knowledge about people. This corroborates Mortier's observation in [7], when mentioning the conflict in making public such algorithms that are intellectual property of companies. However, it was possible to identify important mechanisms related to Legibility aspects, such as the classification of monitored data, the identification of those who want to access personal data, the forms of collection used and the intended goals of the interested parties. However, from the possible communication problems reported in S2, we noticed the more timid use of this concept by the Privacy Badger tool.

In relation to the concept of Agency in HDI, traces of this concept were also perceived in both tools. We can see that the comprehensive use of Legibility offers a more propitious context for the adoption of Agency aspects. In other words, if people are not aware of who they are interested in collecting their data, what their intended data are, their collection methods and desired goals, then the ability of people to act on their data is limited, so that there will be no important information available to users in order to support them in their decisions or to assist in the creation of more adequate criteria and controls on the access and use of their data by third parties. Therefore, we note that the MyActivity tool communicates these aspects more clearly, reflecting the concept of the Agency more comprehensively.

Finally, no traces of the application of the aspects related to the Negotiability concept were found in the strategies identified. This may point to the difficulty in defining mechanisms that allow identifying characteristics that are expressed in data and are likely to change over time, such as individual attitudes and interests, for example. However, these issues are relevant in the context of processing and generating inference from personal data. The Table 1 summarizes the answers by each research question.

TABLE I
SUMMARY OF THE ANSWERS TO OUR RESEARCH QUESTIONS OBTAINED IN S1 AND S2.

RQ(i)	RQ(ii)
CS1: Provide different categorizations of the collected data.	Legibility
CS2: Show monitored activities and the level of use of products/services.	Legibility
CS3: Offer forms of intervention on data collection.	Agency
CS4: Provide alerts of actions performed by the user	Legibility
CS5: Provide means to report problems or collaborate with ideas.	Agency
CS6: Offer forms of intervention regarding the use of data.	Agency
CS7: Show the third-party domains identified.	Legibility

V. TRIANGULATION

Triangulation is a standard procedure in the validation of qualitative research results [24]. In this case, we validated the results obtained from applying the SIM (S1 and S2), comparing them with the results obtained from a set of interviews and observation sessions with users, thus seeking to identify convergences and divergences, ensuring the scientific validity of the results.

The triangulation step counted on 5 (five) participants. Before starting the tests, a consent form was presented to the

participants, with information about the data collection and use conditions performed in this research. With due acceptance, the next step was to make a brief presentation of HDI for each participant, exploring the three main concepts proposed by Mortier *et al.* [7] . Then the tests were started.

We also used the inspection scenario described in sections III-A and III-B, with some adaptations in their tasks to carry out this study with participants, in order to generate a good comparability of the results obtained in S1 and S2. After completing each task, participants responded verbally to a post-test questionnaire. With the participants' speech, it was possible to identify and highlight the convergences and divergences, regarding research questions proposed in this paper. The convergences and divergences identified will be presented in the following subsections A and B. The Table 2 summarizes these results obtained in the triangulation step.

TABLE II
SUMMARY OF THE CONVERGENT AND DIVERGENT RESULTS IDENTIFIED IN THE TRIANGULATION STEP.

	Convergences		Divergences	
CSs	RQ(i)	RQ(ii)	RQ(i)	RQ(ii)
1	P1, P2, P3, P4, P5	P1, P2, P3, P4	-	P5
2	P1, P2, P3, P4, P5	P1, P2, P3, P4, P5	-	-
3	P1, P2, P3, P4, P5	P1, P2, P3, P4, P5	-	-
4	P2, P3, P4	P2, P3, P4	P1, P5	P1, P5
5	P2, P3, P4, P5	P3, P4, P5	P1	P1, P2
6	P2, P3, P4	P2, P3, P4	P1, P5	P1, P5
7	P2, P3, P4	P3, P4	P5	P1, P2, P5

A. Convergences

In this section will be presented, for each communication strategy, the discourse excerpts that show the convergences identified, related to the answers questions RQ(i) and RQ(ii).

As evidence for the CS1, provide different categorizations of the collected data, the following discourse excerpts were collected:

P1: "From my point of view, Google is keeping an eye on my interests, especially those that move my everyday life. So it seeks to monitor my clicks, texts, videos, comments and where I went. This is related to the concept of Legibility."

P2: "... I realize that Google wants to know your location, the places you've visited, what you search for (search engines, Youtube and Play Music) because this will help them in their recommendations. This is related to the concept of Legibility because the tool shows what Google is interested in doing with this data, but not in depth. That is, it tells what to do and with what, but does not say how it will do."

P3: "You can easily see what data Google has about you. In this case, it is the concept of Legibility involved."

P4: "Google has pretty much everything about me, like location, places I've visited and the time that happened, what I watched or did on Youtube, my vocal signature... The Legibility is not completely applied because I realize that there is still a certain lack of transparency on forms of using my data and its purposes."

P5: "Yes, for example the location, types of songs, what I see on Youtube."

As evidence for the CS2, show monitored activities and the level of use of products/services, the following discourse excerpts were collected:

P1: "I can see what Google has recorded about me. This is associated with Legibility, since there are texts or words like "Learn More" that help you understand what has been recorded."

P2: "It is possible to visualize, including the circuit that the user performed in a locality, a kind of history of how you visited a place. Thus, it is possible to associate with the concept of Legibility because you are aware about what it is monitoring."

P3: "It is possible to consult the records by means of histories, such as the one of location, that allows to identify even the route accomplished, in an easy and organized way for the user. So, this view fits into the concept of Legibility, because it is very clear what was recorded i.e., a query."

P4: "MyActivity shows the history of all these types of activities ... so it's Legibility."

P5: "MyActivity logs (in the case of Chrome), not only which site I've visited, but which sections of the site I passed. Here I think it's Legibility because it shows exactly what I did."

As evidence for the CS3, offer forms of intervention on data collection, the following discourse excerpts were collected:

P1: "I didn't trust too much in controlling my data because we don't have a policy or something that makes that control more present in our everyday lives. However, with the option of downloading my data, I was more reassured. This is tied to the concept of Agency."

P2: "I realized that in some cases these options are hidden. But you can control what they can and can't access your data. It's related to the Agency, by allowing action on the data."

P3: "It is possible, for example, to both exclude and prevent them from continuing to monitor you. So there are Agency these options that give you control over your data."

P4: "I can erase history, intervene in some things ... the minimum exists, which is the case of being able to delete, modify, allow or deny. Agency, but in this case, I can not manage for third parties, but only for myself."

P5: "It does provide an option of what I can release or not. I think it's Agency, because it's a way to conduct management over my data, it's you showing what you want to happen."

As evidence for the CS4, provide alerts of actions performed by the user, the following discourse excerpts were collected:

P2: "The pros and cons of letting them monitor their activities are clear. These alerts reinforce what your action will cause, not allowing the user to simply take action, forcing the user to heed it. Thus, the concept of Legibility appears again."

P3: "They provide more in-depth information about your data. Soon, it becomes Legibility."

P4: "It gives me more in-depth information when I click activate or when I want to pause. Legibility."

As evidence for the CS5, provide means to report problems or collaborate with ideas, the following discourse excerpts were collected:

P2: "It is possible to report problems, but it is not known whether this will be answered or not."

P3: "MyActivity even allows screenshots to be sent to help identify the problem. They even allow modifications for legal reasons, and as it is a global company, then it is important to have that same option. It is related to Agency."

P4: "It offers a way to report legal issues, but no issues with third parties. However, the screen passes generic information so I guess it could be for any kind of problem as well. On the other hand, the system provides a 'learn more', so maybe it is the case to go and read more about it. It would be the Agency."

P5: "In case it is the feedback option. This option allows me to participate, either to complain or to suggest or make some kind of contribution. So I think it's more related to the Agency than to the other two concepts."

As evidence for the CS6, offer forms of intervention regarding the use of data, the following discourse excerpts were collected:

P2: "I've been able to distinguish the intentions of third-party domains by colors and by some domain names, i.e. those that are interested in my behavior or not. This identification can relate to the concept of Legibility."

P3: "The total he gives you easily before you click. And when you click, it gives you the names, so it's very easy to identify these things in it. This relates to Legibility (a bit), because it gives you just that this domain is trying to monitor you, but you do not know what it's registered for, i.e., it's the minimum level of Legibility."

P4: "I found it very good to be able to identify because he showed me many domains that may be monitoring me, but on the other hand lacked more transparency because he does not describe very well the purpose of these possible trackers are acting. I think if I knew that, I could allow monitoring if it was to benefit people. It is related to the principle of Legibility."

As evidence for the CS7, show the third-party domains identified, the following discourse excerpts were collected:

P2: "I can enable and disable, for example, third-party domains. This is related to the concept of Agency. But the concepts applied here are weak, because it could provide more information about what each third-party domain wants to do, or what behavior the third-party domain has presented to the PB to block."

P3: "I managed to block some domains, but since it does not provide information on the consequences of this action, it might impact the functioning of the page. It is related to the concept of Agency, but only to the part of managing access."

P4: "The Privacy Badger allows me to perform the blocking or the release, so it is the minimum of management. Therefore, it is associated with the Agency concept."

B. Divergences

The divergences identified in the triangulation step can be classified into two cases: The first case comprises the participants who didn't identify some communicative strategy and, therefore, couldn't answer the RQ(i) and RQ(ii). The second case comprises the participants who were able to identify communication strategies as mentioned in the SIM, but associating them with other concepts of HDI, different from those pointed out during the SIM. In this case, the perceived divergences apply only to RQ(ii).

Regarding the first case, some communicative strategies were not perceived by a few participants, for many reasons. The examples identified were: P1 and P5 had contact with CS4. However, they did not consider it as an alert, but rather as another textual information presented in the view. P1 was also unable to locate CS5 nor understand CS7, as follows: *"The identified names of third-party domains made no sense to me."*. In addition to these examples, P5 couldn't to evaluate the Privacy Badger, because this tool did not offer the language desired by the user, as evidenced in his report: *"The interface does not make sense to me, because I do not know English, so I would not manipulate this program."*

Regarding the second case, for example, from the textual description presented for each activity category in MyActivity through CS1, P5 understood that activity monitoring allows users to better understand themselves over time through a possible processing of their historical records. This new understanding may be relevant in subsequent data exchanges, allowing the sharing of new reassessments in relation to their behavior or interests expressed in data, with others interested in their data. So, P5 linked CS1 with Negotiability concept, as reported: *"I think it's Negotiability, because we may want to be monitored for certain types of activities at some point, maybe for benefits. By knowing what types of activity I am monitored, it allows me to set up your profile and perceive, over time, changes in behavior or interests about me. It may be interesting that other people might know about these changes."*

In a second example, P2 considers that CS5 is related to the concept of Negotiability, because it understood that reporting problems can serve not only to report system failures, but also to express its considerations about the use of its data by third parties, as reported: *"It may have to do with Negotiability, because it allows you to negotiate about your data, saying what you disapprove of the use of data."*

There were two other divergences over CS7. P1 considers it to be related to the concept of Legibility because Privacy Badger automatically performs a possible block on a third party domain. Thus, the tool suggests that it is not necessary for the user to modify such controls, but only to observe the type of constraint applied to a particular third-party domain, as reported: *"The option to block a third-party domain may refer to the concept of Legibility because it is in my discretion to block it or not."* Finally, P2 considers that CS7 is linked to the Negotiability concept because *"it allows me to not only block or enable, but make a middle ground by blocking*

only the cookie and letting the third party domain perform its functionality on the page."

VI. Conclusion

This work aimed to identify and evaluate the communicability of strategies that enable Human-Data Interaction, through the application of the main HDI concepts proposed by Mortier *et al.*[7], in the context of personal data management. For this, the Semiotic Inspection Method was used to inspect the MyActivity and Privacy Badger tools. A methodology was established based on Semiotic Engineering theory [14], contributing to the consistency of the evaluation and results obtained.

The tools were inspected by two evaluators and, through the results obtained, it was possible to answer the research questions raised in this work. We found out evidences (see section IV) of high and low communicability in some Strategies to communicate Legibility, Agency and Negotiability. For instance, in CS1, because they are not explaining the algorithms and methods used in generating inferences from the user data (low legibility). In CS3, because the user has mechanisms to determine what types of data can be accessed and collected (high agency). The results also allowed us to present problems related to the partial application of the inherent aspects to the concepts of Legibility and Agency. However, it was observed that, even without the identification of all concepts related to HDI, it was possible to observe traits that allow Human-Data Interaction.

This paper brings three main contributions. The first is the application of SIM in an yet unexplored context of HDI related to the use of TETs, showing that the application of the method was relevant in that context. The second contribution is the identification of a set of communicative strategies and the classes of signs used by designers to make the HDI feasible. Such strategies may support designers of other TETs in their decisions about which strategies to use. Finally, the research results of the model proposed by Mortier *et al.* [7] in association with the application of the SIM, i.e., the inspection and evaluation by model, has shown to be promising in the HCI evaluation of applications that seek to provide means for Human-Data Interaction. The results presented here were validated (as is typical in validation of qualitative research) through an endogenous triangulation [25]. This motivates us to carry out new empirical studies with HDI designers to explore the practical effects of designing TETs with the communication strategies.

VII. ACKNOWLEDGMENTS

The authors want to thank the Brazilian funding agencies that support this project in different ways: CAPES, CNPq and FAPERJ. They would also like to express their gratitude to the volunteers who participated in the study.

References

[1] H. Hornung, R. Pereira, M. Baranauskas, and K. Liu, "Challenges for human-data interaction–a semiotic perspective," in *International Conference on Human-Computer Interaction*. Springer, 2015, pp. 37–48.

[2] J. Han, H. Ding, C. Qian, W. Xi, Z. Wang, Z. Jiang, L. Shangguan, and J. Zhao, "Cbid: A customer behavior identification system using passive tags," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2885–2898, 2016.

[3] C. Meurisch, U. Naeem, M. A. Azam, F. Janssen, B. Schmidt, and M. Mühlhäuser, "Smarticipation: intelligent personal guidance of human behavior utilizing anticipatory models," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1227–1230.

[4] A. Doryab, M. Frost, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Impact factor analysis: combining prediction with parameter ranking to reveal the impact of behavior on health outcome," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 355–365, 2015.

[5] F. Zhang, N. J. Yuan, K. Zheng, D. Lian, X. Xie, and Y. Rui, "Mining consumer impulsivity from offline and online behavior," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1281–1292.

[6] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, and J. Crowcroft, "Challenges & opportunities in human-data interaction," *University of Cambridge, Computer Laboratory*, 2013.

[7] R. Mortier, J. Crowcroft, D. McAuley, H. Haddadi, and T. Henderson, "Human-data interaction: The human face of the data-drivem society," 2014.

[8] E. W. Ritter and S. J. Rigo, "Fitdata: A system for monitoring physical activity based on mobile devices," in *Proceedings of the XII Brazilian Symposium - Volume 1*, ser. SBSI 2016. Porto Alegre, Brazil: Brazilian Computer Society, 2016. ISBN 978-85-7669-317-8 pp. 72:550–72:557.

[9] C. Buck and S. Burster, "App information privacy concerns," *AIS Electronic Library - Americas Conference on Information Systems*, 2017.

[10] C. Buck, "Stop disclosing personal data about your future self," *AIS Electronic Library - Americas Conference on Information Systems*, 2017.

[11] J. Angulo, S. Fischer-Hübner, T. Pulls, and E. Wästlund, "Usable transparency with the data track: a tool for visualizing data disclosures," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015, pp. 1803–1808.

[12] C. Bier, K. Kühne, and J. Beyerer, "Privacyinsight: the next generation privacy dashboard," in *Annual Privacy Forum*. Springer, 2016, pp. 135–152.

[13] J. Siljee, "Privacy transparency patterns," in *Proceedings of the 20th European Conference on Pattern Languages of Programs*. ACM, 2015, p. 52.

[14] C. S. De Souza, *The semiotic engineering of human-computer interaction*. MIT press, 2005.

[15] C. F. and Leitão and C. S. De Souza, "Semiotic engineering methods for scientific research in hci," *Synthesis Lectures on Human-Centered Informatics*, vol. 2, no. 1, pp. 1–122, 2009.

[16] F. Cafaro, "Using embodied allegories to design gesture suites for human-data interaction," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 560–563.

[17] A. Crabtree and R. Mortier, "Human data interaction: historical lessons from social studies and cscw," in *Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway*. Springer, 2015, pp. 3–21.

[18] C. S. de Souza, C. F. Leitão, R. O. Prates, and E. J. da Silva, "The semiotic inspection method," in *Proceedings of VII Brazilian symposium on Human factors in computing systems*. ACM, 2006, pp. 148–157.

[19] C. S. de Souza, C. F. Leitão, R. O. Prates, S. A. Bim, and E. J. da Silva, "Can inspection methods generate valid new knowledge in hci? the case of semiotic inspection," *International Journal of Human-Computer Studies*, vol. 68, no. 1-2, pp. 22–40, 2010.

[20] R. O. Prates, C. S. de Souza, and S. D. Barbosa, "Methods and tools: a method for evaluating the communicability of user interfaces," *interactions*, vol. 7, no. 1, pp. 31–38, 2000.

[21] S. Lewis, "Qualitative inquiry and research design: Choosing among five approaches," *Health promotion practice*, vol. 16, no. 4, pp. 473–475, 2015.

[22] "Google myactivity," https://myactivity.google.com/, acessado em 12/11/2017.

[23] "Privacy badger," https://www.eff.org/privacybadger, acessado em 10/01/2018.

[24] N. K. Denzin and Y. S. Lincoln, *The landscape of qualitative research*. Sage, 2008, vol. 1.

[25] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2009.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 907–911
DOI: 10.15439/2018F162
ISSN 2300-5963 ACSIS, Vol. 15

The functional design method for buildings (FDM) with gamification of information models and AI help to design safer buildings

JUKKA SELIN
Department of Information Technology
South-Eastern Finland University of Applied Sciences Ltd
Patteristonkatu 2, 50100 Mikkeli FINLAND
jukka.selin@xamk.fi www.xamk.fi

MARKKU ROSSI
RTD and Services
South-Eastern Finland University of Applied Sciences Ltd
Patteristonkatu 2, 50100 Mikkeli FINLAND
markku.rossi@xamk.fi www.xamk.fi

***Abstract*— We have developed a method that enables better taking into account of need of space of actions of moving objects in a building when Building Information Modeling (BIM) is used. The 3D spatial objects for real space requirements of actions created with our method can be used with various design tools as such in dimensioning or to generate video game colliders defining the need of space. Together with the gamified model of the structure from BIM they enable dimensioning in designs and simulations. We can dimension spatial objects for example related to safety. There are a lot of needs and applications for our methods. It is possible to design buildings and other structures that fit their purpose well.**

I. Introduction

In this article we present the utilization of the Functional Design Method (FDM) "Value Add Data / VAddD" together with the Gamified BIM Information Model during the simulation of emergency evacuation from a supermarket. The simulation relates to the joint R&D and piloting effort of the S-E Finland University of Applied Sciences (XAMK) and our industrial construction partners. The Functional Design Method is based on the Finnish patent by our senior lecturer and Licentiate of Science Jukka-Pekka Selin. The patent was derived from an earlier internal innovation report [1], [2], [3].

The core idea in the FDM is to capture the space requirements of real action in three dimensions and then create a new 3D spatial object. The new 3D objects represent the maximum spatial space requirements of actions. When using Construction CAD these objects can be used in dimensioning different spaces. The goal can be both to fit an action to a space or alternatively ensure that an object can reach all the needed places in a space. If the Information Model in the form of an IFC (Industrial Foundation Classes) file is gamified it becomes possible to use the spatial objects to realise e.g. adaptation capable game colliders around game objects. In that case we can test the fitting of different actions or reaching capabilities of objects to stationary parts of the construction. We can also perform simulations.

In our piloting program we used a real construction project of our industrial partner, a Finnish supermarket [4]. In piloting we simulated the emergency evacuation from a supermarket with a gamified model by utilising AI (Artificial Intelligence) based navigation of moving avatars in the model of the market. In the simulation we had avatars with individual behavioristic profiles and sets of rules. We made the avatars evacuate themselves using an AI engine and individual parameters along evacuation routes that were rapidly chosen after a fire alarm.

II. Methodologies and Software Used in the Research

The maximum space requirement in dimensions x,y and z can be derived from video clips of real human actions. In this method the real actions are videoed with at least two video cameras that are situated at right angles. The need of space from the actions in three dimensions is measured and a corresponding geometrical object is created. This 3D object in IFC file format is compatible with different CAD software. The goal is to ensure that the required actions can fit the space under design [1], [2], [3].

During the research program the application that is BIM compatible and is meant for Lifecycle Management of Building Data, and the application extension Value Add Data of Xamk R&D, were used for creating 3D objects. The application extension realizes the functionality of the FDM. The Colliders describing the space requirements for different actions to the gamified model were created via the execution of the Value Add Data software [5].

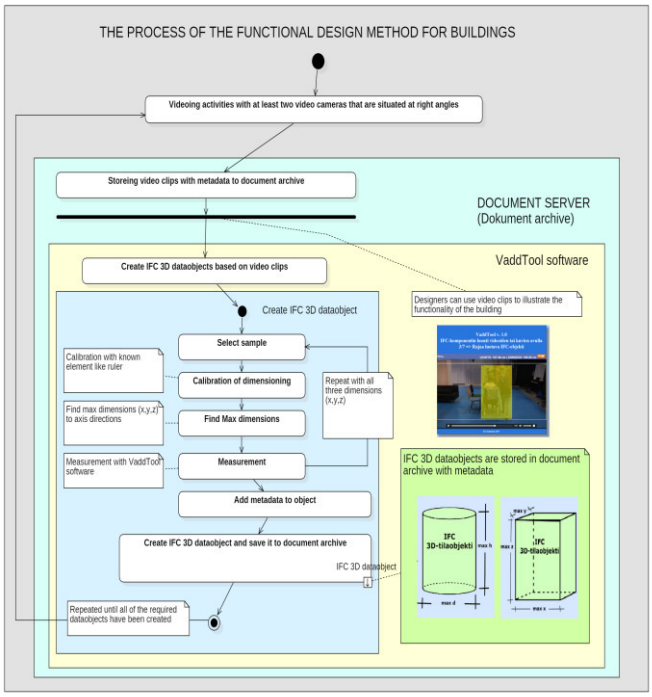


Fig 1. The process of the FDM for Buildings [5].

The Information Model used with the supermarket consists of several submodels that have been created by the designers of architecture, structures and different subsystems, with different CAD tools. As an example, the architecture model has been designed with the ArchiCAD software by Graphisoft SE. All submodels of the supermarket have been received for piloting in the IFC format [4], [6]. The submodels have been converted by the tools of the Open Source software family IfcConvert to the OBJ format [7]. The files that were in the OBJ format have been further optimized and shortened with the Open Source software MeshLab.

The metadata that are in the IFC files have been transferred by using the XML format. The conversion from IFC into XML has been performed with the IfcConvert. The gamification itself has been performed with the Unity game engine. When simulating the emergency evacuation we used the Navigation extension of the Unity engine. This extension brings possibilities to use AI based navigation schemes [8].

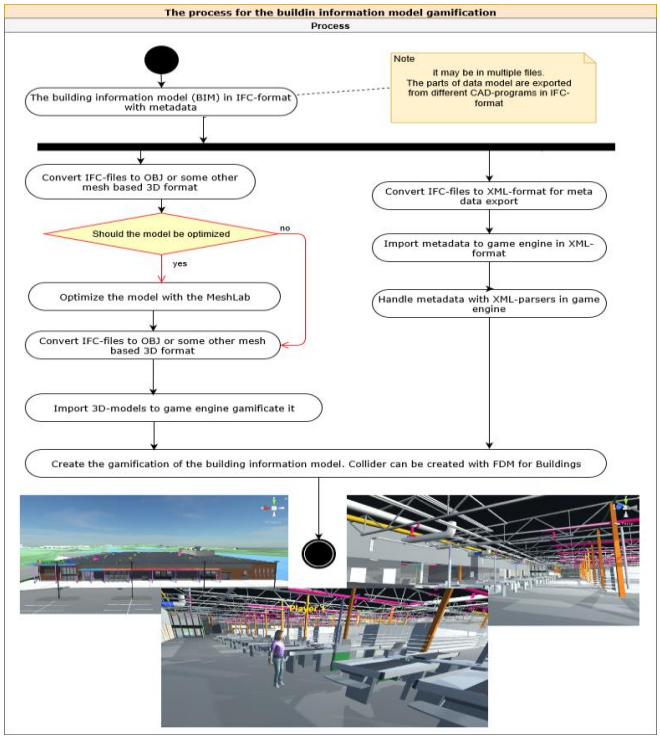


Fig 2. Our process for gamifying Information Models.

The schematic above describes the process developed by us to gamify Information Models of buildings. The basic principle of the process is that the gamification should be possible independent of the choice of the original CAD design tool. The only requirement is that the CAD tool supports the IFC format and can export both data and metadata into an IFC file according to the BIM recommendations [6].

Into a model gamified this way we can bring spatial objects generated by our FDM. These VAddD spatial objects can also support designs as such or if further converted into Game Colliders.

III. The Functional Design Method for Buildings Together with Gamification of Information Models Help to Design Safer Buildings

We created pilots to study different possibilities and applications for the FDM. We utilized an Information Model of a real supermarket. Our partner, the construction company U.Lipsanen Oy is currently building the supermarket. We received the Information Model as several IFC submodels, so we were able to test our procedure to gamify models represented according to the recommendation IFC (Fig 2) so that also the metadata can be input to the gamified model [4].

The main goal was to be able to use open software. For that reason we ended up to use the IfcConvert software from the IfcOpenShell library to convert IFC files into mesh models supported by game engines. Among the formats of mesh model formats we selected the OBJ format. We had a need to optimize and radically reduce the amount of polygons approximating the surfaces, especially related to pipes. To do this we chose the open MeshLab software that has support for the OBJ format. We also used the IfcConvert to create XML files from IFC and to bring the metadata of the models together with the 3D model into the game engine [6], [7].

With the aid of the gamified Information Model we piloted the usage of FDM in the emergency evacuation with an AI navigation control used in a Unity game environment [8]. We placed profiled or individually parametrized avatars randomly inside the supermarket and then commanded the avatars to proceed to the closest emergency exit with the aid of the AI engine. Around each avatar we placed game colliders that described the space needed for the avatar when mobile, according to the avatar profile. Such avatar can use different aids to overcome limited mobility when moving. An avatar using an aid like a rollator usually needs more space than an average person. Each of the avatars also possessed an individual moving speed.

With FDM we generated the 3D spatial objects representing spatial needs for actions. The description of the cases for avatars can be seen in the following table (Table 1).

TABLE I.
MAXIMUM SPACE REQUIREMENTS FOR THE DIMENSIONING OF THE COLLIDERS CREATED BY THE FDM

The dimensioned Action	Space Requirement (m)
A walking person	0.50 x 1.90 x 0.70 (x,y,z)
A running person	0.65 x 1.90 x 0.90 (x,y,z)
The assistant walks with the wheelchair user on behind the wheelchair	1.30 x 1.90 x 1.00 (x,y,z)
The wheelchair user without the assistant	1.00 x 1.50 x 1.00 (x,y,z)
The user with rollator	0.90 x 1.90 x 0.75 (x,y,z)
The user with crutches	0.90 x 1.90 x 1.00 (x,y,z)

From each of the cases we generated an avatar prototype that has colliders corresponding to the table. In this pilot the characteristics of the colliders were specified on a relatively common level. We felt that this accuracy is adequate from

the piloting of simulation method point of view. The avatar profiles could well be much more individual and more different types of avatars representing different customers of the supermarket could be created.

For accurate and realistic simulations corresponding to real situations it is necessary to increase both the number of variables and to increase the number of movement functions that have a partially random outcome. The number of simulation drives should be high when we want to apply statistical methods to the outcomes of the individual simulation drives. The results of the statistical studies would show how well the supermarket supports safety and show hints how to further increase safety by design. Different building designs could be simulated and compared. After iterating with different designs it is possible to reach the functionality requirements.

In our piloting we were mainly interested in the feasibility of the FDM in this kind of dimensioning and simulations. We used in our pilots some randomness in the collisions in between different objects to increase realism. The piloting was performed by using the Unity game engine and the C# programming language. We added a simulation extension to the gamified Information Model of the supermarket. The simulation extension creates a predefined number of avatars with different characteristic profiles. In the beginning of the simulation the avatars are located at their initial locations around the floor area, derived from random values of parameters. The characteristics for different types of avatars are in the table above (Table I).

In the simulation we used the Navigation tools of Unity, based on AI. With the tool we could create a navigation map situated on the floor level of the supermarket. More generally, the map could include any stationary objects of the gaming world. With the aid of the navigation map an avatar can now be commanded to progress to the needed target area so that the AI tool steers the movement of the avatar. The events during the navigation phase can be controlled by numerous different navigation parameters. Among the features that can be controlled are the maximum physical difficulty level of obstacles that can be over- or sidewise passed, and the intensity of effort in movement.

We realized the simulation in a way that the customer avatars with spatial needs according to their profiles are moving and directed by the AI towards emergency exits. The situation can be monitored by virtual cameras situated in the gamified Information Model. Each of the avatars are also carrying their own cameras that can be switched on when necessary.

The development of the situation can now be monitored from all angles. With the simulation application developed You can quickly control e.g. the number of customers, randomness of events and navigation parameters. It is also possible to change and define quickly the target-objects that are the destinations of the avatars. The avatar game objects that represent supermarket customers and are used in the simulation can in principle look like whatever the artist decides, but we increased the realism of the simulation by creating 3D avatars that look like typical shoppers. We also made them move according to kinetic data from an earlier human motion capturing session with cameras. Their movement looks therefor quite natural.

The realism of the simulation can be enhanced by adding different effects. Here we wanted to simulate a fire in a supermarket and we added a smoke effect to the gamified model. The progress of the fire follows a well known Fire Intensity Curve that is a generalized model about how the fires progress (Fire Intensity against elapsed time) [9]. The following picture (Fig 3) visualizes a generalized Fire Intensity curve that represents the intensity of the fire in majority of cases. In addition, the picture visualizes the phases of the evacuation and parameters used in the case of a supermarket.

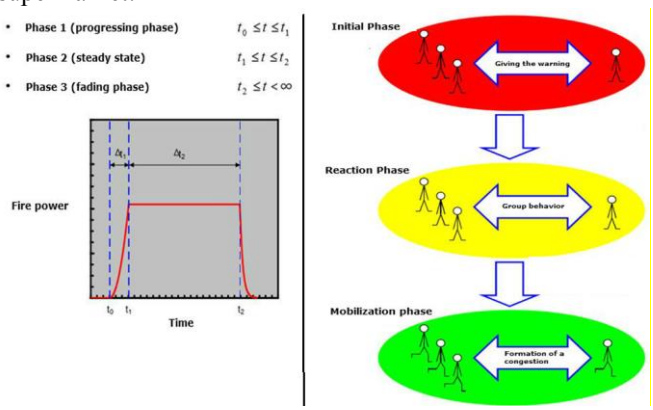


Fig 3. On the left side a generalized Fire Intensity curve that represents the intensity of the fire in majority of cases. On the right side the phases of an emergency evacuation [9].

The fire develops inside a building usually so that the environment becomes intolerable for humans. The cause can be different poisonous and irritating gases and the heat. Because of the fatal environment one should remove humans from the building as quickly as possible [9].

At the University of Lund the research aimed to find out how long it takes before the environment close to the fire becomes intolerable for humans. Also the VTT Technical Research Centre of Finland Ltd has performed research on the topic. In both research programs the critical factor was found to be the smoke occupying the whole volume of the building. The scattering of the research results is however relatively high [9], [10].

Based on the before mentioned research results we set the critical elapsed time after the start of the fire to 25 minutes with a standard deviation of 8 minutes. We put 500 customers into the supermarket corresponding to the daily rush hour. Our simulation can however adapt easily to a wide range of parameters for each case. The evacuation of humans from the buildings can be divided into three main phases. The phases are called the preliminary phase, the reaction phase and the mobilization phase. The total time elapsing in the evacuation follows the formula t = t(initial) +t(reaction) +

t(mobilization). The durations of these times are naturally very individual, so we need to include random variables related to the behaviour of people to our calculations. We realized the simulation application of an emergency evacuation by utilizing scientific studies mentioned above. We set the parameters of the dangerous situation (here a fire) so that the situation can be divided into the above mentioned phases. It was possible to freely change the durations of the phases. We also designed the customer (avatar) profiles so that there is randomness in the profiles and the process follows the phases of an evacuation [9].

The following picture (Fig 4) shows a view from the simulation application on the display during the run of a process performed by the Unity game engine.

Fig 4. A simulation application realized with the Unity game engine. A game collider corresponding to the spatial need of an avatar is visible on the picture.

The event that causes the need for evacuation can be parametrized and the seriousness and location can be varied. We can test a situation where a fire restricts access to a certain area and can even prevent using certain emergency exit.

After the simulation starts the game objects corresponding to profiled supermarket customers begin to move towards the selected emergency exits. The following picture (Fig 5) visualizes a situation where the fire simulation is active and the profiled customers are trying to move towards the emergency exits. Their spatial needs and profile settings affect their moving speed and route when they navigate past shelves and other structures, while interacting with the other customers. We can observe how the customers occasionally prevent each other using the most direct routes.

Fig 5. The simulation is active and smoke begins to enter the interior of the supermarket. The profiled customers (game objects) are rushing in chaos towards the defined emergency exits under the control of the AI engine (the reaction phase and the mobilization phase).

The structure of the simulation application is such that it saves the movements of the profiled customers to a log file. The applications also recorded for every customer the moment when he moved from a phase to the next. We also recorded to the log of the collisions with the other avatars and the moment of time when the customer was able to exit the building. Then the avatar reached the emergency exit that was the target location of the navigation.

The simulation was built so that the profiled customers possessed randomness according to the range found out in the studies [9], [10] and [11] during the initial and reaction phases. The simulation run specific log file generated gave a lot of valuable data about the functioning of the building in an emergency. With the aid of the logs it is e.g. possible to study the success level of different safety increasing measures. The log includes data for e.g. finding out when 90% of the customers have left the building or how much a new exit reduces the time of evacuation and also what the best place for a new exit is. In the exemplary simulation presented here all of the 500 customers could exit the building in 15 minutes. Over 90% of the customers had exited already in less than 10 minutes from the beginning of the fire. The last ones to exit were the individuals who have long reaction times and who are using aids for reduced mobility. When we made the distance between the shelf units narrower the evacuation times increased significantly especially with high numbers of people. Correspondingly, extra exits shortened the evacuation times, as expected. According to this simulation the supermarket was found to be safe enough in fire situations with the designed types of corridors and exits.

We also tested during the simulations a situation where the cameras are videoing from the point of view of the customers. We gave each customer an own camera enabling us to switch to the view from any of these cameras. It was especially interesting to note how an aged customer using mobility supporting aids and moving slowly suffers from collisions with other customers moving move swiftly. The following picture visualizes the emergency evacuation event from the point of view of e.g. an older customer moving slowly and using a rollator. This fact could also be found from the generated log files. The following picture (Fig 6) visualizes the emergency evacuation event from the point of view of e.g. an elderly.

Fig 6. Simulation showing the point of view of a slowly moving elderly using a rollator.

Finally we simulated a situation where the viewer is a TPC-type object (Third Person Controller) who tries to exit the building. Also in this case we have a set role and spatial needs that were dimensioned by using the FDM. We conclude that this could be an excellent way to practice doing the emergency evacuation. By using the multiplayer functions of Unity we could simulate situations where several players are practicing emergency evacuation or generally any co-

operation with other players. The following picture (Fig 7) visualizes a situation where each of the players of the simulation is moving a profiled avatar that has settable parameters. The players can do different roles like rescue or leading the evacuation operation. In the picture the player has entered the supermarket and encounters there customers who are rushing out. This way we could practice beforehand the actions during emergencies from the point of view of different actor roles.

Fig 7. A TPC player (Player1) has entered the supermarket after having received an alarm, performs rescue and assistance work and encounters customers who are rushing in panic.

The approach of this pilot appeared to be very interesting and relevant. The FDM brought clear added value to the piloting, because the different individual spatial needs could be taken into account in the simulation. According to our industrial partner the principle to combine the simulation of emergency evacuation with the usage of the FDM is a very good idea. We are planning to further develop these methods in our current research project.

The insertion of multiplayer functionality and co-operation of a community into this kind of simulation was especially rewarding. When the Information Model is gamified and we crowdsource the emergency evacuation simulation we could get valuable knowledge and ideas from a large group of people to enhance the accessibility and safety of a supermarket. We plan to develop these methods and tools further in the future.

IV. Results and Conclusions

The pilots presented strengthened our view that the Functional Design Method VAddD developed earlier is a useful and practical tool for designing buildings and that it is flexible to cover different purposes and application environments. The method suits also to different design challenges of infrastructure and built environment in addition to building design. It helps to perform the designs based on end users' activities in the buildings. The method is also independent on what CAD ecosystem was used to design the first 3D model. The pilot presented belong to our research program for the construction and building industry.

We conducted a small scale interview among our R&D partners related to the research results. We interviewed the industrial contact persons who are either managers or designers from the construction industry. One of the designers interviewed was specialized on gaming.

As a summary from the results of the interviews the usage of BIM, the gamification of Information Models and the Lifecycle Management of the Information Model were seen very welcome and needed development paths for the construction industry. The interviewed also said that the FDM is a very good idea for taking the human actions into account in all kinds of design challenges better than before. All new methods belonging to this category of methods were seen as welcome development. The partners thought that the simulation of the emergency evacuation scene was successful and stated that the FDM brings clear added value and new feasibility because You can now profile the users of the building and create individual characteristics and spatial needs to the avatars used. This way the simulation can be made very realistic and the results really bring new data about the building design and its functionality.

The partners of our R&D program think that the FDM for buildings and the pilots utilizing the methods are interesting, bring added value to the design processes and so are definitely worth further studies. Exactly to this direction should the design of buildings according to their opinion be developed in the future. When this pilot analysis cases are combined with further analysis of other actions in the building we have a large number of analysis results from a complete toolset of analysis fulfilling the needs in structural design, simulation, building automation design, training and even lifetime measurement and control during the whole lifecycle. Our vision is to adapt and expand the methods to develop virtual user interfaces or digital twins to buildings for the whole lifetime, beginning from the design phase.

References

[1] An action space defining object for computer aided design. Finnish Patent 125913 , granted on April 15, 2016. Fourteen claims. Current owner : Xamk. Inventor : Selin, Jukka-Pekka. Is the origin of the PCT patent application WO 2014/154942 A1.

[2] An action space defining object for Computer Aided Design. PCT patent application WO 2014/154942 A1. Applicant Mikkeli University of Applied Sciences Ltd (Jukka-Pekka Selin).

[3] Rossi, Markku J. and Bhargav, Dave, Digitalization and quality enhancement initiatives in sw assisted design processes in building and construction industries. 9th International Conference on Computer Engineering and Applications, Dubai, 2015. ISBN 978-1-61804-276-7.

[4] Construction Company U.Lipsanen Oy, A Building Information Model of Commercial Center in IFC format, Pieksämäki, Finland, 2018.

[5] Selin Jukka., Rossi Markku J., Simulation of universal design by a functional design method and by Gamification of Building Information Modeling. FedCSIS, 2016 Federated Conference, Gdansk, Poland, 2016. ISBN 178-8-3608-1090-3 and 978-1-5090-0046-3, IEEE.

[6] BuildingSMART International. The official website of the IFC-format (ISO 16739:2013). Accessed on April 18, 2018. http://www.buildingsmart-tech.org/specifications/ifc-releaces

[7] The official website of the IfcOpenShell-project. Accessed on April 18, 2018. http://ifcopenshell.org/

[8] Unity Inc. The official website of the Unity game engine. Accessed on April 18, 2018. https://www.unity3d.com/

[9] Paloposki Tuomas, Myllymäki Jukka, Weckman Henry. VTT Oy. Julkaisu. http://www.vtt.fi. (ISBN:951–38–6114–7). Accessed on May 11, 2018.

[10] Magnusson, S.-E., Frantzich, H. & Harada, K. Fire safety design based on calculations – Uncertainty analysis and safety verification. Lund, SE: Lund University, 1995. 120 s. (Report 3078.) ISSN 1102–8246k.

[11] Weckman, H. Rakennuksista poistumisen laskennallinen arviointi. VTT Oy, Julkaisu. ISBN 951–38–5133–8, ISSN 1235–060

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 913–917
DOI: 10.15439/2018F296
ISSN 2300-5963 ACSIS, Vol. 15

An Analysis of Game-Related Emotions Using EMOTIV EPOC

Jerzy Kosiński, Krzysztof Szklanny
Polish-Japanese Academy of Information Technology ul. Koszykowa 86, 02-008 Warsaw, Poland
Email: {s9347, kszklanny}@pjwstk.edu.pl}

Alicja Wieczorkowska
Polish-Japanese Academy of Information ul. Koszykowa 86, 02-008 Warsaw, Poland
Email: alicja@poljap.edu.pl

Marcin Wichrowski
Polish-Japanese Academy of Information Technology ul. Koszykowa 86, 02-008 Warsaw, Poland. Email: mati@pjwstk.edu.pl

***Abstract*— Computer games represent a very popular form of entertainment. Therefore, playing games became an object of interest for researchers. The research on the brain activity of players when playing a game is an experimental contribution to the neurophysiology of the central nervous system, and it also supports marketing research.**

Devices that register electromagnetic waves generated by the brain, e.g. EEG (Electroencephalography) can be used by psychologists studying the impact of games on users when the game. Our goal was to analyze emotion changes while playing video games, based on EEG signal registered with EMOTIV EPOC headset, and identify the strongest emotions accompanying the game. We also wanted to link emotions to particular elements of the game. Game developers, especially educational and therapeutic, can use the outcomes of this work in the practical implementation of the brain-computer interfaces in their products, in order to create better and more engaging games.

I. Introduction

ELECTROENCELOGRAPHY (EEG) is a noninvasive examination of brain function, using electrodes attached to the scalp. These electrodes record the electric field (electric potential) generated by the brain. Analysis of EEG records provides information on the activity of specific areas of the brain. EEG examinations are performed routinely as part of neurological diagnostics of the central nervous system. An important medical application of EEG is the diagnosis of coma and of brain death. EEG is also used as a tool in neurophysiological and psychoneurological tests, for example during the examination of the level of anesthesia, in the study of sleep disorders [1], and the analysis of the effects of drugs on the central nervous system. EEG is also used in neurophysiological research, including studies of cognitive and emotional processes. The emergence of instruments enabling low-cost registration of EEG signals contributes to the increase of interest in the use of such measurements in various fields of study.

The possibility of registering brain activity is also attractive in areas far from medicine. Unlike routine medical procedures, the research on brain activity does not focus on detecting anomalies that may indicate pathological changes, but on linking the recorded activity of the brain with received stimuli, cognitive activities, progress in training various skills [2], degree of relaxation, etc.

As a non-invasive method, and also more widely available recently, EEG is a natural candidate for use in the work on the brain-computer interfaces (BCI). In this work we will discuss some of the results of research in this area.

Game developers, especially educational and therapeutic, are particularly interested in the practical implementation of BCI in their products. EEG registering the reaction of players and the emotions accompanying the game is an interesting source of information for game designers, striving to create a game model that is satisfying for each user. EEG tests are also used by psychologists in studies on the impact of games on users.

II. Consumer EEG devices

Standard EEG measurements are performed using a stationary device, in an outpatient or laboratory setting. At present, various solutions are available on the market that allow recording EEG signals using portable devices [3], [4], [5]. Since these devices are inexpensive and easy to use, they can be applied in many areas, including using them as BCI interfaces in consumer applications. The usability of commercial EEG devices was tested in [6].

The EMOTIV EPOC+ wireless headset provides 14 channel EEG, plus 2 reference channels. Saline based wet sensors allow avoiding sticky gels. The headset is quite flexible, but its plastic structure limits the ability to adapt the device to untypical sizes or the shapes of the head. The signal is sent through Bluetooth or proprietary wireless communication. Battery life is up to 12 hours when using proprietary wireless, and up to 6 hours when using Bluetooth.

III. EMOTIV EPOC in applications and research

The EMOTIV hardware and software evoked great interest. The authors in [7] compared EPOC with the research EEG system (Neuroscan Synamps) for measuring

This work was partially supported by the Research Center of PJAIT, supported by the Ministry of Science and Higher Education in Poland

auditory event-related potentials. Their findings suggest that EPOC compare well with Synamps in such tests, and EPOC is easier to use because of its quick and clean set-up. On the other hand, EPOC may perform significantly worse than a medical device, as shown in [8]. The authors of [8] tested the ANT medical grade system and EPOC on P300 responses. P300 is an involuntary positive potential that is evoked about 300 ms after the user has perceived a relevant and rare stimulus. EPOC indicated lower SNR (signal-to-noise ratio), so the authors suggested choosing EPOC only for non-critical applications such as games, as it is not reliable enough for medical purposes such as prosthesis control.

IV. Experiment design

The aim of the experiment was to register EEG of players while playing computer games, and analyze emotions, depending on the type of game. During testing, EEG of 15 males aged 23-28 were registered. Persons who often play computer games were selected for our experiments.

Two computers were used in the experiments. The EMOTIV headset was connected to one of them and EEG data were recorded. Only the experimenter had access to this computer. The second computer was used for gaming, and this computer was used by the players. Three games with different gameplay characteristics have been selected for these experiments.

At the beginning of each experiment session, the experimenter puts the EPOC headset on the player's head. Next, the experiment adjusts the headset while checking the readings from EPOC, to assure the best possible quality of the signal. Afterwards, the experiment session is performed. The session consists of 7 parts. The player is informed before each part what he is expected to do, and how long it will take. These 7 parts (stages) of the experiment session are listed below:

1. Initial EEG measurements - the tested player closes his eyes, tries to relax, and not think about anything. The purpose of this stage of the experiment is to register EEG readings from the resting-state of the brain, for comparison with reading acquired in the next stages of the experiment.
2. Square game - the player plays the Square game, which is a team based game. His activity is monotonous, neither requiring mental effort nor psychically engaging. The registered EEG are to reflect moderate brain activity, typical of normal computer use.
3. Rocket League played with bots - the test player plays a 5 minute match with a computer-controlled player (bot).
4. Rocket League played online – the player plays a 5 minute match with other players (persons) over the Internet.
5. Watching high level game – the player watches a 5 minute match of the Rocket League, played by advanced players.
6. Super Bomberman game against bots – the player plays 3 matches of the Super Bomberman game against bots.
7. Super Bomberman game with another player – the test player plays a match of the Super Bomberman game with another player (person) against 2 bots.

As mentioned before, we selected 3 games of different characteristics. EEG readings were acquired using two methods. In the first method, each 0.5 the average power of theta, alpha, beta, and gamma waves was acquired (for beta waves, separately for low and high subbands). The wave ranges are as follows: theta 4–8 Hz, alpha 8–13 Hz, beta 13-30 (low 13–20 Hz, high 21–30 Hz), gamma 30–80 Hz.

The second method consisted in collecting emotion data using Emotiv software, for the following emotions: boredom, meditation, frustration, instantaneous excitement, and long-term excitement, scaled from 0 to 1. The Emotiv software is needed to acquire emotion data, facial expressions, and access mental commands.

After the preliminary analysis of the collected emotion data in the first round of tests, we decided to reject boredom and meditation, because of very low level of the obtained signals.

We also experienced technical problems with the EPOC headset we used, as 2 electrodes, A3 and AF3, did not work properly (and the readings from these electrodes had to be discarded, as they represented noise only). Therefore, we had to replace this headset with another one. This shows that this equipment is delicate and prone to failures. Accessories for EPOC are available, but it takes time to obtain these accessories.

V. Results

In our research, we describe emotions in 2-dimensional valence/arousal plane [9]. Based on the readings from the EPOC electrodes, we can calculated valence values, indicating positive or negative emotional states, and estimate arousal levels. Since beta to alpha ratio is a reasonable indicator of arousal state [10], we estimate arousal index E_A, as the indicator of arousal state, according to the following formula:

$$E_A = \frac{E_\beta}{E_\alpha}, \tag{1}$$

E_β - average beta power, E_α - average alpha power.

The valence index can be calculated for a particular electrode, or brain regions. Typically, this index is calculated for frontal (F) and antero-frontal (AF) electrodes, as beta and alpha waves are most pronounced and can be most easily measured in the frontal and the middle part of the brain [11]. Therefore, data from A and AF electrodes are the best choice when estimating valence [12].

It is suggested that greater left frontal activity is associated with positive affect and/or approach motivation, and that greater right frontal activity is associated with negative affect and/or withdrawal motivation, although in [13] the author found that the frontal asymmetry is responsive to

motivational direction and not affective valence. Still, since we had symmetrically placed electrodes at our disposal, we decided to use the following measure of valence E_V, calculated by comparing the alpha power and beta power between the right and left hemispheres [10]:

$$E_V = \frac{E_\alpha^r}{E_\beta^r} - \frac{E_\alpha^l}{E_\beta^l}, \quad (2)$$

r – right hemisphere,
l – left hemisphere.

E_V was calculated for the readings of symmetrically placed electrodes, or for averaged power calculated over selected brain parts. Positive E_V values indicate positive emotions, and negative values indicate negative emotions.

In the presented graphs we illustrate outcomes for the data collected in the 3rd stage of experiment, i.e. when playing Rocket League with bots:

- E_V for readings from electrodes F3, AF3 and also T7 for E_α^l and E_β^l, and readings from electrodes F4, AF4 and also T8 for E_α^r and E_β^r
- Instantaneous Excitement calculated using Emotiv software
- arousal index E_A for readings calculated for frontal parts of the brain

	Stimulus	Avg. Ex.	Ex. var.	Avg.Fr.	Fr. var.	TE	TF
1	Meditation	0.541	0.153	0.541	0.09	12%	3%
2	Squares	0.549	0.207	0.5	0.127	16%	4%
3	„Rocket League", bots	0.564	0.244	0.526	0.186	21%	10%
4	„Rocket League", persons	0.504	0.228	0.512	0.168	14%	8%
5	Watching „Rocket League"	0.518	0.266	0.483	0.153	20%	3%
6	„Super Bomberman", bots	0.621	0.2	0.568	0.144	22%	7%
7	„Super Bomberman", persons	0.579	0.224	0.537	0.153	23%	7%

Avg. Ex - Average Excitation; Ex. var. - Excitation variance; Avg. Fr. - Average Frustration Fr. var.- Frustration variance; TE - % time with excitation > 0.8; TF - % time with frustration > 0.8

We chose F3 and F4 as suggested in [10], and we additionally chose another symmetrically placed pair, AF3 and AF4, located close to F3 and F4. Additionally, electrodes T7 and T8 (temporal part of the brain) were chosen, as suggested in [14], where the authors indicate that information from the temporal part of the brain is significant in emotion classification.

The comparison of E_V values for various pairs of electrodes, i.e. F3 and F4, AF3 and AF4, T7 and T8, shows differences in brain activity for frontal and temporal parts of the brain.

Instantaneous Excitement calculated using Emotiv software shows large variability, and the results are directly related to game events.

The valence measure E_V and the arousal index E_A for frontal electrodes are not related directly to particular game events. The calculated values do not indicate stimulate from the game.

Therefore, we decided to use Emotiv software in further investigations. Instantaneous Excitement for the game and Long-Term Excitement for meditation show, that Excitement is related to game events.

The assessment of the emotion readings was done on the basis of average Instantaneous Excitement and Frustration, as well as Meditation and Boredom, calculated using Emotiv software. Average values and variance, as well as the percentage of time when Frustration or Excitation level was high (above 0.8) are shown in Tab. I. Initially, we also calculated these values for the Emotiv headset with damaged electrodes, but the results were inconsistent, so we decided to use another headset.

The values were averaged over all test participants, separately for each stage of the experiment.

As we can see, our experiment consisted of non-engaging (mentally, technically) as well as of highly engaging parts. Meditation, playing Square, or watching a game was rather non-engaging, whereas active playing games requiring high mental and motoric involvement (Rocket League, Super Bomberman) was highly demanding.

Excitement results show that average Excitation and the time of high Excitement when playing with other persons are greater than when playing with bots. For less engaging activities, the test results are similar, but variance is lower.

Frustration remained approximately constant during engaging activities, but variance increases when playing Rocket League with other persons, compared to playing with bots. Frustration results for non-engaging activities are a bit surprising; this is in line with results obtained in other investigations.

The highest meditation level was obtained for mentally non-engaging activities, but the difference in meditation level is small.

The highest boredom level was obtained when watching Rocket League match. This agrees with the relations of the tested players, as they emphasized their boredom during this part of the experiment.

Emotion levels when playing Rocket League match and EV values calculated for various electrodes are shown in Fig. 1. Unfortunately, the results from Emotive software are not perfect neither, as some exciting parts of the match are not reflected with the Excitement graph. For instance, only slight emotion changes can be observed between 180th and 230th second of the match, and rapid change can be seen during the scored goal only, whereas during this part of the match the ball was very close to the goal, and the player shot several times. After the 320th second of the match the excitement

suddenly drops almost to zero, whereas it is not reflected in the game, as the player was constantly defending goal during this part of the match. Frustration graph shows similar changes, but frustrating actions like loosing goal do not evoke frustration in this graph. In the final part of the match both scoring and defending goal correspond to surprisingly high frustration in this graph. The meditation and boredom levels also do not correspond to the gameplay.

The comparison of meditation and boredom levels registered in various experiments shows that these values are averaged over long runs. Therefore, such results are not very useful for estimating emotions while playing computer games.

VI. DISCUSSION

Our experiments confirm that playing a game causes brain activity changes, and these changes can be associated with strong emotions evoked by some parts of the gameplay. Similar observations were made in [16], with the focus on the changes of brainwaves power during high intensity and low intensity game events.

The analysis of the acquired data shows that the equipment and software are not sufficient to identify game-evoked emotions.

The literature we analyzed suggests that satisfactory results might be achieved only after complicated analysis and processing of the EEG data [17][18][19]. The use of neural networks is suggested. Deep neural networks are very often used recently, but since neural nets require large amount of data, are data are not sufficient for such analysis.

VII. SUMMARY AND CONCLUSIONS

The Emotiv software quite successfully detects excitement and frustration. Still, it is difficult to assure if the readings were accurate enough. The analysis of video recordings of the experiments is necessary to assess which parts of the gameplay correspond to the observed changes of emotions, and how to interpret these changes.

The algorithms implemented in the Emotiv software are not publicly available. Therefore, we cannot verify directly how various emotions are processed by these algorithms and what readings are reported for particular emotions. The values returned by these algorithms as meditation level (Meditation) and boredom level (Boredom) do not show large variability. The returned levels of excitement and frustration present larger variability, related to the gameplay as expected, but for each of these readings we can find parts of the play when these readings do not reflected emotions associated with playing the game. Since we cannot access these algorithms, nor formulas for calculating the readings, we are not able to assess whether the obtained values are correct or random. Similar problem was described by Harrison in [20].

The analysis of literature on emotion identification shows problems which must be addressed in order to achieve efficient detection and identification of emotions while playing computer games, based on EEG data. One of the issues to solve is to objectively identify game-related emotions. Various persons subjected to similar stimuli may experience different emotions. In the case of computer games, the player preferences may influence the experienced emotions. In the literature in the psychology domain [9] difficulties in the objective classification of emotions were reported, as the subjects may attribute various meaning to the same labels of emotional states, so the reported emotions may be ambiguous.

The method of evoking emotions and the emotion time should also be taken into account, no matter what emotion

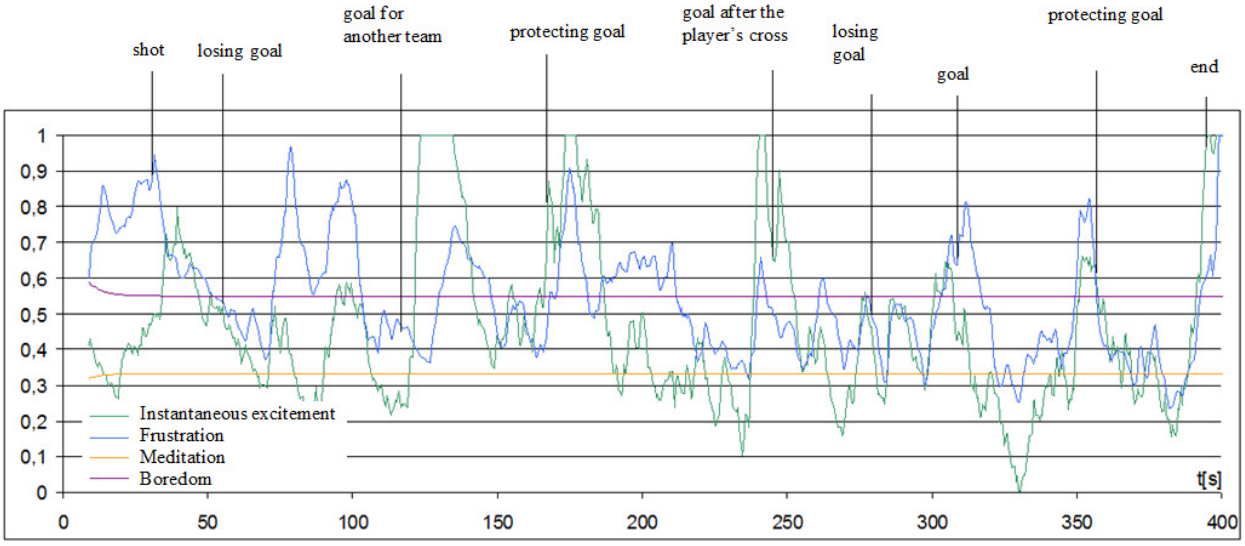


Fig.1. Emotions during the Rocket League match, calculated using Emotiv software

classification method was applied.

Depending on the type of stimulus and other experiment conditions, the registered EEG may represent emotional states of various intensity, duration, and co-occurrence of motor or cognitive activity. Even successful classification of emotional states, based on EEG data, can be less accurate for other stimuli or experiment conditions.

Computer games represent quite a specific type of stimuli, evoking emotions. Various types of gameplay may evoke different emotions in various players; the intensity of these emotions may also vary between players. In the case of typical computer games, intensive cognitive activity accompanies the playing, and at least medium motor activity. The player has to understand the gameplay, and actively control the game by means of the appropriate controller (at least until BCI interfaces are not efficiently and broadly implemented as game controllers). In typical games, there are numerous unforeseen events, to avoid experiencing boredom. Therefore, intensive emotions are expected, and they must quickly change in time.

As mentioned above, we cannot rely on the emotions reported by players when we want to analyze their emotions experienced while playing a computer game. This was reported by Salminen and Ravaja in [15], as well as McMahan et al. in [16]. Also, negative emotions experienced while playing, e.g. fear or frustration, cause experiencing positive emotion of satisfaction at the end of the game, which is specificity of computer games. The player is aware of the game convention and he or she is distanced to the experienced emotion, so the player subjectively assesses emotions as less intense. Without further investigations it is difficult to answer the question whether these differences can be seen in EEG.

Depending on the type and dynamics of the game, and also on the type of emotions to be analyzed, the analysis should be performed on shorter or longer time segments. The analysis of the power of brainwaves assumes averaging of the registered potentials, both over time and over frequency. In our experiment, we registered 2 readings of the Emotiv EPOC headset per second. Therefore, we can perform the analysis of psycho-neurological phenomena of medium duration. In the literature, short time EEG readings are also reported, i.e. of order of hundreds of readings per second, also in investigations on emotion detection. The Emotiv EPOC headset can also provide such data, when using appropriate software library [7], [8]. However, such analysis requires much more complicated experiment planning, as precise synchronization of the EEG readings with stimuli is necessary [19]. Additionally, in order to obtain high signal to noise ratio, numerous repetitions of stimuli are necessary. On the other hand, quick repetition of similar stimuli should be avoided. Therefore, such experiments are behind the topic of this work for now.

REFERENCES

[1] E. Weaver, M. Gradisar, H. Dohnt, N. Lovato, and P. Douglas, "The effect of presleep video-game playing on adolescent sleep," *Journal of Clinical Sleep Medicine*, vol. 6, 2010, pp. 184–189.

[2] S. Wolf, E. Brölz, D. Scholz, A. Ramos-Murguialday, P. Keune, M. Hautzinger, N. Birbaumer, and U. Strehl, „Winning the game: brain processes in expert, young elite and amateur table tennis players", *Frontiers in Behavioral Neuroscience,* vol. 8, 2014, pp. 11–12.

[3] T. B. Cedro, and A. Grzanka, "CeDeROM Brain Computer Interface", *Information Technologies in Biomedicine*. Springer LNCS 7339, 2012, pp. 219-231.

[4] W. D. Hairston, K. W. Whitaker, A. J. Ries, J. M. Vettel, J. C. Bradford, and S. E. Kerick, K. McDowell, "Usability of four commercially-oriented EEG systems," *Journal of Neural Engineering,* vol. 11, 2014, pp. 1-14.

[5] T. S. Grummett, R. E. Leibbrandt, T. W. Lewis, D. DeLosAngeles, D. M. W. Powers, J. O. Willoughby, K. J. Pope, and S. P. Fitzgibbon, "Measurement of neural signals from inexpensive, wireless and dry EEG systems," *Physiological Measurement*, vol. 36, 2015, pp. 1469-1484.

[6] W. D. Hairston, K. W. Whitaker, A. J Ries, J. M. Vettel, J. C. Bradford, S. E. Kerick, and K. McDowell, *J. Neural Eng.,* vol. 11 046018, 2014.

[7] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC EEG gaming system for measuring research quality auditory ERPs", *PeerJ* 1:e38, 2013

[8] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the Emotiv Epoc headset for P300-based applications", *BioMedical Engineering OnLine*, vol. 12, no. 56, 2013.

[9] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[10] R. Ramirez and Z. Vamvakousis, "Detecting emotion from EEG signals using the Emotive Epoc device," in *Brain Informatics*, LNCS vol. 7670 LNAI, 2012, pp. 175–184.

[11] T. Matlovič, "Emotion Detection using EPOC EEG device", in IIT SRC 2016, available at https://www.pewe.sk/wp-content/uploads/2016/01/20_iitsrc_matlovic.pdf

[12] Y. Liu and O. Sourina, "Real-time subject-dependent EEG-based emotion recognition algorithm," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8490, pp. 199–223.

[13] E. Harmon-Jones, "Clarifying the emotive functions of asymmetrical frontal cortical activity", *Psychophysiology*, vol. 40, pp. 838–848, 2003.

[14] W. L. Zheng and B. L. Lu, "Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks," *IEEE Trans. Auton. Ment. Dev*., vol. 7, no. 3, pp. 162–175, 2015

[15] M. Salminen and N. Ravaja, "Increased oscillatory theta activation evoked by violent digital game events," *Neurosci. Lett.*, vol. 435, no. 1, pp. 69–72, 2008.

[16] T. McMahan, I. Parberry , and T. D. Parsons, "Modality specific assessment of video game player's experience using the Emotiv," Entertainment Computing, vol. 7, pp. 1-6, 2015.

[17] N. A. Badcock, P. Mousikou, Y. Mahajan, P. de Lissa, J. Thie, and G. McArthur, "Validation of the Emotiv EPOC ® EEG gaming system for measuring research quality auditory ERPs," *PeerJ*, vol. 1, p. e38, 2013.

[18] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the Emotiv Epoc headset for P300-based applications," *Biomed. Eng. Online*, vol. 12, no. 1, 2013.

[19] A. Sivanathan, T. Lim, S. Louchart, and J. Ritchie, "Temporal multimodal data synchronisation for the analysis of a game driving task using EEG," *Entertain. Comput.*, vol. 5, no. 4, pp. 323–334, 2014.

[20] T. Harrison, The Emotiv mind: Investigating the accuracy of the Emotiv EPOC in identifying emotions and its use in an Intelligent Tutoring System. Department of Computer Science and Software Engineering, University of Canterbury 2013

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 919–923
DOI: 10.15439/2018F175
ISSN 2300-5963 ACSIS, Vol. 15

Exploring EMG gesture recognition—interactive armband for audio playback control

Mikołaj Woźniak*, Patryk Pomykalski*, Dawid Sielski, Krzysztof Grudzień, Natalia Paluch, Zbigniew Chaniecki
UbiCOMP Engineering Club
Institute of Applied Computer Science
Lodz University of Technology Lodz, 90-924 Poland
Email: mikolaj@pawelwozniak.eu, pspomykalski@gmail.com,
dawid.sielski@outlook.com, kgrudzi@iis.p.lodz.pl, nat.paluch@gmail.com, z.chaniecki@iis.p.lodz.pl
* Equally contributing authors

***Abstract*—This paper investigates the potential of using an electromyographic gesture recognition armband as an everyday companion for operating mobile devices in awareness-requiring contexts and suggests the fields, in which further developments are advisable. The Myo armband from Thalmic Labs is a fully functional motion controller, based on gesture recognition through EMG muscle sensing. The device has been applied for audio control, and the usability and relevance of the gestural interaction have been examined. Participants were asked to operate on a recording while cycling, and a reference group performed similar task in leisure context. The gathered answers suggest decent potential of gestural interaction manner for environments requiring high visual attention, eg. driving or cycling. However, the current state of the solution acts in too sensitive way, as processing numerous misinterpreted gestures highly decreases the system's usability. Moreover, gestures employed are perceived as too apparent and intrusive for social interactions.**

I. Introduction and Related Work

NOWADAYS, mobile devices are common everyday companion for various types of activities. However, in some of those contexts, the user's attention shall not be drawn towards the device operation, maintaining the control almost unnoticeable, both for the user and the environment. Therefore, various interfaces are being developed to meet those expectations.

Eyes-free Interface Technology

While mobile and wearable computing have been rapidly growing fields in the recent years, the approaches towards user interaction were dominated by vision-based techniques. However, restricted input and output capabilities of mobile devices, as well as limited screen space cause those approaches to be of a reduced usability [1]. Moreover, the visual perception of information is likely to be disadvantageous or even intrusive due to number of factors, such as: 1) *Competition for Visual Attention* - crucial in multiple mobile contexts eg. driving, running, cycling; 2) *Inconvenience* - as using the display requires the user to reach the device from a pocket or a bag; 3) *Technological Limitations* - such as reducing the battery life or the screen being illegible in bright sunlight [2].

Therefore, multiple interaction techniques employing senses other than sight have been developed in recent years. The most popular approaches make use of speech and gesture as the means of communication with the system [3]. Acoustic and haptic modalities have been adopted for assisting the interaction with mobile devices, both in terms of feedback and command, emerging the branch of eyes-free interfaces.

The most classical approach uses camera-based gesture recognition [4]. While real-time image processing and gesture recognition is no longer a challenging issue, the attitude still posses several constraints. The system intended for everyday use would need to be robust against environmental noise and obstructions of vision range, require minimal or first-use calibration and analyse comprehensible set of gestures and react with minimal delay. Moreover, a huge issue is connected to lightning conditions required by the system, which are out of control for rapidly changing environments. An attempt to bring such solution into practice was performed by Akyol et al. [5] for automobile use of acoustic messages. However, the necessity of assembling a complex camera-based setup makes this approach inappropriate for mobile scenarios.

An insightful endeavour towards applying gestures to audio control for on-the-go contexts was performed by Brewster et al. [1]. The study shown that metaphorical gestures are an intuitive and efficient way of interacting with an MP3 Player. However, the presented solution employed a device that the user was to swipe his finger at, therefore not solving the issue of inconvenience of hand-occupying device. A related solution was proposed by Zhao et al. [2], employing touch-input and auditory feedback for the purposes of menu browsing. This concept has been further explored by Kajastila and Lokki [6], providing further insights on menu operation using wii-like controller for eyes-free operation.

User Experience

The technological means of control are not the only challenge in designing appealing mobile interaction. It is the user experience and the emotions it brings that would make the novel interface become an everyday companion. Therefore, the social and behavioral aspects of operation are the essential issues to be addressed. Yi et al. [7] present an extensive analysis of key factors determining typical user motivations for using eyes-free interfaces. Apart form technological limitations mentioned in [2], these are the effort and social

reception of the performed actions which significantly affect the system perceiving. Hence, the users indicated that it is desired to obtain low perceived effort and entertaining manner of operation. Moreover, the users specified a strong need for the manner that would not interrupt social activities or cause anxiety. The command over the device shall be organised in a way that would not draw undesired attention of the people in the nearest environment, which is the challenge that speech-recognition interfaces fail to successfully address [8] [9]. Those concerns are especially crucial in on-the-go scenarios of mobile device operations. Further, wireless and hands-free devices are desired, employing gestures with little human-to-human discursive meaning and being difficult to misinterpret [10].

EMG Gesture Controllers

The above-mentioned requirements may be met through employing gesture-recognition system that would enable to perform subtle and discreet movements, with minimal attention required. The technology which is capable to offer that functionality is an electromyographic sensor. EMG can convey information about muscle electrical activity, which could be applied for designing intimate mobile interfaces. An extensive study on the capability of this approach in terms of social-acceptance and user perception has been performed by Constanza et al. [11]. However, the system used, while offering satisfactory and promising interaction pattern, maintained of a hardware that was highly inconvenient through the necessity to apply electrodes on the users' bodies.

The most recent commercial solution of an appealing EMG controller have been delivered by Thalmic Labs, the Myo Armband [12]. The device consists of mutually aware set of electromyographic electrodes. Moreover, the device is equipped with 3D gyroscope, 3D accelerometer and a magnetometer for extended movement analysis. The band communicates via Bluetooth connection and enables controlling systems using set of predefined gestures. The precision and capability of the solution have been examined during the evaluation of the band for musical interaction application [13].

In our study, we explore the potential of gesture-capturing armband for the purposes of everyday interaction, using the case of audio playback control. The research aimed to verify whether EMG-driven solution is truly capable to meet the requirements of the subtle and convenient interaction, while maintaining sufficient precision and resistance against misinterpretations and environmental noise. The armband has been tested in two different context - one while performing moderately attention-requiring activity and the other in leisure context.

II. System Evaluation

In order to consider possible application of the system for everyday purposes, several criteria must be taken into account. The examined solution of an armband shall perform as an efficient tool for playback control, offering the desired user experience. Therefore, we decided to investigate following questions and concerns:

- establishing whether the Myo armband can perform as efficient audio controller,
- assessment of the usability of the gesture-based operation manner,
- examining the perceived effort and attention needed for Myo operation,
- inspecting the risk of performing undesired action due to gesture misinterpretations,
- studying social perception and emotional comfort of the gestural interaction,

Experimental Setup

The armband was configured to control a simple audio player. For this purpose only EMG sensors was used. The experimental setup employed three gestures to be interpreted: 1) *double tap* - resulting in play/pause toggle, 2) *wave in* - for switching to previous file/rewinding the recording 5 seconds back and 3) *wave out* - for switching to next file/rewinding the audio 5 seconds forward. These gestures were chosen because the device is excellent at recognizing them. The gestures are presented in figure 1.

Fig. 1. Gestures used in system: left - double tap, middle - wave in, right - wave out.

Prior to the test, the Myo armband was calibrated for the user and quick introduction was given. Then, first task for evaluation was introduced. The player was turned on into music preset - so the song switching gestures were made available. Each user was asked to operate the player for a while, pausing and playing songs and switching between them.

Following, the main podcast-based task was introduced - the player was preset in a way that rewinding became assigned to wave in/out gestures. The user was briefly introduced to the topic of the 12-minute long podcast on airplane crew cooperation [14]. The users were asked to listen to the recording and establish the answers to four content-relevant questions. The task was not time-limited and the users were encouraged to rewind or pause the recording if the find it necessary.

1) Scenario A - cycling activity: The users were asked to perform the task while cycling around the park. In such setting, additional muscle activity is present due to the necessity to control and steer the bicycle and overcome obstacles. Further, the issue of competition for users' attention and ease of control is emphasized. Moreover, the context enables analysis of the system's sensitivity towards noise and whether non-intended gestures are recognized. The setup described is depicted in figure 2.

Fig. 2. Setup used in bicycle scenario.

2) Scenario B - leisure activity: The control group were asked to perform the same task in leisure setting in a cafe. This scenario highlights the aspect of social perception of interaction and enables comparison for the assessment of the undesired gesture recognition influence on the usability (Fig. 3).

Fig. 3. Setup used in cafe scenario.

After completing the task, users were asked to fill-in System Usability Scale [15] and NASA TLX [16] questionnaires. Further, a wider post-study interview was performed. Participants were asked to share their views on the following issues:

- whether the system is conducive for everyday use,
- whether the gestures are intuitive and convenient,
- whether the system performed as desired and what problems were determined,
- to describe their feelings and perceived effort,
- whether they liked the system operations,
- to mention other contexts of potential use,

and encouraged for some free discussion about the tested device and their experience.

III. Results

The experiment was performed on a group of $N = 11$ participants, of various age and gender. Each participant took part in a single scenario - 7 participants in scenario A and 4 participants in scenario B. All of the participants from scenario A performed the activity on a sunny day between 12 and 16 PM. The assessment was performed based on the questionnaires answers and the interview commentary provided by the users.

A. Scenario A - cycling activity

Overall, the cycling task was rated as less challenging than the leisure one, with the exception of perceived temporal demand. However, participants felt less successful with fulfilling the task requirements. Averaged results of NASA TLX [16] task analysis are presented in figure 4.

Quantitative analysis of system usability using System Usability Scale [15] brought an average score of 69,64 points (STDEV 13,18), with the range of gathered scores between 47,5 - 80 points. This result suggests moderate usability, which comprises with qualitative analysis of the interview answers.

Users performing the cycling task appreciated lack of reaching the device out of a pocket. The gestures applied were commented as intuitive and easy to learn, and have been perceived as non-intrusive. The most significant drawback of the system was the tendency for gesture misinterpretations - the users complained that numerous undesired operations were recorded by the armband while steering the bicycle, especially while riding on rough road surfaces. Users felt lost, as multiple actions were performed one after another with no clear feedback on their sequence. Users commented: *Once I got into bumpy road, it messed up totally. I lost track of the podcast as it paused and resumed randomly. Every time I wanted to take a sharp turn left, it rewinded the recording*. On the other hand, the users appreciated that the operation of the recording requires relatively low attention and they can keep their vision focused on the road - *It was cool not to have to slow down or stop to safely operate the recording*.

Concerning the experience of armband operation, the participants stated that they feel confident with operating the device, does not feel ashamed or embarrassed using the device. However, some users commented that the gestures might be misinterpreted as turn signifying in traffic contexts. Multiple users shared the view that the device might be a convenient companion to control audio during driving - where the chance for undesired gestures is reduced, while the advantage of maintaining the sight on the road is even more significant. Yet, the same users confessed that they are not willing to use the device while on foot, as they believe it would misinterpret common gestures and might become inconvenient during social interactions, especially while not covered by the clothing.

B. Scenario B - leisure activity

Surprisingly, the task was perceived as more effort-demanding and more frustrating during operation in the leisure environment, while the users rated themselves more successful. The detailed analysis of the task effort measured is depicted in figure 4.

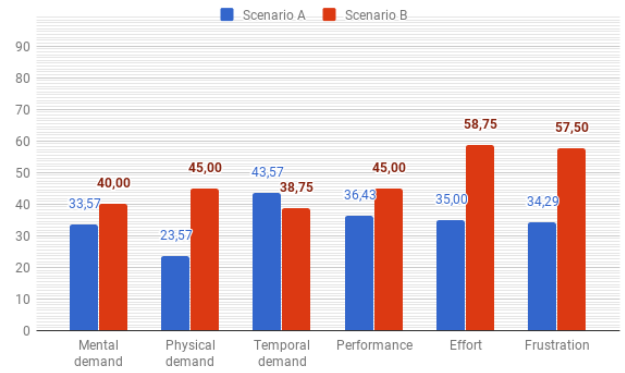


Fig. 4. NASA TLX Results - Scenarios A and B compared.

The system usability measured provided the average score of 71,25 points (STDEV 35,38), with the range of gathered scores being 20-98. The spread between assessments is significant, as issues connected to device calibration highly affected the operation manner for individual participant. It were the anatomic factors which seemed to be significant for proper system setup, as similar problems were encountered for different applications, such as [13]. Therefore, the score calculated on basis of median results seem to depict this assessment more accurately, providing the score of 82,5.

The most significant difference in regard to cycling scenario is the number of misinterpreted gestures reported - in the cafe environment undesired actions are less significant and limited to other movements performed by the user. The set of interaction gestures was commented as intuitive, while being easily mis-performed during common movements - eg. holding a mug, grabbing items. Moreover, some users felt uncomfortable when made to perform a particular gesture multiple times, which drawn attention of other people in the room. The users commented: *I feel quite weird sitting here and waving a hand to myself.*; *I might use it at home, while cooking or reading in my couch, but I don't feel alright with hitting the street with it*. Asked about other contexts of use, they suggested using the armband during jogging, cycling and driving.

The general impression from all participants consisted of the device being too sensitive and performing a lot of undesired operations. A group of 7 users reported that they do not feel comfortable enough with system operation to use it in public, as they would feel embarrassed. Moreover, when informed about the price of the device, majority of users stated that they would not spend that much money, even if the device were applied for contexts suggested by themselves. On the other hand, supporting phone and audio control while driving was the most frequent idea for possible application of the device.

IV. Discussion and Future Work

The performed study clearly highlights several drawbacks of the current state of the technology. While the EMG-driven armband offers satisfactory precision and technological capability [13], it does not offer the desired user experience. The metaphorical gestures chosen by the Myo designers act properly, offering decent level of intuitiveness and enabling quick learning of system principles. However, the social reception of system control draws too much of an undesired attention and may cause anxiety during interacting with others, which is a huge drawback when concerning the device as possible daily companion [7] [2]. Therefore, further improvements in recognition of more discrete gestures shall be applied to enable convenient control of the device in social contexts.

Furthermore, the task assessment shows that using the device is seen as more demanding when the eyes-free interaction manner is not a clear advantage. Occasional need for performing a particular gesture multiple times, combined with quite apparent and explicit movements is perceived more demanding than habitual on-screen operation [17]. This reception results in extended frustration while the device is misoperating, as a well-known alternative naturally arises. However, the effect is no longer present when engaging the sight is clearly unfavorable [18].

A drawback reported to be the most frustrating during armband operation was the excessive sensitivity, causing numerous undesired actions to be performed unconsciously. Misinterpretations turned out to be the most significant flaw noticed, highly affecting the comfort of use - while the users rated the concept as interesting and worth exploring, they were not willing to use the device in the current state due to difficulty of maintaining control over its reactions.

In terms of improving usability, a new concept for decreasing the number of redundant tensions being processed is necessary. Reduced sensitivity is clearly not a suitable solution, in presence of the need for more subtle gesture recognition. Hence, deeper investigation is advised to facilitate greater control over the armband. One of the possible approaches employs using an interpretation window triggered with a highly unique gesture. However, such method requires extensive investigation both in terms of efficiency and the effort and social - related criteria [19] [7].

Analysis of the gathered results enables to observe that the armband performs efficiently when used with decent focus on its operation. Therefore, it might be advantageous to apply the device as a controller in direct contexts, such as interacting with large displays [20] or for controlling rapid visualisation techniques for dynamic real-time process imaging [21] [22] [23] [24] [25] [26] [27], semi-automatic data analysis [28] [29] and crowdsourcing analysis of industrial images [30] [31]. High precision of the movement processing may be employed for mapping the movements of users' forearm onto other kinds of motile devices, such as drones and robots [32]. Further investigation is advocated in the field of ergonomics and potential consequences of long-term operation of the armband, as mid-air interaction pose several challenges in terms of arm fatigue, as signified in [33].

V. Conclusions

The concept of an armband is a successful endeavour towards providing an accessible EMG gesture-recognition controller, having several interesting and efficient applications. However, the examined device poses additional challenges to its users in terms of everyday use. Excessive actions performed due to misinterpreted muscle activity highly affect the system's versatility and decreases the comfort of use. Nevertheless, the armband presents enormous potential for developing eyes-free interactions, employing exceptionally capable technology for gesture recognition. Further efforts for designing more subtle and restrained user experience are necessary for providing efficient and convenient eyes-free interface, suited to support mobile contexts where vision-awareness shall be drawn towards other factors. The proposed solution is promising in terms of alleviating users' problems with operating screen-based devices in contexts where it is unfavorable, but also create an engaging and enjoyable experience of operation.

References

[1] S. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker, "Multimodal 'eyes-free' interaction techniques for wearable devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, (New York, NY, USA), pp. 473–480, ACM, 2003.

[2] S.Zhao, P.Dragicevic, M.Chignell, R.Balakrishnan, and P.Baudisch, "Earpod: Eyes-free menu selection using touch input and reactive audio feedback," in *Proc. of the CHI Conference on Human Factors in Computing Systems*, CHI '07, (New York, NY, USA), pp. 1395–1404, ACM, 2007.

[3] M. T. Vo and A. Waibel, "Multi-modal hci: Combination of gesture and speech recognition," in *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, CHI '93, (New York, NY, USA), pp. 69–70, ACM, 1993.

[4] S. Rümelin, C. Marouane, and A. Butz, "Free-hand pointing for identification and interaction with distant objects," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '13, (New York, NY, USA), pp. 40–47, ACM, 2013.

[5] S. Akyol, U. Canzler, K. Bengler, and W. H. T, "Gesture control for use in automobiles," in *In IAPR MVA Workshop*, pp. 349–352, 2000.

[6] R. A. Kajastila and T. Lokki, "A gesture-based and eyes-free control method for mobile devices," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, (New York, NY, USA), pp. 3559–3564, ACM, 2009.

[7] B. Yi, X. Cao, M. Fjeld, and S. Zhao, "Exploring user motivations for eyes-free interaction on mobile devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, (New York, NY, USA), pp. 2789–2792, ACM, 2012.

[8] A. Pirhonen, S. Brewster, and C. Holguin, "Gestural and audio metaphors as a means of control for mobile devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, (New York, NY, USA), pp. 291–298, ACM, 2002.

[9] S. Brewster, "Overcoming the lack of screen space on mobile computers," *Personal Ubiquitous Comput.*, vol. 6, pp. 188–205, Jan. 2002.

[10] S. S. P. M. C. S. A. Feldman, E.M. Tapia, "Reachmedia: On-the-move interaction with everyday objects," in *Proceedings of the 2005 9th IEEE Int. Symnposium on Wearable Computers (ISWC'05)*, 2005.

[11] E. Costanza, S. A. Inverso, and R. Allen, "Toward subtle intimate interfaces for mobile devices using an emg controller," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, (New York, NY, USA), pp. 481–489, ACM, 2005.

[12] ThalmicLabs, "Myo armband." www.myo.com.

[13] M. K.Nymoen, M.Haugen and A.Jensenius, "Mumyo - evaluating and exploring the myo armband for musical interaction," in *Proc. of the Int. Conference on New Interfaces for Musical Expression*, NIME 2015, (Baton Rouge, Louisiana, USA), pp. 215–218, 2015.

[14] *A320 Podcast, Episode TAP015*. http://a320podcast.co.uk.

[15] J. Brooke, "Sus: A quick and dirty usability scale," 1996.

[16] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload* (P. A. Hancock and N. Meshkati, eds.), vol. 52 of *Advances in Psychology*, pp. 139 – 183, North-Holland, 1988.

[17] T. Li and D. Tsekouras, "Reciprocity in effort to personalize: Examining perceived effort as a signal for quality," in *Proceedings of the 14th Annual International Conference on Electronic Commerce*, ICEC '12, (New York, NY, USA), pp. 1–8, ACM, 2012.

[18] N. Henze, A. Löcken, S. Boll, T. Hesselmann, and M. Pielot, "Free-hand gestures for music playback: Deriving gestures with a user-centred process," in *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, MUM '10, (New York, NY, USA), pp. 16:1–16:10, ACM, 2010.

[19] A. Jude, G. M. Poor, and D. Guinness, "Grasp, grab or pinch? identifying user preference for in-air gestural manipulation," in *Proceedings of the 2016 Symposium on Spatial User Interaction*, SUI '16, (New York, NY, USA), pp. 219–219, ACM, 2016.

[20] L. Lischke, S. Mayer, A. Preikshat, M. Schweizer, B. Vu, P. W. Wozniak, and N. Henze, "Understanding large display environments: Contextual inquiry in a control room," in *2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, (New York, NY, USA), pp. LBW134:1–LBW134:6, ACM, 2018.

[21] K. Grudzień, Z. Chaniecki, B. Matusiak, A. Romanowski, G. Rybak, and D. Sankowski, "Visualisation of granular material concentration changes, during silo discharging process, using ect large scale sensor," *Image Processing and Communications*, vol. 17, no. 4, 2012.

[22] K. Grudzien, "Visualization system for large-scale silo flow monitoring based on ect technique," *IEEE Sensors Journal*, vol. 17, pp. 8242–8250, Dec 2017.

[23] K.Grudzien, A. Andrzej, and R.A.Williams, "Application of a bayesian approach to the tomographic analysis of hopper flow," *Particle and Particle Systems Characterization*, vol. 22, no. 4, pp. 246–253, 2006.

[24] K. Grudzien, A. Romanowski, D. Sankowski, and R. Williams, "Gravitational granular flow dynamics study based on tomographic data processing," *Particulate Science and Technology*, vol. 26, no. 1, pp. 67–82, 2007.

[25] A. Romanowski, K. Grudzien, Z. Chaniecki, and P. Wozniak, "Contextual processing of ECT measurement information towards detection of process emergency states," in *Hybrid Intelligent Systems (HIS), 2013 13th International Conference on*, pp. 291–297, 2013.

[26] A. Wojciechowski and R. Staniucha, "Mouth features extraction for emotion classification," in *FedCSIS'16, ACSIS, vol. 8. IEEE*, p. 1685–1692, IEEE, 2016.

[27] A. Wojciechowski and K. Fornalczyk, "Exponentially smoothed interactive gaze tracking method," in *Computer Vision and Graphics*, (Cham), pp. 645–652, Springer International Publishing, 2014.

[28] A. Romanowski, "Big data-driven contextual processing methods for electrical capacitance tomography," *IEEE Transactions on Industrial Informatics*, vol. doi:10.1109/TII.2018.2855200, p. in press, 2018.

[29] M. Skuza and A. Romanowski, "Sentiment analysis of twitter data within big data distributed environment for stock prediction," in *2015 FedCSIS'15*, pp. 1349–1354, Sept 2015.

[30] C. Chen, P. W. Woźniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, and M. Fjeld, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 52:1–52:25, 2016.

[31] I. Jelliti, A. Romanowski, and K. Grudzien, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *FedCSIS'16, ACSIS, vol. 8. IEEE*, p. 1613–1619, 2016.

[32] P. A. Romanowski S.Mayer L.Lischke K.Grudzien T.Jaworski I.Perenc P.Kucharski M.Obaid T.Kosinski, "Towards supporting remote cheering during running races with drone technology," in *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI EA '17, (New York, NY, USA), pp. 2867–2874, ACM, 2017.

[33] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani, "Consumed endurance: A metric to quantify arm fatigue of mid-air interactions," in *Proc. of the Conference on Human Factors in Computing Systems*, CHI '14, (New York, NY, USA), pp. 1063–1072, ACM, 2014.

2nd International Conference on Lean and Agile Software Development

THE evolution of software development life cycles is driven by the perennial quest on how to organize projects for better productivity and better quality. The traditional software development projects, which followed well-defined plans and detailed documentations, were unable to meet the dynamism, unpredictability and changing conditions that characterize rapidly changing business environment. Agile methods overcame these limits by considering that requirements are not static but dynamic, while customers are unable to definitively state their needs up front. However, the advent of agile methods divided the software engineering community into opposing camps of traditionalists and agilists. After more than a decade of debate and experimental studies a majority consensus has emerged that each method has its strengths as well as limitations, and is appropriate for specific types of projects, while numerous organizations have evolved toward the best balance of agile and plan-driven methods that fits their situation.

In more recent years, the software industry has started to look at lean software development as a new approach that could complement agile methods. Lean development further expands agile software development by adopting practices from lean manufacturing. Lean emphasizes waste elimination by removing all nonvalue-adding activities.

TOPICS

The objective of LASD is to extend the state-of-the-art in lean and agile software development by providing a platform at which industry practitioners and academic researchers can meet and learn from each other. We are interested in high quality submissions from both industry and academia on all topics related to lean and agile software development. These include, but are not limited to:

- Combining lean and agile methods for software development
- Lean and agile requirements engineering
- Scaling agile methods
- Distributed agile software development
- Challenges of migrating to lean and agile methods
- Balancing agility and discipline
- Agile development for safety systems
- Lean and agility at the enterprise level
- Conflicts in agile teams
- Lean and agile project and product management
- Collaborative games in software processes
- Lean and agile coaching
- Managing knowledge for agility and collaboration
- Tools and techniques for lean and agile development
- Measurement and metrics for agile projects, agile processes, and agile teams
- Innovation and creativity in software engineering
- Variability across the software life cycle
- Industrial experiments, case studies, and experience reports related to all of the above topics
- Gamification
- Affective Software Engineering

EVENT CHAIRS

- **Przybyłek, Adam,** Gdansk University of Technology, Poland

PROGRAM COMMITTEE

- **Ahmad, Muhammad Ovais,** University of Oulu, Finland
- **Akman, Ibrahim,** Atilim University, Turkey
- **Alshayeb, Mohammad,** King Fahd University of Petroleum and Minerals, Saudi Arabia
- **Angelov, Samuil,** Fontys University of Applied Sciences, The Netherlands
- **Anslow, Craig,** Victoria University of Wellington, New Zealand
- **Bach-Dąbrowska, Irena,** WSB Gdańsk, Poland
- **Bagnato, Alessandra,** SOFTEAM R&D Department, France
- **Belle, Alvine Boaye,** École de Technologie Supérieure, Canada
- **Bhadauria, Vikram,** Texas A&M International University, United States
- **Binti Abdullah, Nik Nailah,** Monash University Malaysia, Malaysia
- **Biró, Miklós,** Software Competence Center Hagenberg and Johannes Kepler University Linz, Austria
- **Blech, Jan Olaf,** RMIT University, Australia
- **Borg, Markus,** SICS Swedish ICT AB, Sweden
- **Buglione, Luigi,** Engineering Ingegneria Informatica SpA, Italy
- **Carreira, Paulo,** Instituto Superior Técnico, Portugal
- **Chatzigeorgiou, Alexandros,** University of Macedonia, Greece
- **Cruzes, Daniela,** SINTEF ICT, Norway
- **Daszczuk, Wiktor Bohdan,** Warsaw University of Technology, Poland
- **Dejanović, Igor,** Faculty of Technical Sciences, Novi Sad, Serbia

- **Derezinska, Anna,** Warsaw University of Technology, Institute of Computer Science, Poland
- **Diebold, Philipp,** Fraunhofer IESE, Germany
- **Dutta, Arpita,** NIT ROURKELA, India
- **Escalona, Maria Jose,** Universidad de Sevilla, Spain
- **Essebaa, Imane,** Hassan II University of Casablanca, Morocco
- **Figueira Filho, Fernando Marques,** Universidade Federal do Rio Grande do Norte, Brazil
- **Goczyła, Krzysztof,** Gdańsk University of Technology, Poland
- **Godboley, Sangharatna,** NIT ROURKELA, India
- **Gonzalez Huerta, Javier,** Blekinge Institute of Technology, Sweden
- **Górski, Janusz,** Gdańsk University of Technology, Poland
- **Gregory, Peggy,** University of Central Lancashire, United Kingdom
- **Hohenstein, Uwe,** Siemens AG, Germany
- **Janes, Andrea,** Free University of Bolzano, Italy
- **Järvinen, Janne,** F-Secure Corporation, Finland
- **Jarzębowicz, Aleksander,** Gdansk University of Technology, Poland
- **Jovanović, Miloš,** University of Novi Sad, Serbia
- **Kakarontzas, George,** Aristotle University of Thessaloniki, Greece
- **Kaloyanova, Kalinka,** Sofia University, Bulgaria
- **Kapitsaki, Georgia,** University of Cyprus, Cyprus
- **Karpus, Aleksandra,** Gdańsk University of Technology, Poland
- **Kassab, Mohamad,** Innopolis University, Russia
- **Katić, Marija,** School of Computing, Engineering and Physical Sciences, United Kingdom
- **Knodel, Jens,** Fraunhofer IESE, Germany
- **Kropp, Martin,** University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
- **Kuciapski, Michał,** University of Gdansk, Poland
- **Landowska, Agnieszka,** Gdansk University of Technology, Poland
- **Lehtinen, Timo O. A.,** Aalto University, Finland
- **Lenarduzzi, Valentina,** Free University of Bolzano, Italy
- **Liebel, Grischa,** University of Gothenburg, Sweden
- **Luković, Ivan,** University of Novi Sad, Serbia
- **Lunesu, Ilaria,** Università degli Studi di Cagliari, Italy
- **Mahnič, Viljan,** University of Ljubljana, Slovenia
- **Mangalaraj, George,** Western Illinois University, United States
- **Marcinkowski, Bartosz,** Department of Business Informatics, University of Gdansk, Poland
- **Mazzara, Manuel,** Innopolis University, Russia
- **Mesquida Calafat, Antoni-Lluís,** University of the Balearic Islands, Spain
- **Miler, Jakub,** Faculty of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
- **Misra, Sanjay,** Covenant University, Nigeria
- **Mohapatra, Durga Prasad,** NIT ROURKELA, India
- **Morales Trujillo, Miguel Ehecatl,** University of Canterbury, New Zealand
- **Mordinyi, Richard,** Vienna University of Technology, Austria
- **Muszyńska, Karolina,** University of Szczecin, Poland
- **Nguyen-Duc, Anh,** University College of Southeast Norway, Norway
- **Norta, Alex,** Tallinn University of Technology, Estonia
- **Noyer, Arne,** University of Osnabrueck and Willert Software Tools GmbH, Germany
- **Oktaba, Hanna,** National Autonomous University of Mexico, Mexico
- **Ortu, Marco,** University of Cagliari, Italy
- **Oyetoyan, Tosin Daniel,** SINTEF Digital, Norway
- **Özkan, Necmettin,** Kuveyt Turk Participation Bank, Turkey
- **Panda, Subhrakanta,** Birla Institute of Technology and Science, Pilani, India
- **Pereira, Rui Humberto R.,** Instituto Politecnico do Porto - ISCAP, Portugal
- **Poniszewska-Maranda, Aneta,** Institute of Information Technology, Lodz University of Technology, Poland
- **Przybyłek, Michał,** Polish-Japanese Academy of Information Technology, Poland
- **Ramsin, Raman,** Sharif University of Technology, Iran1ry
- **Ristić, Sonja,** University of Novi Sad, Faculty of Technical Sciences, Serbia
- **Rybola, Zdenek,** FIT CTU in Prague, Czech Republic
- **Salah, Dina,** Sadat Academy, Egypt
- **Salnitri, Mattia,** University of Trento, Italy
- **Schön, Eva-Maria,** University of Seville, Spain
- **Sedeno, Jorge,** University of Seville, Spain
- **Śmiałek, Michał,** Politechnika Warszawska, Poland
- **Soares, Michel,** Federal University of Sergipe, Brazil
- **Soria, Álvaro,** ISISTAN Research Institute, Argentina
- **Spichkova, Maria,** RMIT University, Australia
- **Stålhane, Tor,** Norwegian University of Science and Technology, Norway
- **Stettina, Christoph Johann,** Leiden University, The Netherlands
- **Swacha, Jakub,** University of Szczecin, Poland
- **Taibi, Davide,** Free University of Bolzano, Italy
- **Tarhan, Ayca,** Hacettepe University Computer Engineering Department, Turkey
- **Thomaschewski, Jörg,** University of Applied Sciences Emden/Leer, Germany
- **Torrecilla Salinas, Carlos,** University of Seville, Spain
- **Unterkalmsteiner, Michael,** Blekinge Institute of Technology, Sweden
- **Wardziński, Andrzej,** Gdańsk University of Technology, Poland
- **Weichbroth, Paweł,** WSB University of Gdansk, Poland
- **Werewka, Jan,** AGH University of Sci. and Technology, Poland

- **Winter, Dominique,** University of Applied Sciences Emden/Leer, Germany
- **Wróbel, Michał,** Gdańsk University of Technology, Poland
- **Yilmaz, Murat,** Çankaya University, Turkey
- **Zarour, Nacer Eddine,** University Constantine2, Algeria
- **Łukasiewicz, Katarzyna,** Gdańsk University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 929–938
DOI: 10.15439/2018F72
ISSN 2300-5963 ACSIS, Vol. 15

Lessons Learned on Communication Channels and Practices in Agile Software Development

Muhammad Ovais Ahmad[1], Valentina Lenarduzzi[2], Markku Oivo[1] and Davide Taibi[2]
[1] M3S Research Unit, University of Oulu, Oulu, Finland
{ovais.ahmad; markku.oivo}@oulu.fi
[2] Tampere University of Technology, Tampere, Finland
{valentina.lenarduzzi; davide.taibi}@tut.fi

***Abstract*—Communication plays an important role in Agile Software Development (ASD). In each ASD practice (e.g., stand-up or retrospective meetings), different communication practices and channels are adopted by different companies. Several works have analyzed the impact of communication channels and practices. However, there are no secondary studies summarizing their impact on ASD. This study presents a Systematic Mapping Study (SMS) that aggregates, summarizes, and discusses the results of 25 relevant primary studies concerning the impact of communication channels and practices in ASD. We followed the well-known systematic mapping methodology in software engineering and analyzed empirical studies published before the end of June 2018. The results of our study have yielded several strategies that can be adopted by practitioners. Communication practices are context dependent. In the case of a distributed team, blended usage of rich-media communication tools, such as shared mind-map tools, videoconferencing, and promoting the exchange of team members between teams, is beneficial. In conclusion, communication can be expensive if teams do not apply the right strategies. Future research direction is to understand how to maximize product quality while reducing communication cost and how to identify the most beneficial communication strategy for the different stages of ASD.**

***Keywords*—communication practices, communication process, agile software development, scrum, extreme programming**

I. Introduction

Constant communication and sharing information about a project's development among the whole team is essential. Agile methods preaching the empowered and self-organizing development teams where focus is on constant communication and information sharing [18]. One of the most important factors for project success is continuous and active communication with the customer and the team members [16], effective communication among developers, operations, support, customers, management, and business areas [19]. During the development process, communication plays an important role in terms of coordination among the different teams involved, in order to manage dependencies between the actors in the process [6] [7] [8]. Communication can be considered as a mediating factor that influences both coordination and control activities during the development process [9].

In software development, it is of crucial importance to have effective communication starting from the project beginning, from the definition of the Minimum Viable Product [5]. Poor team communication is often leads to failure for engineering projects. Such poor communication become more complex in distributed software development. When the teams are unable to find good communication strategies or channels, it affects the quality of the product. However, communication is not always beneficial and can even decrease productivity [17].

Communication in Agile teams can be formal or informal. Formal communication includes specification documents and review meetings [10], while informal communication takes place via conversations among the teams within a company [10] and is usually based on telephone or video calls, audio or video conferences, email, and face-to-face meetings [11]. As suggested by Henttonen and Kirsimarja [11], face-to-face meetings can increase trust among the team members and, consequently, the quality of the development process and the final product.

ASD requires constant communication and information sharing within the team and between the team and the customers [20], [21]. Communication in ASD can be further classified as internal and external. In internal communication, developers and project leaders are the main actors involved in the process, while the development team and the stakeholders are the ones mainly involved in external communication [11]. Regarding the type of communication used, we can distinguish between active and passive. Active communication is mainly based on physical and synchronous approaches, such as face-to-face meetings, while passive communication is based on asynchronous approaches, such as email [20].

Various studies highlighted the problem of identifying the most effective communication channel [16][18][19]. Moreover, besides selecting the appropriate channel, practitioners still face the issue of selecting the most appropriate communication processes and practices [19][22].

In recent years, research has focused on the communication aspects in the Agile development process [12],[13],[14],[15], investigating key success factors [17],[18],[19] and the most effective communication channels that lead to positive effects on the process [22]. To the best of our knowledge, no secondary studies comparing the advantages and disadvantages of communication channels in the context of ASD exist to date.

Therefore, the aim of this study is to provide a systematic mapping review on the benefits and issues of the different communication practices and channels adopted in ASD. We also identify best practices and lessons learned in order to help teams working with Agile development processes in order to communicate more effectively.

Other reviews investigated communication practices in ASD from different point of views. Rizvi et al. [27] investigated distributed agile software engineering for years 2007–2012. Their main goal was to investigate reasons for adopting agile methods to GSD as well as risks. Alzoubi et al. [28] investigated the communication challenges in distributed teams that adopt agile, classifying challenges in six categories, but their focus is different to ours because agile practices are not taken into account. Hossain et al. [29] conducted an SLR on Scrum and GSD, but in contrast to our SLR their focus is limited to Scrum. Vallon et al. [30] performed a SLR on the application of agile practices in GSD. Finally, Hoda et al.[31] performed a tertiary study providing an overview of the SLRs on ASD research topics.

This paper is structured as follows. Section 2 presents the systematic mapping review protocol. In Section 3 we report the obtained results. In Section 4, we describe the threats to validity, and in Section 5, we draw conclusions and present an outlook on future work.

II. RESEARCH METHODOLOGY

The section outlines the adopted systematic mapping study (SMS) process, which follows the established guidelines and procedures proposed by Petersen et al. [2]. The motivation to conduct a SMS is to focus on the "*classification and thematic analysis of literature on a software engineering topic*" [23] [24]. SMS guidelines involves the following tasks: (1) defining the research questions; (2) outlining the search strategy; (3) extracting and analyzing the data.

A. Goal and Research Questions (RQs)

The goal of this study is to investigate the role of communication in ASD, focusing on channels and practices used during the communication process. We aim at identifying best practices and lessons learned in order to help ASD teams to communicate more effectively. To achieve goal of this study we addressed the following Research Questions (RQs):

(RQ1) What is currently known about communication in ASD?
(RQ2) Which communication channels have been studied in ASD?
(RQ3) What are the best communication practices commonly adopted in ASD?

A. Search Strategy

The search strategy adopted in this SMS is depicted in Figure 1. We first identified the bibliographic sources, then we defined the inclusion and exclusion criteria and selected the search keywords. Based on these, we carried out the selection process and finally extracted the data from the selected papers.

*Bibliographic Sources Identification***.** We selected the list of relevant bibliographic sources suggested by Kitchenham [4]: ACM Digital Library, IEEE-Xplore, Scopus ScienceDirect, Citeseer Library, Inspec, Springer. The selected databases are pertinent to this study as they are adopted by most of the literature reviews.

Search Keywords Definition. We defined search keywords based on the PICO structure [4]. We extracted the keywords from Population and Intervention terms. As suggested by Kitchenham [4], the Outcome and Comparison terms cannot always be considered in software engineering if the research focuses on general investigation.

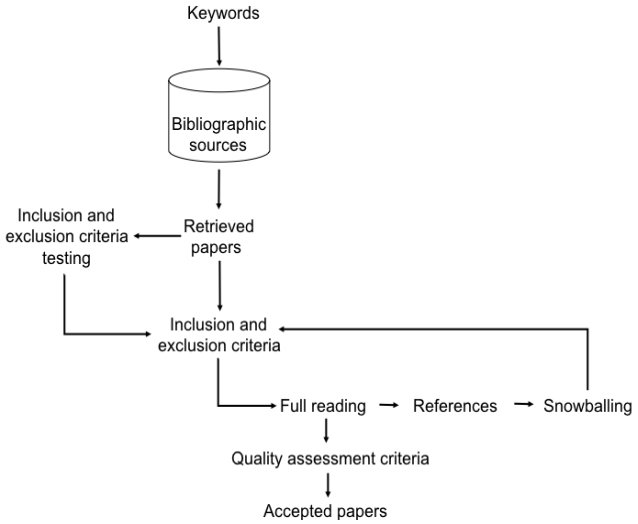


Fig. 1. Search Strategy process

Table 1. Search keywords

Population	**Intervention**
P: Agile development	**I:** Communication
P terms: Agile, Scrum, Extreme Programming	**I terms:** Communication

Based on the identified search keywords, we derived the following query:
(agile OR scrum OR "extreme programming") AND communication)

Inclusion Criteria Definition. Papers were eligible for inclusion in the SMS if they presented data related to communication in ASD. The inclusion criteria used were:

- Papers reporting communication practices in Agile;
- Papers reporting lessons learned, advantages, or disadvantages regarding the usage of communication channels;
- Papers reporting the influence of communication on the process;
- Study written in English and published before 07/2018;
- If published in more than one journal/conference, the most recent version of the study.

Exclusion criteria. Papers not fulfilling any of the inclusion criteria were left out, according to the following criteria:

- Papers not written in English;
- Duplicate articles;
- Not peer-reviewed scientific papers (books or book chapters, presentations, prefaces, gray literature, etc.);
- Simulation studies (e.g., mathematical modeling);
- Papers adopting the term "Agile" for purposes other than ASD (e.g., Agile manufacturing)
- Short papers, workshop papers, and work plans (papers not reporting results).

Search and Selection Process. The application of the search keywords in the selected bibliographic sources returned 2042 unique papers. Next, we applied the inclusion and exclusion criteria to the retrieved papers, regarding both title and abstract. As suggested by Kitchenham [4], we tested the applicability of the inclusion and exclusion criteria as follows: 90 papers randomly selected out of the 2059 were used as a sample; The three authors applied the inclusion and exclusion criteria to 60 papers each (each paper was reviewed by two authors); There was disagreement on 9 of the 90 selected papers (10%). For these 9 papers, the third author provided his/her opinion.

We applied the inclusion and exclusion criteria, for both title and abstract, to the remaining 1969 papers, and we included in this step 600 papers. Then, after the full reading process we selected only 82 papers.

In order to retrieve the most relevant papers, we integrated the procedure, taking into account also forward and backward systematic snowballing [3] on the 74 remaining papers. Regarding backward snowballing, we considered all the references in the papers retrieved, while for forward snowballing, we evaluated all the papers referencing the retrieved ones. We added another two papers and thus ended up with a new set of 84 papers as primary studies.

Assessing the Suitability of the Papers. In this last step, we checked whether the quality of the selected papers was sufficient to provide the required information needed to support the goal of our study. We considered all 76 papers obtained from the search and selection process. Each paper had to provide us with the following information:

- **Communication processes.** The paper had to report the communication process adopted during various development process phases and the research method adopted.
- **Communication channels.** The paper had to report the channels used for communication among the team, the stakeholders, and each partner involved in the process.
- **Best Practices or lessons learned** on the usage of communication channels in ASD.

The final dataset was reduced to 25 papers for the review, as reported in Table 2. In Appendix A, we present the list of the selected Primary Studies (PS).

Table 2. Search keywords

Step	# Papers
Retrieval from bibliographic sources	2059
Inclusion and exclusion criteria (title)	-1266
Inclusion and exclusion criteria (abstract)	-193
Full reading	-518
Snowballing process	+2
Assessing the papers' suitability	-59
Primary Studies (PS)	**25**

Data Extraction and analysis. Once the primary studies are selected, two of the authors analyzed and extracted the data, and the third one verified the correctness of the extraction. This practice helps to avoid researcher bias which is validity threat. The primary studies were analyzed based on study properties:

- **Context Data:** Study type of the paper; paper's goal(s); purpose of the communication; type of teams (distributed teams, collocated, distributed different times, different countries); communication stakeholders.
- **Process Data:** Communication definition; communication frequency; communication channels; communication tools; reported communication challenges for each communication channel.
- **Outcome**: Future direction from each paper; strength of evidence for each communication channel; Best Practices.

All primary studies have been analyzed separately by each author and then a combined peer-review has been conducted. In cases of disagreement, the third author was requested for his input. Finally, the author four ensure consistency in the analysis and consolidation of the results.

III. RESULTS

This section presents the results from the analysis of the 17 primary studies, which is based on the research goal previously mentioned in Section II. The results represent summary of results regarding the communication channels adopted in the primary papers, and then classify the lessons learned. The results identified the usage of four synchronous and two asynchronous communication channels as shown in Table 3.

The PS highlighted communication channels that are the main medium of contacting and exchanging information in ASD teams. In synchronous communication channels, face-to-face communication is the most frequently adopted channel, including both formal and informal communication. Phone calls are used mainly by project managers, while other roles commonly rely on video-conferencing, chat, or email [P1] [P2] [P16]. In case of asynchronous communication, ASD teams often use email, even in the case of co-located teams, while documentation is rarely adopted.

The PS reported that in Global Software Development (GSD), temporal distance challenges play an important role. Coordination between teams increases in complexity with increasing difference between time zones. Face-to-face communication is seldom employed in GSD, except in rare cases such as kickoff meetings. However, unexpectedly, even in GSD, pair programming (via videoconference) can be easily applied and turns out to be highly beneficial [P2].

For instance, Holmström et al. [P2], exemplified that developers with up to eight hours-time differences can work efficiently in pairs, supported by video-conferencing, but they need to shift their working hours so that they can have at least six overlapping hours per day. There is not silver bullet for communication channel, each channel has its own purpose. However, the PS recommended to experiment and blend various tools based on the team requirements and projects purpose to solve information sharing issues in ASD.

3.1 General Communication Channels Benefits in ASD

The PS reported that Agile teams use a variety of channels for communication to ensure open and multidirectional interaction.

Table 3. Classification and description of the communication channels adopted in the studies

Channel	Description	Primary studies (PS)
Synchronous		
Face to face	Can be either formal or informal. Includes communication during Agile ceremonies, pair programming, or any other face-to-face communication related to the project under development.	P1, P2, P3, P6, P8, P9, P10, P12, P13, P16, P17, P18, P19, P21, P22, P23, P24, P25
Video-conference	Used as medium for high-level discussions between non-co-located teams or between teams and customers.	P1, P7, P10, P11, P23, P24
Telephone/ Audio conference	Includes phone calls and any other audio conference tools adopted, such as Skype conferences (only audio) or phone calls.	P1, P2, P4, P5, P7, P10, P11, P14, P16, P19, P19, P21
Online chat	Instant Messaging applications such as Skype or Facebook chat. Mainly used as a means for exchanging technical information between co-located teams (e.g., passing information such as short configuration files) or for quick and informal communication between distributed teams.	P2, P7, P8, P5, P15, P16 P22
Asynchronous		
Documentation	Documentation is another form of communication that occurs between developers who need to report a set of written information. In the two PS, Wiki was the only tool reported and was used to communicate with other teams during project implementation or to keep track of technical choices.	P1, P10, P19, P20, P21, P24, P25
Email	Email is used for different purposes in ASD. Can be used both formally and informally.	P1, P2, P4, P5, P7, P8, P10, P16

Table 4 exhibits benefits, presented in the PS. It is important to identify tools for communication in early phases of a project, based on context and needs. Customers play a critical role in terms of identifying the communication tools to be used in a project. Such early identification is beneficial for Agile teams to achieve optimal performance and strengthen the relationship with the customer.

In all PS it is discussed that face-to-face communication yields a lot of positive outcomes compared to the use of other communication channels. For example, in case of requirement gathering is preferred way of communication [P5]. The face to face communication helps to reduce the capability of conveying ambiguous information [P1]. However, when face-to-face communication is not possible, online communication tools can be used efficiently. Video conferences, supported by rich media such as mind-mapping tools or desktop sharing [P15], improve the quality of the communication, while voice calls (Skype or telephone) are not as effective and should only be used for unofficial meetings. Chat is deemed to be more effective and useful for daily, informal information exchange or asking question from an expert about software functionality [P5].

Email was recognized as more formal way of communication and is more effective in case of getting approval on documents or requirements from the customer (where an email message constitutes a sort of contract) [P5]. On the other hand, the PS highlighted it as concern regarding effectiveness in the use of email for person-to-person communication or formal approval of documents [P5]. However, blended usage of different tools for different purposes can solve most information-sharing issues [P5]. Further, continuous Integration tools useful and helps to facilitate and to communicate the project status from development to final delivery.

Communication effectiveness decreases paired with the level of interaction provided by the communication channel.

3.2 Communication Benefits and Challenges in Agile Practices

ASD includes several practices prescribed by the different Agile approaches, such as Scrum, Extreme Programming (XP), and others. Some practices are shared by different approaches while others are not. In Table 5, we report the list of Agile practices along with their communication related benefits from PS.

Many software companies practicing pair programming. One reason is that pair programming helps to improve individual commitment and efficient way to implement code review [P2]. However, it is very challenging to use them on a daily basis as they are time and resource consuming [P6].

All PS highlighted that scrum meeting (e.g. daily standup, retrospectives) yields various benefits such as help to keep track of the project status, increase communication, enhance collaboration, reduce temporal distances and culture barriers. It is very important to note that the meeting which are conducted in front of project board are appreciated. One reason is that visibility and open discussions in the team/organization helped the spread the information and solve issues quickly. For example, in daily standup meeting, the visual board with different color cards is useful for keeping track of different types of stories [P12].

Table 4. Findings regarding communication benefits and issues.

Practices	Findings and PS
Open Communication	- Open communication must be encouraged and assured in order to get the benefits of Scrum [P11]. - Direct peer-to-peer communication between developers must be enabled to achieve successful results [P4]. - Frequent communication may be a symptom of a good and trusting relationship [P9]. - Team members communicate more with those they are aware of or with those they know can help [P8]. - Collocated teams over-communicate overheard problems [P12]. - A multicultural environment stimulates and increases productivity and creativity [P5].
Face-to-Face	- Preferred communication for collecting requirements [P5]. - Informal face-to-face communication should be encouraged to increase knowledge transfer [P14].
Videoconference	- Distributed teams should be equipped with video-conferencing instead of only using audio-conferencing or telephone [P7][P14][P18][P19][P21]. - Should be accessible to the entire team for regular meetings [P11].
Telephone, Chat	- Useful for unofficial meetings [GSD P11]
Email	- When face to face is not feasible, use email to increase the chance of response and encourage prompt, useful, and conclusive responses [GSD P4]. - Asynchronous communication due to temporal distance impacts coordination mechanisms [GSD P7]. - Effective to approve customer requirements when email message constitutes a sort of contract [P5].

Tool-supported Communication	- Use globally available project management tools to record and monitor project status on daily basis [P3]
	- Blended usage of different tools for different purposes can solve most information-sharing issues [P5].
	- Mind-mapping tools increase communication effectiveness and help mediate issues between distributed teams [P15].
	- Continuous Integration tools are
	  - Useful to communicate the current status of the project to testers, thus making system testing activities easier [P6];
	  - Facilitate testing. Help quality engineers get information on the status of the end product.

There is misconception that documentation is not important in ASD. The PS pointed out the importance of documentation, it helps both current and future team to work more efficiently and understand the logic easily. Code documentation is an important channel when there is needs to modify the code, helps traceability and test validation [P17]. In general, Agile teams rely on ad hoc communication and dynamic patterns of knowledge sharing [P8].

3.2.1 Team-related Communication Practices

Table 7 summarizes the findings of the PS on team-related communication practices. Leaders should be aware of a variety of culturally sensitive behavior and values. At the same time team members should respect the leader's views and should not be underestimated. A mutual trust is essential and it is duty of management to build such trust between developers and their first level manager [P5][P8].

It is also very important to exchange members in teams which are working on same or similar projects. The PS highlighted that a visiting engineer or outside expert is highly beneficial to support inexperienced teams during the first iteration [P10]. It is important for both co-located and distributed teams. Frequent exchange visits of team members are beneficial at the beginning of the project or in critical phases to get in touch with other members and learn how to work together [P11].

The primary studies stress that open communication should be encouraged among software development team members. This is helpful in various ways. For example, it improves team interaction and fosters good understanding between project team and management; in multicultural environments, it stimulates and increases productivity and creativity.

To increase interaction between teams, it is good to exchange team members in distributed project or interdependent teams. This helps them to interact more closely and fosters interpersonal relationships within teams. The use of emergent members helps to spread/share knowledge. Further, pair programming helps to increases mutual understanding and collaboration within and between teams [P2] as well as reduces social and cultural distances [P2]. However, it is difficult and problematic practice for daily use [P6].

Customer communication and close collaboration is crucial for development team and project success. During requirement elicitation customer absences bring challenges for challenge and it more difficult to perform remotely [P5]. The situation become more complex in distributed teams, where customer requirements are presented by other teams. It is also argued that upfront fixed requirements should be less ambiguous than deliberately vague agile requirements.

Table 5. Communication Benefits and Challenges in Agile practices.

Practice	Findings
Pair Programming	+ Increases time overlap and reduces temporal distance [P2] + Efficient way to implement code reviews [P6] + Increases mutual understanding and collaboration within and between teams [P2] + Reduces social and cultural distances [P2] - Difficult and problematic practice for daily use [P6]
Scrum/Sprint Planning Meetings	+ Set the scene for iterations involving negotiations with the customer [P12] + Reduce geographical distances[P2][P23] + Help to keep track of the project status [P6] + Increase awareness of the next iteration in the whole project team [P6][P23] + Provide close interaction among distributed project stakeholders [P7] + Help to minimize misunderstandings and misinterpretations regarding project standards [P2] [P19] [P21] [P23] + Increase mutual understanding and collaboration within and between teams [P2] + Help to reduce the confusion about what to developed from both the developers' and customer perspectives[P6]
Reflection Retrospective Reviews	+ Provide an efficient way to deploy and improve Agile practices [P6] + Good to use for assessing teamwork in completed sprint [P7] + Help to understand project standards among distributed project stakeholders [P7] + Increase project visibility and transparency [P7] + Help project managers with more efficient project supervision [P7]
Daily Standup Meetings	+ Increase awareness of project status among developers, product leaders, and customers [P6] + Help to quickly respond to changes in the project [P14] + Reduce coordination breakdown caused by temporal and geographical distance [P7] + Reduce cultural issues (e.g., perception of authority/hierarchy, frames of reference) [P7] + Convey strategy to the stakeholders [P7] + Increase knowledge sharing, thereby fostering a collaborative approach to problem solving [P12]
Story/Task Board	+Provides project status information to all stakeholders [P6] [P12]
Story Cards	+ Different color cards are useful for keeping track of different kinds of stories [P12]
Test-driven Development	+ Helps to maintain a shared standard view [P7] + Improves understanding of the functionalities required from the customer perspective [P7]
Refactoring	+ Improves communication, simplifying it and adding flexibility [P7]
Documentation	- Code documentation is an important communication channel when the customer needs to modify the code [P17] [P19] [P21] [P23] [P24] [P25]. - Documenting decisions is important in order to communicate them to future team members [P17]. - Test documentation helps to communicate information about traceability and test validation [P17].
Note: + (plus sign) indicates benefits; – (minus sign) indicates a challenge.	

Korkala et al [P1] exemplify that due to some reason project manager and customer group did not help developers and architects to analyze the requirements as well as deliberately hiding information. The companies should carefully plan their practices and recommended to follow people- vs. process-oriented control strategy [P1]. Further, some tools and practices can improve collaboration with the customer even in the case of geographical distance. In any case, the customer's role must be defined upfront and the customer should be enabled to make conclusive decisions regarding the project's functionality and scope [P3].

3.2.2 Organizational Responsibilities

The primary studies reported that in order to utilize the optimal capacity and skills of Agile teams, it is essential to provide proper method, process, and tool training (see Table 6).

Management should provide support along with a combination of internal and external coaching [P10] [P13]. Management should also provide access to everything that is necessary for the team's work, so that dependencies can be avoided easily. In the case of distributed development teams, the manager needs to understand the languages in which the various stakeholders communicate and needs to be sensitive to culture differences. The management need to make developers aware that they should be careful about other cultures. For instance, nobody felt the need to point out any cultural factor that would be disturbing or (even more surprising) stimulating [P5].

The introduction of new practices should be clearly communicated to the whole team paying attention to adapt

the process to the team needs, instead of forcing the team to adopt to the process as prescribed. The management needs to avoid practices which are process- vs. people-oriented control. "*Weak customer relationship and organizational politics that restrict information sharing may cause any communication medium to become inefficient*". The management should focus on creating an efficient customer relationship and environment that enables effective communication [P1]. Afterwards, management can focus on communication channels and tools.

Another interesting point highlighted by primary studies is an open office environment. In the case of co-located development work, it helps to decrease documentation and reduces the number of email communications. In an open office environment, the team members can easily keep up to date with the whole project view and notice obstacles in their colleagues' work [P6] [P17].

Table 6. Organizational responsibilities

Practices	Findings
Management role	-Training on ASD is needed if the team has never used it before. -Reading documentation is not enough [GSD - P10] [P13]. -Necessary resources (corporation intranet, documentation …) must be available to all team members [GSD [P5]. -The project manager should speak all the languages of the developers involved in the project [GSD P3]. -The project manager should collaborate on a daily basis with all the distributed teams [GSD P3]. -Project managers are key players in distributing information to others and being aware of others [GSD P8]. -The role of project managers is essential to the development of project requirements [GSD P8]. -Focus and practice people- vs. process-oriented control [P1].
Open office spaces	-Help to decrease the need for documentation [P6] -Enable everybody to have knowledge of the project status and common goals [P6] -Co-located office spaces close to the customer improve customer relationships since they foster communication [P17].

Table 7. Team-related communication practices and benefits.

Practice	Findings
Leadership & Trust	-Team leaders need to trust other team members [P5]. -The role of the team leader cannot be underestimated [P5].
Effective management	-Effective management helps to balance power [P5]. -Lack of necessary resources (corporation intranet, documentation) can lead to frustration and lower the motivation of the team [P5]. -Effective management makes projects successful [P5].
Emergent team members	-On-demand involvement of emergent members in a team helps to smooth out difficulties rather than getting stuck on a certain point that may lead to delay and failures [P8]. -Gathering information from outside members (i.e., support team, management team, executives…) is more necessary at the start of the project [P8].
Exchange of team members	-A visiting engineer or outside expert is highly beneficial to support inexperienced teams during the first iteration [P10]. -Frequent exchange visits of team members are beneficial at the beginning of the project or in critical phases to get in touch with other members and learn how to work together [P11]. -Exchange visits of team members and visiting schedules must be properly planned without the focus on saving traveling budget [P11]. -Agile teams need to consist of experts. Novices should be introduced to stakeholders gradually [P13].
Customer communication	-The customer must be readily accessible for communication with the development team [P3]. -Daily communication with customer reduces effort overrun [P9]. -Information hiding and lack of efficient customer relationship may lead to inefficient communication [P1] [P10][P19][P20][P21]. -Face-to-face communication between customer and development team is beneficial during project inception to discuss project goals [P14].

IV. DISCUSSION

Communication plays a crucial role in software, services, and systems development. The literature suggests synchronous and asynchronous communication channels in both co-located and distributed ASD. Many communication

challenges can be avoided with the right strategies. When face-to-face, synchronous communication is infeasible, the use of email increases the chance of getting a response and encourages prompt, useful, and conclusive responses. The blended approach of using various kinds of technology-mediated communication helps to avoid communication gaps among various development sites and/or teams. Software development teams need to promote, and be encouraged to maintain, healthy cooperation. Daily stand up meetings with the aid of various communication tools ensures a synchronous communication environment, as result build mutual understanding among distributed project stakeholders [P7].

The "teamness" and one-team attitude is a good strategy. It brings together team members across different locations and encourages cooperation between the team and the customer. Scrum planning meetings help increase "teamness" and reduce geographical distance. To get effective communication among team members, non-verbal communication is very important. In this regard, communication in co-located and distributed teams is not the same. Body language and hand gestures are difficult to observe in distributed teams. However, we can address such challenges to some extent by using technology-mediated communication. Minimizing the physical distances and using heavily technology-mediated communication can help software development teams build trust and work efficiently. The use of video-conferencing, the exchange of developers between different sites, and the deployment of emergent members, which helps in knowledge sharing, play a significant role in enabling effective communication in ASD teams. Furthermore, the PS discussed that Extreme Programming is useful for the more technical and coding aspects of GSD projects, whereas Scrum practices are good for GSD planning and tracking [P2].

Management needs to understand that spending some of the budget on exchanging team members among various sites fosters understanding of different cultures and facilitates communication. The challenges related to communication do not always stem from the use of the various communication media themselves, but may also be due to other reasons, such as fixed requirements, process-oriented control, lack of efficient customer relationship, proxy customer with no conclusive decision power. This results in inefficient communication and reduced efficiency of the communication media. These challenges can be avoided through various strategies, such as: defining the role of the customer up front and providing conclusive decision power regarding the project's functionality and scope; customer being readily accessible; customer having a vested interest in the project. Furthermore, the use of globally available project management tools is recommended in order to record and monitor the project status on a daily basis. To avoid low motivation of co-located and distributed software development teams, they should be granted access to the necessary resources (e.g., corporation intranet, email, product documentation, etc.). Additionally, it is recommended providing the necessary methodology training (e.g., Scrum, Kanban) and allow teams to experiment or pilot the new method in their work. Such training should be followed up with internal coaching to reap maximum benefits.

V. THREATS TO VALIDITY

In this section, we report the threats to validity, applying the structure suggested by Yin [1]. We identified and how we mitigated them based on SMS guidelines [2-5]. Moreover, the guideline proposed by Petersen et al. [2] suggests an objective checklist for assessing the quality of a study. The checklist considers information about activities conducted in the review, the need of the review, the search strategy adopted and its evaluation, extraction and classification process. We achieved an excellent score of 72% compared to the average (33% - 48%) of similar studies [2]. This value is the ratio of the number of actions taken in a review compared with the total number of actions required by the checklist.

Internal Validity. We defined the protocol based on the guideline proposed by [4] in a rigorous manner. As this protocol is the one most frequently used by researchers in the software engineering domain, we are sure that we have avoided any possible bias regarding the design of the methodology.

External Validity. Regarding the representation of the state of the art on communication in Agile development processes, we avoided this issue in our search and selection strategy by using a combination of automatic search in the bibliographic sources and backward-forward snowballing on the references of the selected studies. We did not consider papers that were not peer-reviewed in order to obtain high quality in our results.

Construct Validity is about bringing the right measures for the concept being investigated [2]. In order to reduce this threat, a data collection process was designed as suggested by Kitchenham and Charters [4]. We iteratively refined the inclusion and exclusion criteria by selecting a set of initial papers on which we tested their performance with regard to our goal. We also guaranteed inter-researcher agreement during the search and selection process.

Reliability. The results obtained from the selected papers allowed us to answer the defined research questions in the best possible way. This means that the data extraction process was well designed. Performing our SMS according to the guidelines [3] and providing raw data, will allow other researchers to easily replicate this study.

VI. CONCLUSION

In this work, we investigated communication channels and practices adopted in Agile software development using a Systematic Mapping Study. The 25 primary studies provide a detailed background regarding communication in ASD. It is identified that synchronous communication (i.e. face to face and phone calls) is dominantly used in ASD compare to asynchronous communication channels. It is evident that, even in GSD, pair programming with help of

videoconference easily applied and turns out to be highly beneficial. Various Agile practices such as Scrum/Sprint planning Meetings, reflection, retrospective reviews and daily standup meetings are beneficial in both ASD and GSF. However, the team managers and leaders should know that variety of culturally sensitive behavior and values. Further, along with the identified communication practices and channels, we found the following strategies for promoting effective team interaction in development teams:

- *Leadership & Trust: proactive role of management;*
- *Promoting the exchange of team members between sites;*
- *Active role of customer with conclusive power;*
- *Establishment of open office spaces in the case of co-located teams;*
- *Necessity of good information sharing tools selection at the beginning of project especially in case of GSD;*
- *Blended use of technology-mediated communication channels;*
- *Management should avoid in assigning work beyond capacity because teams easy burn and excused;*
- *Practice people- vs. process-oriented control*

The primary studies highlighted that practicing these strategies promotes team interaction between members from different sites/locations/units. These strategies are not the only ones for promoting communication; other strategies might exist but did not emerge from our analysis. Future works will include a set of industrial case studies and surveys to validate the results presented here.

REFERENCES

[1] Yin R.K. , "Case Study Research: Design and Methods", 4th edition, Sage, 2009.

[2] Petersen, K., Vakkalanka, S., Kuzniarz, L.,"Guidelines for conducting systematic mapping studies in software engineering: An update". Information and Software Technology. vol. 64, pp. 1-18. 2015.

[3] Wohlin, C., "Guidelines for snowballing in systematic literature studies and a replication in software engineering". 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14), Article 38. 2014.

[4] Kitchenham, B., Charters, S., "Guidelines for Performing Systematic Literature Reviews in Software Engineering", Version 2.3. 2007.

[5] D. Taibi and Lenarduzzi, V. , "MVP explained: A Systematic Mapping on the Definition of Minimum Viable Product", in Proceedings of the 42th Euromicro Conference Series on Software Engineering and Advanced Applications, 2016

[6] Espinosa J.A. and Carmel E., "The impact of time separation on coordination in global software teams: a conceptual foundation". Software Process: Improvement and Practice. vol. 8(4), pp. 249–266. 2003.

[7] Harbring, C. "The effect of communication in incentive systems—an experimental study". Manage. Decis. Econ., vol. 27. pp. 333–353. 2006.

[8] Malone T. W. and Crowston K., "The interdisciplinary study of coordination". ACM Comp. Surv. v. 26(1), p. 87-119. 1994.

[9] Carmel E. and Agarwal R., "Tactical Approaches for Alleviating Distance in Global Software Development". IEEE Softw. vol. 18(2), pp. 22-29. 2001.

[10] Herbsleb, J.D. and Mockus A., "An empirical study of speed and communication in globally distributed software development". IEEE trans. on Soft. Eng. v. 9(6), p. 1-14. 2003.

[11] Henttonen K. and Kirsimarja B., "Managing distance in a global virtual team: the evolution of trust through technology-mediated relational communication". Strat. Change. vol.14. pp. 107–119. 2005.

[12] Korkala M., Abrahamsson P., and Kyllonen P., A case study on the impact of customer communication on defects in agile software development. AGILE 2006, pp. 76-88, 2006.

[13] Melnik, G., and Maurer, F., Direct Verbal Communication as a Catalyst of Agile Knowledge Sharing. AGILE 2004, 2004.

[14] Sarker, S., and Sarker, S., Exploring Agility in Distributed Information Systems Development Teams: An Interpretive Study in an Offshoring Context. Information Systems Research, Vol.20(3), pp.440-461, 2009.

[15] Wang, X., Conboy, K., and Pikkarainen, M., Assimilation of agile practices in use. Information Systems Journal. Vol 22(6), pp. 435-455, 2012.

[16] Pikkarainen, M., Haikara, J., Salo, O., Abrahamsson, P., and Still, J., The impact of agile practices on communication in software development. Empirical Software Engineering. Vol. 13(3), pp. 303-337, 2008.

[17] Koskela, J., and Abrahamsson, P., On-Site Customer in an XP Project: Empirical Results from a Case Study. Torgeir Dingsøyr (Ed.) Software Process Improvement, Springer, Berlin Heidelberg, pp.1-11, 2004.

[18] Mishra, D., and Mishra, A., Effective communication, collaboration, and coordination in eXtreme Programming: Human-centric perspective in a small organization. Human Factors and Ergonomics in Manufacturing & Service Industries. Vol 19(5), pp.438-456, 2009.

[19] Mishra, D., Mishra, A., and Ostrovska, S., Impact of physical ambiance on communication, collaboration and coordination in agile software development: An empirical evaluation. Information and Software Technology. Vol 54(10), pp.1067-1078, 2012.

[20] Bhalerao, S., Puntambekar, D. and Ingle, M., Generalized agile software development life cycle. International Journal of Computer Science and Engineering. Vol I (3), 2009.

[21] Turner, R. and Boehm, B., Balancing Agility and Discipline: A Guide for the Perplexed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 2003

[22] Taibi, D., Lenarduzzi, V., Ahmad, M.O., and Liukkunen, K.,. "Comparing Communication Effort within the Scrum, Scrum with Kanban, XP, and Banana Development Processes". EASE. pp. 258-263. 2017.

[23] Ahmad, M.O, Denis, D, Kieran, C, and Markku, O. "Kanban in software engineering: A systematic mapping study." Journal of Systems and Software 137 (2018): 96-113.

[24] B.A. Kitchenham, D. Budgen, O.P. Brereton. "Using mapping studies as the basis for further research–a participant-observer case study" Inf. Softw. Technol., 53 (6) (2011), pp. 638-651.

[25] D. Taibi, Lenarduzzi, V. , Janes, A., Liukkunen, K. , and Ahmad, M. Ovais, "Comparing Requirements Decomposition Within the Scrum, Scrum with Kanban, XP, and Banana Development Processes" XP 2017

[26] B. Rizvi, E. Bagheri, D. Gasevic, D. “A systematic review of distributed Agile software engineering”. Journal of Software: Evolution and Process, 27(10), 723–762. 2015.

[27] E. Hossain, M.A. Babar, H.Y. Paik “Using scrum in global software development: a systematic literature review”. ICGSE 2009. pp. 175-184.

[28] Y.I. Alzoubi, A.Q. Gill, A. Al-Ani. “Empirical studies of geographically distributed agile development communication challenges: a systematic review” Inf. Management, 53 (1), pp. 22-37. 2016

[29] R. Vallon, B.J.d.S. Estácio, R. Prikladnicki, T. Grechenig. “Systematic literature review on agile practices in global software development” Information and Software Technology, Vol(96), pp. 161-180 2018

[30] R. Hoda, N. Salleh, J. Grundy, H. Mien Tee.. Systematic literature reviews in agile software development. Inf. Softw. Technol. 85, C (May 2017), 60-70. 2017.

[31] Ahmad, M. O., Markkula, J., & Oivo, M. (2013, September). Kanban in software development: A systematic literature review. In Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on. pp. 9-16.

APPENDIX A: SELECTED PRIMARY STUDIES (PS)

P1. Korkala, M., Abrahamsson, P., and Kyllonen, P. “A Case Study on the Impact of Customer Communication on Defects in Agile Software Development”. In Proceedings of the conference on AGILE 2006 (AGILE '06). 2006.

P2. Holmström, H., Fitzgerald, B., Ågerfalk, P.J., and Conchúir O.E. ”Agile Practices Reduce Distance in Global Software Development”. Information Systems Management. vol. 23(3), pp. 7-18. 2006.

P3. Layman, L., Williams, L., Damian, D., Bures, H. “Essential communication practices for Extreme Programming in a global software development team”. In Information and Software Technology, Volume 48, Issue 9, Pages 781-794. 2006.

P4. Korkala, M., and Abrahamson, P. “Communication in Distributed Agile Development: A Case Study“. EUROMICRO. 2007

P5. Cichocki, P., and Maccari, A. “Empirical Analysis of a Distributed Software Development Project”. In Balancing Agility and Formalism in Software Engineering. Vol. 5082, pp.169-181. 2008.

P6. Pikkarainen, M., Haikara, J., Salo, O., Abrahamsson, P., and Still, J. “The impact of agile practices on communication in software development”. Empirical Softw. Engg. Vol. 13,(3), pp. 303-337. 2008.

P7. Emam, H., Babar, M.A., and Verner, J. “How Can Agile Practices Minimize Global Software Development Co-ordination Risks?” Software Process Improvement: 16th European Conference, EuroSPI 2009, pp 81-92. 2009

P8. Inayat, I., Marczak, S., Salim, S.S., and Damian, D. “Patterns of Collaboration Driven by Requirements in Agile Software Development Teams”. 23rd International Working Conference, REFSQ 2017. pp.131-147. 2007.

P9. Molokken-Ostvold, K., and Furulund, K.M. “The Relationship between Customer Collaboration and Software Project Overruns”. AGILE 2007. pp. 72-83. 2007.

P10. Korkala, Mikko, Minna Pikkarainen, and Kieran Conboy. “Distributed Agile Development: A Case Study of Customer Communication Challenges”. Agile Processes in Software Engineering and Extreme Programming: 10th International Conference, XP, pp.161-167. 2009.

P11. Paasivaara, M., Durasiewicz, S., and Lassenius, C. “Using Scrum in Distributed Agile Development: A Multiple Case Study”. In Proceedings of the 2009 Fourth IEEE International Conference on Global Software Engineering. 2009.

P12. Sharp, H., and Robinson, H. “Three `C's of Agile Practice: Collaboration, Co-ordination and Communication”. Agile Software Development: Current Research and Future Directions. pp. 61-85. 2010.

P13. Gulliksen, V. Stray, V., Moe, N.B., and Aurum, A. “Investigating Daily Team Meetings in Agile Software Projects”. EUROMICRO 2012.

P14. Dorairaj, S., and Noble, J., and Malik, P. “Effective Communication in Distributed Agile Software Development Teams”. Agile Processes in Software Engineering and Extreme Programming XP 2011, pp. 102-116. 2011.

P15. Persson, J.S., Mathiassen, L., and Aaen, I. “Agile distributed software development: enacting control through media and context”. Info. Sys. Jour. Vol. 22(6). pp. 411-433. 2012.

P16. Inayat, I.N., Muhammad A., and Zubaria I. “Facilitating an Off-Site Customer in Product-Based Agile Software Development: An Industrial Case Study”. Emerging Trends and Applications in Information Communication Technologies, IMTIC 2012, pp. 210-221. 2012.

P17. Hummel, M., Rosenkranz, C., and Holten, R. “The Role of Social Agile Practices for Direct and Indirect Communication in Information Systems Development Teams”. Communications of the Association for Information Systems: Vol. 36, Article 15. 2015.

P18. Lenarduzzi V., Lunesu I., Matta M., Taibi D. “Functional Size Measures and Effort Estimation in Agile Development: A Replicated Study”. XP 2015. LNBIP, vol 212. 2015

P19. Taibi D., Lenarduzzi V., Pahl C. "Processes, Motivations, and Issues for Migrating to Microservices Architectures: An Empirical Investigation," in IEEE Cloud Computing, vol. 4, no. 5, pp. 22-32, September/October 2017.

P20. Munari S., Valle S., Vardanega T. (2018) Microservice-Based Agile Architectures: An Opportunity for Specialized Niche Technologies. In: Casimiro A., Ferreira P. (eds) Reliable Software Technologies – Ada-Europe 2018. Ada-Europe 2018.

P21. D. Taibi and Lenarduzzi, V. , “On the Definition of Microservice Bad Smells”, IEEE Software , v. 35, no. 3, 2018.

P22. Zykov, S. "Agile Patterns and Practices." pp. 107-134. Springer, Cham, 2018.

P23. D. Taibi, Diebold, P., and Lampasona, C. , “Moonlighting Scrum: An Agile Method for Distributed Teams with Part-Time Developers Working during Non-Overlapping Hours”, in ICSEA - International Conference on Software Engineering and Advances, Venice (ITALY), 2013.

P24. D. Taibi, Lenarduzzi, V., Diebold, P., and Lunesu, I. , “Operationalizing the Experience Factory for Effort Estimation in Agile Processes”, in 21th Evaluation and Assessment in Software Engineering (EASE), 2017.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 939–948
DOI: 10.15439/2018F358
ISSN 2300-5963 ACSIS, Vol. 15

Model Driven Architecture and Agile Methodologies: Reflexion and discussion of their combination

Imane ESSEBAA
Computer Science Laboratory of Mohammedia,
Faculty of Sciences and Technics Mohammedia,
Hassan II university of Casablanca, Mohammedia
Email: imane.essebaa@gmail.com

Salima CHANTIT
Computer Science Laboratory of Mohammedia,
Faculty of Sciences and Technics Mohammedia,
Hassan II university of Casablanca, Mohammedia
Email: salima.chantit@gmail.com

***Abstract*—Model Driven Architecture (MDA) and Agile Methods (AM) are two principal domains that are in the way of improvement and evolution in order to facilitate the development of IT projects. However, these areas evolve separately despite the great number of research that focuses on improving project' development techniques. Thus, our proposal aims to provide a method describing how can Agile Methodologies benefits from MDA, and how MDA can automate activities within AM. In this paper, we present a state of the art of existing works that combine a Model Driven Architecture approach and Agile Methodologies. Then we present our analysis of this combination to identify with which Agile methods, the MDA approach is more adequate. We also propose our vision about how to combine MDA with the selected Agile Methodologies and evaluate the strengths and weaknesses of each methodology during its combination with MDA and finally we present a case study of Rental Car Agency.**

I. Introduction

THE constant evolution of information system leads companies to search how to improve their productivity, their effectiveness and their profit margins in order to stay competitive. They are in a permanent quest for reliable tools that will cover the maximum of features.

In this context, two major domains have emerged in recent years and made an important place in companies' business: Model Driven Architecture and Agile Methodologies. On one hand, MDA has emerged as a new paradigm of software development that tends to use models as main artifacts in a higher level of abstraction, as well as the separation of the functional and technical specification. Indeed MDA has changed the view of software development: while classic methods concentrate on writing code, the MDA proposes a new method that focuses on analysis phases.

On the other hand, Agile Methods focus on the definition of best practices of information systems programming and their integration in the development process. It is an approach that defines a disciplined management of software development projects: Agility recommends an iterative and incremental method to develop software systems.

Agility puts the customer at the center of the development process of a project [1] and aims to develop software projects in the shortest possible time that satisfies all customer requirements and take into account requirements change, which is a fundamental principle of Agility.

Several works have been made on these two domains that help them to evolve and improve but separately. However, few of them focus on how to combine MDA and Agility.

The main idea in this paper is to present a state of the art of previous studies made in the context of combination of MDA and Agility.

We propose in this paper an analysis of Agile methodologies to define which are more appropriate to combine with MDA. We also give propositions on how to combine MDA and some Agile methodologies that we find appropriate to this combination.

This paper is organized as follows: after this introduction, the second section contains an overview of concepts in which this work is based, namely: MDA, Agile Methodology and RUP. Section 3 presents related works made in the context of combination of MDA and different agile methodologies. In section 4 we describes our analysis and proposed approach to combine MDA with V lifecycle in Scrum sprints. Section 5 is reserved to some perspectives of our future works, and finally we finish by a conclusion.

II. Overview of Concepts

A. Model Driven Architecture

The MDA (Model Driven Architecture) is an initiative of the OMG (Object Management Group) released in 2000 [2] The basic idea of the MDA approach is the separation of the functional system specifications and its implementation on a particular platform.

The MDA approach lies in the context of the Model Driven Engineering which involves the use of model and metamodels in the different phases of development lifecycle of an application[3], MDA defines three viewpoints:

- CIM (Computation Independent Model): the objective of this model is to represent the application in their environment independently of any computation information.
- PIM (Platform Independent Model): the role of the PIM is to give a static and dynamic vision of the application regardless of the technical conception of it.
- PSM (Platform Specific Model): This model depends on technical platforms, it represents a template of code that facilitates code generation.

B. Agile Methodologies

The ' Agile Manifesto ' published in in February 2001 [4] based on analysis of previous experiences that allow to propose good practices to developers, The agile principle introduced by the agile manifesto is related to time invested in analysis and design.[5]

Agility, a paradigm for a new vision of an organization, asserts itself as an alignment and coherence tool between internal forces and external challenges that give dynamism to an enterprise [6]. Agile methodology is a loom to project management, classically used in software development to manage IT projects development. It helps teams to respond to the changeability of building software through incremental, iterative work cadences, known as iterations.

III. STATE OF THE ART

The basic idea behind both Model Driven Engineering and Agile Methodologies is to create systems that can respond quickly to frequent changes, they propose different approaches resolve mentioned requirements; the agility focus on a methodological aspects that concerns an individual product, while Model Driven Engineering is more concerned by an architectural aspect defined by its specific variant MDA (Model Driven Architecture) that aims to separate system features from its implementation in technical platform.

Being aware of the importance of the agility and MDA in the development of software system, many works were focused on combining these areas, in this section we present some works previously made on this context.

In their paper [7] S.Hansson and al. collects different works made in practice on the context of the Model driven Agile Development and analyse the result of each approach, this approach consider that the basic idea behind the Model Agile Development approach is to benefit from practices proposed by agile methodologies, authors of this paper summarize the empirical literature made in the MAD context in order to extarct the lacks of this domain.

P.Cáceres and al. proposes in their paper [8] a case of study of an Agile Model Driven Development integrated in MIDAS framework which combines Model Driven Architecture approach and Agile practices based on eXtreme Progrming (XP). MIDAS is a model driven methodology for Web Information Systems (WIS) agile development, the architecture of MIDAS combines both MDA and n-tiers architecture, this combination is in order to propose a model of a WIS respecting the independence of the platform according to the MDA and in the same time take account of a middleware architecture of the Web services development platforms, while the process of the MIDAS methodology framework is based on Agile Modeling Driven Development practices that aims to write code progressively in agreement with the model. We mention that authors in this paper details the architecture of the MIDAS framework while it does not explain how the Agility is integrated in the process of MIDAS tool, moreover we note that the XP practice is dedicated specifically to the development phase during the software system development, which allows us to note that this approach does not implement all the aspect of the MDA approach.

In their paper [9] M.B.Nakicenovic presents an Agile Model Driven Development process developed in consideration of lean and agile practices, the paper aims to provide an approach that shows that MDD and agility can work together exploiting the benefits of each domain, the approach is applied on both forward and reverse engineering in order to respond to two issues; accelerating the re-engineering process of the MDD solution, how benefit from agility and lean while producing MDD solution within a short time frame. The paper describes an approach that combines MDD and agility based on lean, the implementation of the approach was made on the Market Server Capabilities (MSC) project proposed by SunGard company.

F.P.Basso and al. presents an approach that combines a Model Driven Architecture approach and Agility in the context of Rapid Application Prototyping (RAP) [10]; RAP allows the validation of software requirements before acceptance tests which in order to obtain a quick feedback from clients. This approach aims to take account of the agility principles in the context of MDA based on RAP methodology to generate front end and models based on MVC pattern, the implementation of this approach was applied on the generation of the Web Information System based on scrum methodology and MDE practices. The authors aims to ensure several benefits within this approach; better organization of source code, simplicity of changing the source code, simple and rapid design of models, and they still expecting to benefits from reuse of designed models. We note that this approach can't be generalized to all types of software systems, indeed it was dedicated to develop Web Information System, we also mention that in this approach authors do not detail how to integrate the MDA and scrum, i.e. they do not propose where to use each level of MDA in scrum methodology.

V.Kulkarni and al. discuss and argue in their paper [11] why agile methodology can't be used with Model Driven Engineering, then they propose a modification to make on agile methodologies in order to combine them with MDE. Indeed this paper describes a new Software Development process that combines Scrum and MDE, in this approach authors proposed the use of Meta-Sprints that run in parallel to Sprints in order to validate models, they suggest two to three months as timescales for meta-sprints where clients must provide feedback on models and prototyping , which is opposite to agility principles; indeed agility recommends that the feedback

of clients must be in period less than what was proposed in this approach.

H.Alfraihi in its paper [12] analyses the challenge of combining Agility and Model Driven Development, the paper describes an approach that aims to increase the adaptability of these domains by proposing a framework that facilitate Agility and MDD, this approach proposes recommendations, guidelines, and procedure to can use Agile MDD in practice. We note that even if this approach proposes some practices to implement the Agile MDD it does not take account of the architecture of the MDD, Model Driven Architecture, and how to benefit from the different abstraction levels to produce sustainable software systems.

In the paper, H.Wegener [13]presents a study made on the context of the combination of agility and Model Driven Development, then to propose issues that show how this combination affect organizations, process and architecture, this paper presents a comparison of different approaches proposed to use Agility and Model Driven Development.

In their paper [14] V.Mahe and al. presents their first reflections about the fusion of the MDA and Agility in order to have a combination with improved properties than the additions of the two approaches, they propose a canvas based on processes and agile practices in both modeling and meta-modeling level.

V.Nikulsins and al. propose an implementation of the MDA into RUP as presented in figure bellow:

- CIM covers the Business Model and requirements as well as planning for the development of the PIM metamodels.
- Elaboration is the main phase impacted by the MDA project, which is covered by PIM.
- PIM also cover part of the construction.
- In the construction phase the transformation of PIM model to PSM model starts.
- PSM covers transition phase

Burden et al. [15] have conducted a systematic literature review by proposing two research questions with the goal to investigate the empirical evidence of the state of art of integration of agile and MDD approaches and what is lacking in that area. The study shows that Agile MDD is still in its early stages and there is a need for detailed experience reports.

In their paper [16], H.Alfraihi and al. presented a systematic literature review to complement the results of [15] where fifteen primary studies were reviewed. The main characteristics of Agile MDD approaches besides their benefits and problems are highlighted. Both systemic literature review studies provide broader coverage, but they are less in-depth than interview study.

Eliasson and Burden [17] have conducted an exploratory study at Volvo Car Corporation (VCC). At VCC, individual teams adopt Agile practices with MDD while the organisation at large still use plan-driven process. The aim of this exploratory study is to investigate how Agile practices can be extended to the organisation level. In specific, it aims to answer the following question: Which are the challenges and possibilities for a more Agile software development process on a system level?. They interviewed 17 engineers to identify the challenges of the current process at VCC and how to improve it. The results of the interviews revealed two main challenges: first, the developers have to wait long before getting feedback which forced them to make premature assumptions leading to unwanted side-effects and faults; second, the use of MDD tools force developers to employ a waterfall process. The main finding of this study is that there is a need for a more Agile way of working to obtain earlier and faster feedback. They conclude that Agile MDD can be useful in automotive development. However, their study is context specific to VCC to examine its case and limited to automotive domain which is difficult to be generalised.

IV. OUR PROPOSITIONS

To ensure the good combination of MDA and agility we have to answer four questions:

- What are the reasons that motivate integrating Agile practices and MDD processes?
- How are Agile practices and MDD integrated?
- What are the benefits of integrating Agile and MDD on development process?
- What are the challenges of integrating Agile and MDD?

In their paper R.Matinnejad [18] defined that the integration of Agile in MDD process can be developed by:

- MDD-based: introducing Agile method to a current MDD process which is called.
- Agile-based: applying MDD process to an agile method.
- Assembly-based: integrating some fragments from Agile and other from MDD to develop the process

Although, both Agile and MDD processes have been introduced for more than a decade, the basic principles on how to integrate them together are not well-known. In this regard, the use of agile and MDD do not follow a well-defined process or systematic guidelines to guide through development. As a consequence, development teams introduce practices of Agile and MDD in an ad-hoc manner or based on their personal experiences.

In following, we discuss our reflection and analysis about different software methodologies and also agile ones and the possibility of their combination with MDD depending on each methodology criteria. To this end, we choose the most popular and used methodologies:

- eXtreme Programing
- 2TUP
- RUP
- V Life Cycle
- Scrum

A. *eXtreme Programing*

eXtreme Programming is a discipline of software development based on values of simplicity, communication, feedback, and courage. It works by bringing the whole team together in the presence of simple practices, with enough feedback to enable the team to see where they are and to tune the practices to their unique situation.

The eXtreme Programming is a software development discipline that organize people to develop higher quality software system and be more productive. XP defines four activities that are performed with software development process: Coding, testing, listening and designing. [19]

Analysing the definition and principles of the eXtreme Programming we deduce that this process is more appropriate to development phase in the process of software development.

According to XP and MDA approaches principles we deduce that it is not possible to combine both of these domains taking into account the fact that XP is dedicated to manage the development phase while MDA approach is an architecture based on different level of conception using models to finally generate automatically the source code.

B. 2TUP: Two Track Unified Process

2TUP is an instantiation of the UP (Unified Process) proposed by Valtech company, it takes into account constraints of rapid business changes of software system.

The fundamental principle of the 2TUP is to decompose and treat at the same time the business requirements, following two axis; functional one and technical one. At the end of the evolutions of the functional model and technical architecture, software development phase consists on combining the results of these two branches of the process as described in the figure bellow.

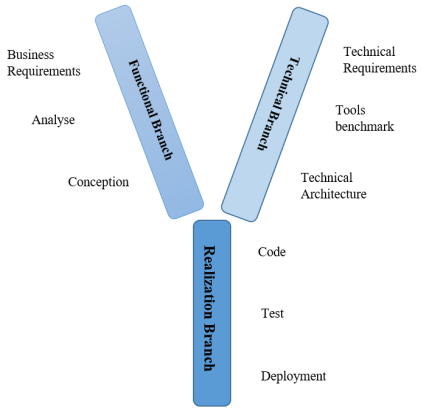


Fig. 1: 2TUP methodology

Analyzing 2TUP methodology we conclude that it can't be combined with MDA, indeed the MDA approach proposes to generate technical conception from functional one through model transformations which are the core of MDA approach. There are two types of model transformations; Model to Model (CIM to PIM and PIM to PSM) , and Model to Text (PIM to source code), while in 2TUP methodology the functional and technical conception work in parallel to finally combine their results in the development phase.

C. RUP: Rational Unified Process

The Rational Unified Process is a Software Engineering Process. It provides a disciplined approach to assigning tasks and responsibilities within a development organization. Its goal is to ensure the production of high-quality software that meets the needs of its end-users, within a predictable schedule and budget. [20][21]

It is closely related to the Unified Modeling Language (UML); this model proposes to manage the IT developments using a model of activities divided in 4 categories described in table 1:

TABLE I: Description of RUP phases

Phase	Description
Inception	• Establish a business case of the system • Define different actors and use case
Elaboration	• Analyze the problems of the domain • Establish the system architecture • Define the project plan
Construction	• Develop, test and integrate the components of the project
Transition	• Forward the project to users

The general architecture of the Rational Unified Process is characterized by two dimensions:

- The horizontal axis represents time and shows the progress of the lifecycle of the process; this first dimension reflects the dynamic aspects of the process that is expressed in terms of cycles, phases, iterations and milestones.
- The vertical axis represents major patterns of activities which include activities according to their nature; this second dimension reports the static aspect of the process that is expressed in terms of components, processes, activities, artifacts and workers.

Analyzing the RUP method and its structural phases, we conclude that it is possible to combine it with MDA approach, in the following part of this section we propose our view of implementation of MDA into RUP that consists on:

- CIM level covers the inception phase, this level is modeled by Use Case diagram to represent a static and functional aspect, and Activity diagram to represent a dynamic aspect.
- The elaboration phase is covered by the PIM level obtained from a vertical transformation of CIM model, this level is presented by Class Diagram representing a static view while a dynamic one is represented by a Sequence Diagram.
- PSM level covers a construction phase after enriching a PIM models and transform them using vertical transformation into Design Class Diagram and Interaction Diagram.
- Transition phase is covered by a code obtained with Model to Text transformation.

We summarize our approach in the following table:

TABLE II: Description of RUP phases

RUP/MDA	Inception	Elaboration	Construction	Transition
CIM	• Use Case diagram • SBVR			
PIM		• "Business" Class Diagram • "System" Sequence Diagram		
PSM			• "Design" Class Diagram • "Detailed" Sequence Diagram	
Code				• Source Code

D. Scrum

The Scrum method is an agile method, founded in 2002, It relies on the Division of project into iterations still named "sprints". A sprint can have a duration which varies generally between two weeks and a month.

The estimation of task in time and complexity is made before each sprint using several methods, in order to plan the deliveries and also estimate the cost of each task for the customer.

Features called user stories are the subject of the sprint that constitute a "sprint backlog" that can be a deliverable product at the end of the sprint.

There is difference between the sprint backlog "product backlog", corresponding to all of the features expected for the product on all of the sprints.

The Scrum method is also characterized by a "melee" daily, called "morning" or "stand up", in which employees (project managers, developers and functional managers) in turn indicate tasks that they have performed the day before, the difficulties and finally this whereupon they will continue their work the next day. This allows to evaluate the progress of the project, resources where they are most needed, but also to provide assistance to workers facing difficulties when these have already been encountered previously by other members of the team.

To combine MDA and Scrum we can use in each sprint of the project the MDA principles, i.e. in each sprint we apply our approach of generating source code from business requirements, in scrum we have a sprint backlog that describes system' features, the combination can be described as follow:

- As first step we model the requirements in sprint backlog by the UseCase Diagram and OMG standard SBVR to represent the CIM level.
- After modeling the CIM level, these models are transformed automatically to PIM level which is modeled by "Business" Class Diagram and "System" Sequence Diagram.
- Models in PIM level are also transformed into "Design" Class Diagram and "Detailed" Sequence Diagram in the PSM level
- The last step is the automatic generation of Source Code from PSM level.

This combination of MDA in every sprint of Scrum methodology have many advantages:

- Reduce the duration of the sprint that influence also the total duration of the project.
- Facilitate the management of requirements' changes of the system view that models and code are generated automatically.

E. Combination of MDA and V lifecycle in Scrum

As known scrum methodology proposes best practices to develop a qualitative software system respecting Agility principles, most of time developers prefer to combine scrum with life cycles to ensure the good management of the project development; in each sprint of Scrum, V life Cycle is used to define steps of the development of system, in this part of this section we will propose a combination of MDA and Scrum+V life Cycle.

V model means Verification and Validation model. Just like the waterfall model, the V life cycle is a sequential path of execution of processes. Each phase must be completed before the next phase begins; based on the requirement document that contains specifications of the system, developer team started working on the design and after completion on design start actual implementation and testing team starts working on test planning, test case writing, test scripting. Both activities are working parallel to each other.

Before presenting the combination of MDA and V lifecycle in scrum we present our MDA approach that consists on :

- Describing system requirements in CIM level by a structured English using SBVR.
- Transforming automatically Business Vocabulary and Business Rules of SBVR into Use Case Diagram (UCD) in CIM level [22]
- Applying transformation rules to generate Business Class Diagram (BCD) and System Sequence Diagram (SSD) from SBVR and UCD of CIM level to represent the PIM level [23].
- Generating PSM level of MVC architecture represented by Detailed Class Diagram (DCD) and Detailed Sequence Diagram (DSD) from PIM level using automatic transformation rules.
- Generating application source code from PSM level.

We mention that the automation of the two last steps of the approach will be implemented as an eclipse plugin which is the continuity of the previous one.

Choosing previous diagrams to model MDA levels is depending on different aspects that each level should cover according to OMG specifications:

- For CIM level, we define three aspects; Static and Dynamic that are covered by SBVR, while the Functional aspect is covered by UCD.
- For PIM level, we define two aspects; Structural aspect covered by BCD and Dy-namic one covered by SSD.
- For PSM level, we define in our approach 4 aspects; Static aspect covered by Model classes of DCD, Structural one covered by Class Diagram, Dynamic aspect covered by Controller classes and Behavioural one represented by DSD.

In this approach, we automate the two types of transformations Model-to-Model (M2M) and Model-to-Text (M2T). For M2M, we use QVT language while for M2T we use Acceleo transformation language. Transformation rules between the different levels of MDA are implemented as an Eclipse plugin to ensure automation and traceability of transformation rules. The figure below describes an overview of our approach:

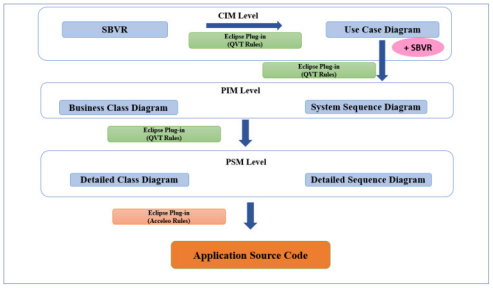


Fig. 2: Overview of our MDA transformation approach

Combining MDA and V life cycle in Scrum can be described as follows:

- Covering Requirements and functional specifications steps in V life cycle by CIM level of MDA which is represented in our approach by SBVR. The UCD and Business Rules generated at the CIM level are then used generate "Validation tests" to validate if the developed system responds to described requirements.
- Generating the High-level design represented in our approach by the PIM level which is generated automatically from CIM level (To generate PIM level from CIM one, we use our approach defined in our previous works (Essebaa and al, 2017). This step is represented by BCD and SSD for each use case element. We then generate "Integration tests" from these diagrams (BCD and SSD) to test the correct functioning between different elements of the system.
- Generating the low-level design represented by PSM level which is modelled by CD and DSD (The approach we propose to automate transformations between PIM and PSM levels will be discussed in our future works). We generate "Unit tests" from this level to test the generated code.

The figure 4 below describes the presented approach:

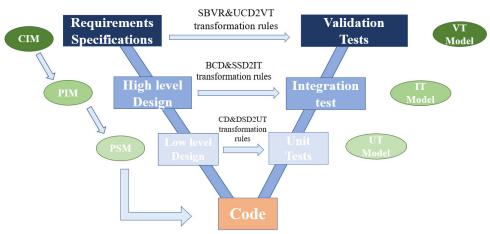


Fig. 3: Combination of MDA and V life cycle

To generate tests from Models in our approach, we defined three main rules that are detailed in our previous work [24]; Rule 1: Generate Validation tests from CIM level: Validation tests are generated from SBVR and Use Case Diagram. Rule 2: Generate Integration tests from PIM level: Integration tests are generated from PIM level which is represented using Class Diagram and System Sequence Diagram. Rule 3: Generate Unit tests from PSM level: Unit tests in our approach are generated from PSM level using Detailed Sequence Diagram and Detailed Class Diagram. The table 3 below summarize these rules:

TABLE III: Test generation rules

Rule	Model	Target
SBVR&UCD2VT	Use Case Element	Requirement to validate
	Fact Type	Sub feature to test
	Business rules of a fact type	Validation tests
BCD&SSD2IT	Actor and DataObject lifecycle	Classes to test
	Relationship between classes	Integration tests
DCD&DSD2UT	Messages	Operation to test
	Operation in classes	Unit tests

In the previous part we present how we automate transformations in MDA and combine with V lifecycle to generate different type of tests, in this part we will present our approach of managing system' requirement using MDA inside a V lifecycle in scrum method. Our proposal is divided into 5 main steps:

- Step 1: Defining system requirement by Backlog Product.
- Step 2: Planning features in a RoadMap.
- Sprint to Begin:
  - Step 3: Apply our approach that combine V lifecycle, MDA and MBT presented in parts 3.1 and 3.2 of section 3.
  - Step 4: Adding code missing parts manually preceding them with "@added" annotation.
- Step 5: Validation and planning of following sprints. In this step we define two cases:
  - If there is no system evolution:
    - Restart from step 3 for the next sprint
  - If there is an evolution:
    - Restart from step 1 and keep the old code except the parts preceded by "@added" annotation of features that still exist in the system (added code of deleted features is deleted automatically after the new execution)

In the figure 5 below we describe an overview of the presented approach in this paper:

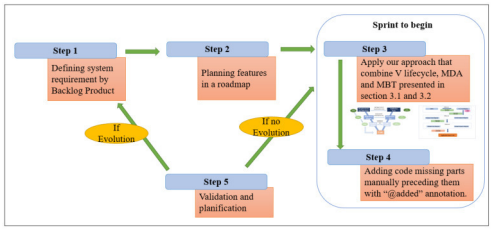


Fig. 4: Overview of a combination of MDA, MBT, V lifecycle in scrum

V. Implementation of our approach

To automate defined transformation rules in our approach, we need tools that allow to create input elements (UML Diagrams), and tools that support the generation of output elements (UML diagrams for PSM level and Source code). After analysing and testing existing tools, we decided to use Eclipse platform with different needed plugins to implement our approach, for example we choose Papyrus Modelling plugin because it supports all UML diagrams elements. According to this we choose to implement our solution as an eclipse plugin that will be a continuity of the previous developed plugin presented in [23] that automate transformations from CIM to PIM. Implementing our approach as an eclipse plugin is in order to facilitate the use of our transformation approach by designers and developers, and also to benefit from existing plugin in Eclipse.

To implement our transformations, we have to use a transformation language; there exist many models of transformations language for both types of transformations M2M and M2T, in our case we chose to use QVT language for M2M transformations and Acceleo language for M2T transformations.

These transformation rules are automated in our Eclipse plugin that take in an input a CIM level and generate PIM, PSM and source code.

The figure 6 shows "Transforms" menu that contains different items that allow the automation of rules:

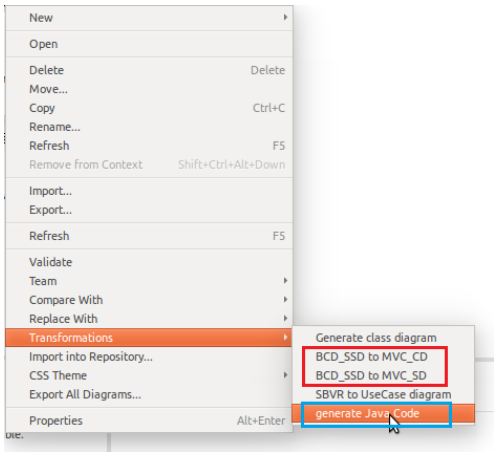


Fig. 5: Transforms menu of our plugin

VI. Case Study

To well illustrate our approach and transformation rules defined, we present in this section their application on a Rental Car Agency system. The case study must provide the following features:

- Visualization of available cars.
- Customers subscription.
- Cars booking.
- Visualization of reservations.
- Management of reservations (accept/decline) by a manager.
- Management of cars.
- Management of customers' accounts.
- Management of Managers' accounts.

The application has three users' profiles that have different privileges:

- Customer: A person who can view the cars available in the agency, rates and promotions and may subscribe. A client must register and authenticate in the system to search for available cars and book a car by indicating the reservation date and time.
- Manager: A Manager must also authenticate to view all cars, add, edit or remove cars. He can also view the bookings made by customers waiting for validation to decide to accept or refuse them.
- Administrator: Once authenticated into the system, the administrator has the privilege of modifying and deleting a customer account, as well as the management of managers account (add, change or delete)

We can also define some management rules as below:

- A customer can rent at least 1 car.
- A car can be rented by at least 1 customer.
- A manager can manage at least 1 car.
- A car is managed by at least 1 manager.
- An administrator can manage at least 1 customer account.
- An administrator can manage at least 1 manager account.

In the following part we present an application of our approach' steps on Rental Car Agency System example:

1) Defining a Backlog product by system requirements: After analyzing system requirements, the first step in our approach is to define the backlog product of the project then plan the RoadMap that describes different sprints of first project' requirement before any evolution, in this example we plan three sprints to develop the system, we define 3 sprints where each one takes 2 weeks. The figure 7 below describes the roadmap of Rental Car Agency system:

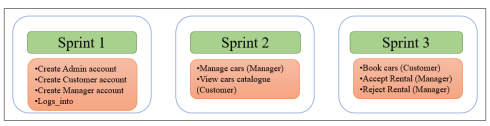


Fig. 6: Scrum RoadMap of Rental Car Agency System

2) Modelling user stories of the first sprint by SBVR and UCD to cover CIM level of MDA: The next step after dispatching features on sprints is to describe CIM level of first sprint by Business Vocabulary and Business Rules using SBVR standard as described in following figure 8.

```
account
login
    General_concept: text
password
    General_concept: text
account has login
    Synonymous_form: login is_property_of account

account has password
    Synonymous_form: password is_proprty_of account

customer owns account
    Synonymous_form: account is_owned_by customer

customer manages account
    Synonymous_form: account is_managed_by customer

admin owns account
    Synonymous_form: account is_owned_by admin

admin manages account
    Synonymous_form: account is_managed_by admin

manager owns account
    Synonymous_form: account is_owned_by manager
It is possible that customer owns at most 1 account.
It is possible that admin owns at most 1 account.
It is permitted that admin manages account.
It is possible that manager owns at most 1 account.
```

Fig. 7: Examples of SBVR of the first sprint of Rental Car agency

In the same level of MDA, we apply horizontal transformation rules, implemented as an eclipse plugin, to automatically generate UCD from SBVR. The figure 9 below represents the generated UCD of the first sprint for Rental Car Agency system.

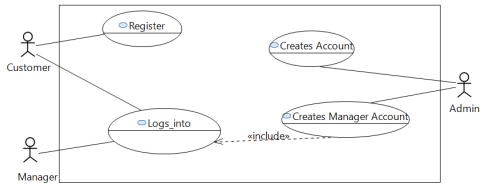


Fig. 8: UML Use Case Diagram of the first sprint of Rental Car Agency System

3) Generating Validation tests from CIM: As in our approach we use V lifecycle process combined with MDA, after defining CIM level, we can generate Validation tests from this level as described for "Logs-into" feature in table 4 below:

TABLE IV: Validation tests generation from PSM level

Source			Target		
Use Case element	Fact Type	Business rule	Requirement	sub feature	validation test
Logs_into	System requests user credential	It is obligatory that the system requests user credential if customer logs into system	The system must allow the customer to logs_into the system	Requests user credential	The system must request user credential if a customer try to logs into the system
	customer sends user credentials	It is necessary that customer sends user credentials if system requests user credential		sends use credentials	The system must allow customer to send user credential
	System verifies user credentials	It is obligatory that the system verifies user credentials if customer sends user credential		verifies user credential	The system must be able to check user credential
	System accepts user credentials	It is possible that system accepts user credential		accepts user credential	The system must be able to accept user credential

4) Applying transformation rules on CIM to generate BCD and SSD of PIM level: Generating PIM level is the first vertical Model-to-Model transformation that aims to automatically generate BCD and SSD from CIM level for Sprint 1 using our Eclipse plugin that implements transformation rules, the figure 10 below represents a BCD of PIM level of Rental Car Agency system:

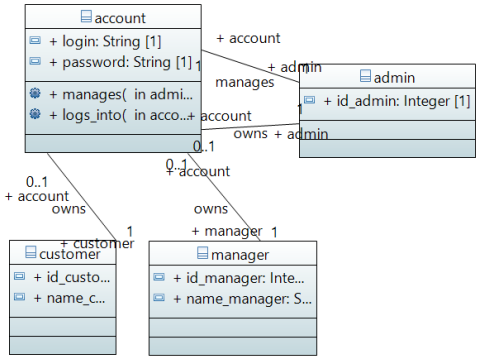


Fig. 9: Generated BCD of PIM level of the first sprint of Rental Car Agency System

The dynamic aspect of PIM level in our approach is represented also by different System Sequence Diagrams, the figure 11 describes System sequence diagram of "logs_into" feature in rental car agency.

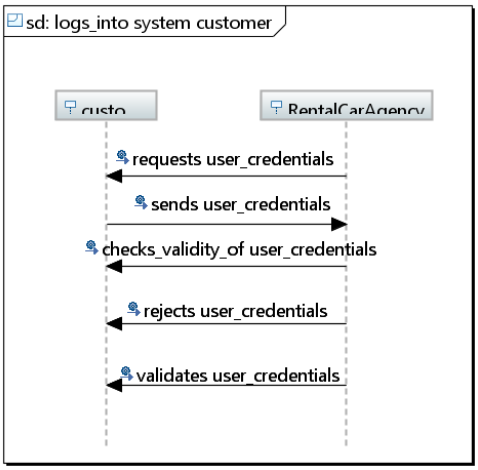


Fig. 10: Generated SSD of PIM level of the first sprint of Rental Car Agency System

5) Generating Integration tests from PIM: According to V lifecycle used in our approach, Integration Tests are automatically generated from PIM level that covers high level design of V lifecycle, the table 5 below describes Integration test for "logs_into" feature in sprint 1:

TABLE V: Integration test generation form PIM level

Source		Target	
Requirements	SD Connection	Classes	Integration tests
Logs_into	The operation requires connection between "Customer" and "Account	Customer	Customer owns 1 account
		Account	Account belongs to 1 customer

6) Applying transformation rules to generate DCD and DSD of PSM level from PIM: The last level before code is PSM level, which is the result of M2M transformations applied on PIM level to automatically generate DCD and DSD, the figure 12 below defines DCD of PSM level of sprint 1:

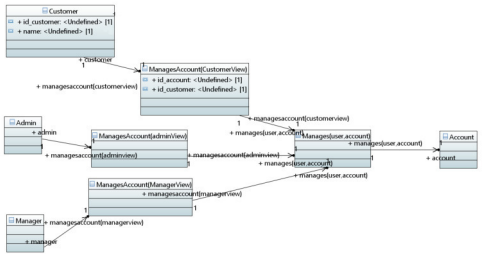


Fig. 11: Generated DCD of PSM level of the first sprint of Rental Car Agency System

The dynamic aspect of PIM level in our approach is represented also by different System Sequence Diagrams, the figure 12 describes Detailed sequence diagram of "logs_into" feature in rental car agency.

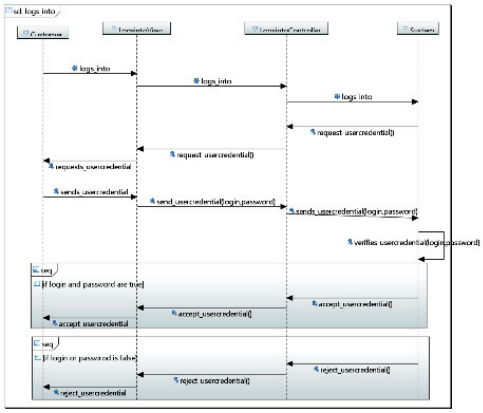


Fig. 12: Generated DSD of SIM level of the first sprint of Rental Car Agency System

7) Generating Unit tests from PSM: Unit tests in our approach are generated from low level design step of V lifecycle covered by PSM level, the table below describes example of Unit test of "logs_into" feature in sprint 1:

TABLE VI: Unit tests generation from PSM level

Source		Target	
Requirements	SD Messages	Operation to test	Unit tests
Logs_into	System requests User_credential	Request(User_credential)	Test "requests" operation
	Customer sends User_credential	Sends(User_credential)	Test "sends" operation
	System verifies User_credential	Verifies(User_credential)	Test "verifies" operation
	System accepts User_credential	Accepts(User_credential)	Test "accepts" operation

8) Generating application source code: The last transformation in our approach is automatic code generation which is the result of M2T transformations that takes as an input a DCD and DSD of PSM to generate as an output source code for MVC web application.

9) Adding missing parts of code manually: The last step in each sprint is to add manually the missing parts of code to complete the system' feature, this code must be preceded by "@adding" annotation.

Evolution of Rental Car Agency system' requirements

In this section we will make some evolutions to the system (addition, deletion and modification of features) in order to visualize the process of models' transformation and test generation in V lifecycle combined in scrum, the evolution will be as follow:

- *Modifications*: "View car catalogue" feature will be available for all users not only customers, this modification engender a new actor "User" that it will be a generalization of "Customer" actor, this modification requires changing the actor of "register" method too.
- *Addition*: In the new system, "Customer" will be able to validate its rental by "payment", The addition of a feature may engender some modifications to old ones, for example the verification of car availability will be made automatically by a system.
- *Deletion*: The addition of "payment" feature requires to delete "Manage rental" feature of "Manager" that allowed him to accept or reject the rental, in the new system the customer can validate its rental from the system, before proceeding to payment option the system must be able to check if the chosen car is available for date specified by the customer.

After studying system requirements' evolution, we have to make another feature dispatching on next sprints as presented in figure 14:

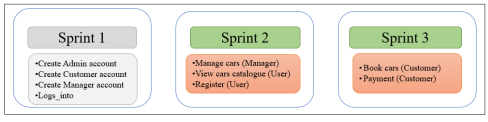


Fig. 13: Scrum RoadMap of Rental Car Agency system after requirements' changes

The next step is to develop the second sprint following same steps previously presented in an example for the first sprint even test cases that will be automatically generated.

VII. PERSPECTIVES AND FUTURE WORKS

This paper is an introduction to our future works on which we aim to combine MDA and Agility in order to propose an new method to develop software systems.

To combine MDA and Agility we aim in our future works to:

- Ensure the traceability between levels to manage requirements changes, as we know among agility principles is the adaptation to changes
- Propose an approach of Agile Model Driven Development appropriate to different types of Software systems (Web, Desktop,...)
- Propose a new Agile methodology that regroup all best practices proposed by other methodologies.

VIII. CONCLUSION

The basic idea in this paper is to combine the MDA approach and Agility in order to propose a new software development method, to well identify and situate the idea in the context, we describe in this paper a state of art of previous approach made in this context.

After analyzing previous works we conclude that proposed methods wasn't implemented to be applied to all types of software systems, we also present our analysis and reflexion on some existing agile methodologies in order to choose those more appropriate to be combined with MDA.

In this paper we propose to combine MDA and Scrum agile methodology in order to improve sprints of scrum and benefit from MDA principles, in this work we proposed to use V life cycle in each sprint of the project where we combine another variant of MDE, to generate automatically different tests applying MBT principles.

As previously mentioned this paper is an introduction to our future works on which we aim to propose a new methodology base on Agility and MDA.

REFERENCES

[1] A. Przybylek and M. Zakrzewski, "Adopting collaborative games into agile requirements engineering," in *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE,*, INSTICC. SciTePress, 2018, pp. 54–64.

[2] J. Miller and J. Mukerji, "Mda guide version 1.0.1." 2003.

[3] R.Soley, "Model driven architecture (mda)," in *http://www.omg.org/cgibin/doc?omg/00-11-05,*, 2000.

[4] K.Beck and al., "Agile manifesto," 2001–2015.

[5] T.Dyba and T.Dingsoyr, "What do we know about agile software development?" vol. 46, no. 5. Software, IEEE, 2009, pp. 6–9.

[6] J. P. Vickoff, "Agile why not?" in *www.entreprise-agile.com*, 2001.

[7] a. H. B. S. Hansson, Y. Zaho, "How mad are we? empirical evidence for model-driven agile development," in *Proceedings of XM 2014, 3rd Extreme Modeling Workshop*. CEUR, 2014.

[8] P. Cáceres, F. Díaz, and E. Marcos, "Integrating an agile process in a model driven architecture," in *INFORMATIK 2004 - Informatik verbindet, Band 1, Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Ulm, 20.-24. September 2004*, 2004, pp. 265–270.

[9] M. B. Nakićenović, "An agile driven architecture modernization to a model-driven development solution," *International Journal on Advances in Software*, vol. 5, no. 3,4, 2012.

[10] F. P. Basso, R. M. Pillat, F. Roos-Frantz, and R. Z. Frantz, "Combining mde and scrum on the rapid prototyping of web information systems," *Int. J. Web Engineering and Technology*, vol. 10, no. 3, 2015.

[11] V. Kulkarni, S. Barat, and U. Ramteerthkar, "Early experience with agile methodology in a model-driven approach," in *Model Driven Engineering Languages and Systems, 14th International Conference, MODELS 2011, Wellington, New Zealand, October 16-21, 2011. Proceedings*, 2011, pp. 578–590.

[12] H. Alfraihi, "Towards improving agility in model-driven development," in *Joint Proceedings of the Doctoral Symposium and Projects Showcase Held as Part of STAF 2016 co-located with Software Technologies: Applications and Foundations (STAF 2016), Vienna, Austria, July 4-7, 2016.*, 2016, pp. 2–10.

[13] H. Wegener, "Agility in model-driven software development? implications for organization, process, and architecture," in *OOPSLA 2002 Workshop on Generative Techniques in the Context of Model Driven Architecture,*, 2002.

[14] V. Mahe, B. Combemale, and J. Cadavid, "Crossing model driven engineering and agility – preliminary thoughts on benefits and challenges,," 2010.

[15] H. Burden, S. Hansson, and Y. Zhao, "How MAD are we? Empirical Evidence for Model-driven Agile Development," in *Proceedings of XM 2014, 3rd Extreme Modeling Workshop*, vol. 1239. Valencia, SPain: CEUR, September 2014, pp. 2–11.

[16] H. Alfraihi and K. Lano, "The integration of agile development and model driven development - a systematic literature review," in *Proceedings of the 5th International Conference on Model-Driven Engineering and Software Development - Volume 1: MODELSWARD,*, INSTICC. SciTePress, 2017, pp. 451–458.

[17] U. Eliasson and H. Burden, "Extending agile practices in automotive MDE," in *XM@MoDELS*, ser. CEUR Workshop Proceedings, vol. 1089. CEUR-WS.org, 2013, pp. 11–19.

[18] R. Matinnejad, "Agile model driven development: An intelligent compromise," in *Proceedings of the 2011 Ninth International Conference on Software Engineering Research, Management and Applications*, ser. SERA '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 197–202. [Online]. Available: https://doi.org/10.1109/SERA.2011.17

[19] K. Beck and C. Andres, *Extreme Programming Explained: Embrace Change (2Nd Edition)*. Addison-Wesley Professional, 2004.

[20] V. Jacobson, G. Booch, , and J. Rumbaugh, "Unified software development process." Addison-Wesley, 1992.

[21] P. Kurchten, "Rational unified process – an introduction." Addison-Wesley, 1999.

[22] I. Essebaa and S. Chantit, "Tool support to automate transformations from sbvr to uml use case diagram," in *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: MDI4SE,*, INSTICC. SciTePress, 2018, pp. 525–532.

[23] ——, "Tool support to automate transformations between cim and pim levels," in *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: MDI4SE,*, INSTICC. SciTePress, 2017, pp. 367–378.

[24] ——, "A combination of v development life cycle and model-based testing to deal with software system evolution issues," in *Proceedings of the 6th International Conference on Model-Driven Engineering and Software Development - Volume 1: MODELSWARD,*, INSTICC. SciTePress, 2018, pp. 528–535.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 949–957
DOI: 10.15439/2018F270
ISSN 2300-5963 ACSIS, Vol. 15

Scrum Adoption Challenges Detection Model: SACDM

Ridewaan Hanslo
University of South Africa, School of Computing,
College of Science, Engineering and Technology
Pretoria, South Africa
Email: ridewaan@gmail.com

Professor Ernest Mnkandla
University of South Africa, School of Computing,
College of Science, Engineering and Technology
Pretoria, South Africa
Email: mnkane@unisa.ac.za

***Abstract*—Scrum has been the most widely adopted Agile methodology over the past decade with Scrum and Scrum variants offering alternatives to the old software development methods. While Scrum plays an important role in the success of Agile development, it does come with its own challenges. In previous research challenges have been analyzed at the organizational and team level, primarily via case studies. However, fundamentally, Scrum needs to be adopted at the individual level. Furthermore, challenges such as inexperience, poor communication, specialization, lack of teamwork, low-quality, organizational culture and Scrum compatibility, have been identified as contributors. This paper therefore discusses the Scrum and Agile adoption challenges faced both globally as well as within the South African borders, from the findings of a narrative review. Secondly, a custom model adapted from the Diffusion of Innovation theoretical model was developed to detect the Scrum adoption challenges experienced within software organizations at the individual level. The custom model referred to as the Scrum Adoption Challenges Detection Model (SACDM) consists of four constructs, namely; individual factors, team factors, organizational factors and technology factors. The constructs are composed of nineteen independent variables that assists in understanding which factors contributes towards an individual either adopting or rejecting Scrum within a software organization. SACDM is therefore used to detect the adoption or rejection of Scrum as the dependent variable based on the independent variables being tested within the four constructs. The model can further be used with a survey questionnaire to provide generalized awareness of Scrum adoption challenges allowing software organizations to make more informed decisions when adopting Scrum. Future research is to allow the model to contribute towards Scrum adoption challenges predictive analysis.**

***Index Terms*—Adoption Challenges, Agile Methodologies, Scrum, Software Engineering, Software Organization.**

I. Introduction

"SOFTWARE development has become one of the world's most important practices. The software we produce today is rapidly becoming the embodiment of much of the world's intellectual property. Simply put, our modern world depends on software" [1].

Scrum, in the context of this paper, refers to an Agile methodology with emphasis on project management structure and communication between all stakeholders including clients and business representatives, regularly setting sprint time limits for software completion, reviewing changes and applying retrospection before working on the next product backlog requirements [2]. The software organization we refer to in this paper is any company, firm or organization that has a software development division or group of two or more individuals responsible for developing and maintaining software, for the benefit of the software organization or the client they service.

Scrum was developed in the early 1990's by Ken Schwaber and Jeff Sutherland [3]. Scrum is currently the most widely adopted Agile methodology, based on the 2017 VersionOne survey [4]. The reason for its high adoption rates could be its simplicity, as it can be easily understood by both business and Scrum teams alike.

Agile adoption (which includes Scrum) has its challenges such as work specialization, organizational culture, resistance to change, and lack of communication to name a few, however, what is certain is the fact that successful adoption improves numerous aspects of the business operation. Business operations include project visibility, manage change priorities, better aligned Information Technology (IT), increased productivity and enhanced software quality [4]. As stated by Rogers [5], "A technological innovation usually has at least some degree of benefit for its potential adopters".

What is evident from the reviewed literature is that, whilst there are common problems and challenges identified, there are very few empirical studies on the Scrum adoption challenges experienced by individuals. Most research focused on qualitative methods with emphasis on case studies [6]. A descriptive and explanatory case study done by Noruwana and Tanner [7] on Agile processes with emphasis on Scrum alludes that there is a knowledge base to unearth on adoption challenges.

To get an overview of adoption challenges faced by individuals within software organizations, it is necessary to get to the core of the challenges, i.e. what are the Scrum adoption challenges experienced in practice? Is there a relationship to be discovered with the adoption challenges and Scrum adoption outcomes? Will the knowledge and understanding point to a potential correlational or causal outcome?

Multiple theories, models and frameworks such as Diffusion of Innovation (DOI), Technology Acceptance Model (TAM), Perceived Characteristics of Innovations, and Theory of Planned Action, have been used to better understand the adoption and implementation of methodologies in software development [8].

We looked at the Scrum adoption challenges experienced through the lens of the DOI theoretical model. The DOI theoretical model was chosen instead of alternate theories because it was the only theory at the time to have been used both at the individual level and organizational level of IT adoption research, which meets the author's requirement. The custom model was divided into four constructs identified as individual factors, team factors, organization factors, and technology factors. The four constructs combine to form a holistic representation of the individual's belief, its relation to people, how they perceive management, and their perception of the methodology being used [6].

The aim of this paper is to propose a model which can be used to detect Scrum adoption challenges experienced by individuals within software organizations. The constructs of the custom model are an adaptation of the DOI theoretical model and the conceptual framework of the object-orientated technology (OT) study by Sultan and Chan [31]. The independent variables within the model's constructs are generated using the narrative review method. The proposed model will be used to differentiate adopters from non-adopters of the Scrum methodology, respectively.

Section II provides a brief background on Scrum as an Agile methodology, followed by the global and South African (SA) Scrum adoption challenges compiled from extant literature. Section III lists the constructs with a discussion of the variables with its hypothesized relationship to Scrum adoption and section IV explains the composition of the proposed SACDM. Section V concludes the paper.

II. Scrum Adoption Challenges

The Scrum Guide written by Schwaber and Sutherland [1] states the following about Scrum: "A framework within which people can address complex adaptive problems, while productively and creatively delivering products of the highest possible value. Scrum is:

- Lightweight
- Simple to understand
- Extremely difficult to master"

Scrum is a value-driven method (as opposed to a plan-driven method such as the waterfall method) which is iterative and incremental development [9]. The Scrum value-driven method continuously reassesses the problem while making small software feature increments in short time blocks within small teams [10]. Scrum is so flexible and abstract in its definition and implementation that it is often used outside of the Software Engineering (SE) practice [2].

Adoption challenges, in the context of this paper, refers to the challenges faced by software organizations when choosing and following an Agile methodology [11]. As mentioned in the introduction we used the narrative review method to generate the custom model's independent variables. Before we were able to generate the independent variables, we first had to identify the adoption challenges faced within the global and SA context. The Scrum and Agile adoption challenges were acquired through the narrative review method, which was as follows;

- Data sources was relatively recent i.e. all except one paper was less than ten years of age.
- It has been cited in other literature.
- If not cited, the source must have been published by an accredited publisher, e.g. Springer, Pearson, Institute of Electrical and Electronics Engineers (IEEE), International Journal of Environmental Science and Technology (IJEST), etc.
- If not published by an accredited publisher, the source must have been presented at a known institution, e.g. Agile Africa conference, Johannesburg Centre for Software Engineering (JCSE).
- Alternatively, the source is a recent dissertation or thesis paper.
- The primary search terms were 'Scrum', 'Scrum adoption', 'Scrum challenges', 'Agile challenges', 'Scrum South Africa', 'Agile South Africa' and 'Agile adoption'.
- The sources where carefully perused and relevant literature was ear marked for further investigation.
- These pre-selected literature sources were filtered based on the content it provided, i.e. Do the literature sources contain challenges and issues experienced during Scrum and Agile adoption? Or is the literature describing adoption challenges on irrelevant Software Development Methodologies (SDM)?
- Identified challenges within the literature was collated and the frequency of occurrence was recorded.

Table I is a consolidated list of global Scrum and Agile adoption challenges taken from twenty-one literature studies Stray et al. [12], [32], Asnawi et al. [13], Santos et al. [14], Fægri [15], Marchenko and Abrahamsson [16], Overhage et al. [17], Heikkila et al. [18], Kapitsaki and Christou [38],

Bjarnason and Regnell [39], Irrazabal et al. [40], Hoda et al. [41], [42], Dorairaj et al. [43], Senapathi et al. [44], Ressin et al. [45], [46], Santos and Goldman [47], Kim and Ryoo [48], Ihme [49], and Allisy-Roberts et al. [4], with publication years ranging from 2008 to 2017. Of the listed challenges, the one that is very peculiar comes from the mixed mode study by Heikkila et al. [18], which recorded that cross functional generalist teams were not plausible in the environment. This is contradictory to the Scrum philosophy of well-balanced redundant knowledge teams with the ability to work on various aspects of projects without the dependency of team member specialization.

TABLE I.
GLOBAL SCRUM AND AGILE ADOPTION CHALLENGES

No.	Global Scrum and Agile Adoption Challenges	Frequency
1	Lack of knowledge/training/skills	11
2	Organizational culture/mindset	9
3	Teamwork/communication issues	9
4	Lack of documentation	5
5	Budget and schedule constraint	2
6	Escalating commitment	2
7	Hard to scale	2
8	High management overhead	2
9	Lack of senior support	2
10	Work specialization	2
11	Cross functional generalist teams	1
12	Increase stress and workload	1
13	Lack of quality	1
14	Lack of top management support	1
15	Long time to market	1
16	Low user satisfaction	1
17	Over engineered solutions	1
18	Over optimistic task estimates	1
19	Project team size	1
20	Requirements creep	1
21	Retrospective inadequacy	1
22	Too many meetings	1

Top Management Support (TMS) has been found to significantly affect the user's perception of an IT technology, and the organizations IT adoption and diffusion, respectively [19], [20]. Therefore, the inclusion of lack of TMS is probably expected, considering the impact management support have on IT adoption [21]. It should be noted that although TMS is important for the adoption and diffusion of a methodology, it cannot save a project that is failing, and too much support might hinder the adoption and diffusion success [19].

Table II is a consolidated list of SA Scrum and Agile adoption challenges taken from six literature studies Mnkandla and Dwolatzky [22], Du Toit [23], Tanner and Khalane [11], Tanner and Mackinnon [24], Tanner and Wallace [31], and Noruwana and Tanner [7], with publication years ranging from 2004 to 2013. The rational for displaying the global and SA adoption challenges separately was to allow the authors to compare the two tables and identify if region had a role in challenges experienced. Most of the challenges within the SA literature was experienced globally, however, within the SA literature, organizational challenges were prevalent within the top challenges (by frequency), and while globally team challenges were more predominant.

It should be made clear that when referring to communication problems, it includes clients, and not just the individuals within the software organization. Especially in Scrum, clients are expected to be more collaborative, knowledgeable and representative, and committed towards the projects [26]. The importance of customer's active involvement in the development process is crucial to the success of Agile development [52]. The greater the involvement of customers during the development process the greater the chance of success [6].

Mohan and Ahlemann [27] says that the use of the Information Systems Development (ISD) process is determined by the rational and hierarchy of the organizational culture. Often the needs, beliefs and values of the users of the methodology are not considered, which is like the subjective norm situation, whereby the developer's views are not always the determinant to the Agile methodology adoption decision. As Hardgrave et al. [28] puts it; "Developer's intentions are directly influenced by their perceptions of usefulness, social pressure, compatibility and organizational mandate". Chan and Thong [26] indicates that prior SDM studies focused on the developer views of the SDM such as perceived ease of use and perceived usefulness, however, failing to realize the importance of management (e.g. management style) and people-related (e.g. competency levels) challenges.

TABLE II.
SA SCRUM AND AGILE ADOPTION CHALLENGES

No.	SA Scrum and Agile Adoption Challenges	Frequency
1	Lack of knowledge/training/skills	6
2	Organizational culture/mindset	5
3	Lack of structure/planning	5
4	Requirements creep/story changes	5
5	Communication issues	4
6	Motivational issues	4
7	Lack of resources (labor and non-labor resources)	3
8	Management inefficiencies	3
9	Workload	3
10	Team distribution	2
11	No/lack of individual recognition	1
12	Team size	1

Due to the nature of software development being a social phenomenon, and Agile being at the forefront of this complex human interaction activity [25], expectations that noise or disturbance by team members would have been identified as one of the challenges encountered, are aroused. However, surprisingly this is not the case, and the study by Eccles et al. [29] states, on the contrary, employees welcome it.

The next section discusses the independent variables and their hypothesized relationship with Scrum adoption, which forms part of the custom model's four constructs.

III. The Custom Model Constructs and Variables

The custom model was constructed with DOI being the theoretical base, but we tailored the model to match the context of the application, i.e. Scrum adoption challenges [31]. Not all DOI constructs were used in the study, the three that have been included, due to it being consistently relevant in innovation studies are compatibility, complexity, and relative advantage [28]. The custom model is discussed in section IV. Scrum Adoption Challenges Detection Model (SACDM).

The narrative review method produced the independent variables which were collated and coded, and subsequently used as the input to the model. These independent variables were assigned to either one of the four custom model factors, namely; individual, team, organization and technology. Therefore, the independent variables were tailored towards the specificity of the innovation [30], [31]. The nineteen independent variables affecting Scrum adoption are discussed in this section.

A. Individual Factors

The first set of variables found in the literature deals with the Scrum challenges experienced by individuals within the organization.

Escalation of Commitment: Escalation of Commitment in the software industry context, its defined as continuously assigning resources to projects that indicates signs of failure. Statistics of 30 to 40 percent of software projects that experiences escalation of commitment have been recorded [32]. We have included escalation of commitment to the individual factors construct because it has often been caused by individual developers within Scrum teams who persist with a task even though it is not adding value to the project. The sooner the Scrum team notices this problem (usually in daily stand-ups) the greater the chances of limiting resource wastage.

H1: Thus, it has been hypothesized that escalation of commitment negatively affects the adoption of Scrum within IT organizations.

Experience: While experience may be seen as being knowledgeable and skilled on an event or subject, it also refers to the project team member having mastery of multiple skills sets such as programming languages, management skills etc. The mastery of multiple skill sets is usually obtained by working on various tasks, projects, and teams over a period of time [26]. Experience has also been identified as a contributor to performance of programmers [33].

H2: Experience is therefore hypothesized to have a positive influence on individual willingness to adopt new innovations.

Over Engineering: Over engineering or over engineered solutions can be summarized as software that has more features and functionality added to it than what was required from the client. Reasons that could lead to software being over engineered are lack of communication with stakeholders, bad planning or limited domain knowledge by the Scrum team [14]. This variable has been included as an individual factor because the developers within the development team are responsible for completing the sprint backlog. The development team is included in the sprint planning meeting and if anything related to the backlog item is unclear to the developer during the sprint he or she may liaise with the Scrum team to clear any confusion. We therefore think that over engineering affects innovation adoption negatively.

H3: Over engineering is negatively related to adoption of Scrum. Over engineering will be lower for adopters.

B. Team Factors

The second set of variables is concerned with the individual's perception of team related challenges based on the literature.

Communication: Communication is the act of exchanging information from one individual or group to another using a common system of behavior [26].

H4: Lack of communication therefore have a negative impact on adoption. Communication will be higher for adopters.

Teamwork: Teamwork is the process whereby individuals work together as a team to complete tasks and achieve a common goal or objective [26]. However, teamwork challenges within Agile development methods is a reoccurring problem. Activities which have been documented as important to increase team as well as organizational performance, are, recognizing other's achievement, responding constructively to team member opinions, assisting and supporting others, and showing greater leniency towards team members [12].

H5: More teamwork amongst team members will affect Scrum adoption positively. Adopters will be in a team that show greater signs of teamwork than non-adopters.

Specialization: The term specialization is the process of an individual having a high degree of knowledge and skills within a domain of interest, improving the individual's

proficiency and expertise within his or her role. Agile software development teams prioritize the idea of self-organizing teams in which team members share overlapping skills which improves flexibility. The problem with work specialization is that it doesn't make provision for interchangeable roles [15].

H6: Specialization within a Scrum team negatively affects adoption. Specialization will be less for adopters.

Sprint Management: Sprint management is defined as a time boxed activity that monitors and manages the progress of a sprint. Events that prevents sprint cycles from operating optimally includes scope creep, lack of timeous feedback, lack of planning and lack of team cohesion [11], [24].

H7: The better the sprint management within the team the more likely there will be adoption. Adopters will have better sprint management.

Change Resistant: Resistance to change within the context of the work environment is a process in which the employee sees change as disruptive and intrusive [34]. With Agile process introduction, developers tend to display signs of cautious optimism, skepticism, and enthusiasm with the problem of some developers not welcoming the change, resisting it without much thought put into it [35].

H8: Teams that are reluctant to change their ways of doing things are more likely to be non-adopters. Adopters will have a lower degree of change resistance.

C. Organization Factors

The third set of variables deals with the individual's perception of organizational challenges encountered within literature.

Training: Training is the acquisition of skills and knowledge through teaching and learning which improves the competency areas of the individual or group. The training within this research study applies to employees going for training to achieve the goals and objectives of the organization they represent [26].

H9: It is hypothesized that staff training is higher in organizations that adopted Scrum. Lack of staff training is hypothesized to negatively affect adoption.

Recognition: Recognition from a business point of view, is seen as matching remuneration, rewards and benefits with the productivity levels of the workers [36]. The study by Noruwana and Tanner [7] identified that individuals were unhappy with the lack of recognition for their contributions within the team because the recognition was given on a team level which does not distinguish between team member productivity levels.

H10: Therefore, individual recognition is hypothesized to improve the likelihood of adoption. Recognition contributes positively towards adoption success.

Quality: The quality that is being referred to is that of software quality and how its correctness contributes toward software projects meeting the business requirements and user expectations. There have been many attempts to improve the quality of software project throughput, yet many software projects continue to fail [11].

H11: Higher degree of throughput quality is positively correlated to Scrum adoption.

Resources: Resources in the context of this study refers to any asset or service, whether it is staff, materials, money etc. that allows the organization to operate sufficiently in producing products and services requested by clients. An exploratory case study conducted by Noruwana and Tanner [7] on a SA company identified lack of labor resources ranging from Agile experience, skillsets, and team members having to perform more than their fair share of responsibilities.

H12: Supply of labor and non-labor resources are more for adopters than non-adopters. Lack of resources is hypothesized to have a negative impact on adoption.

Collaboration: Included in the Agile Manifesto is the statement "Customer collaboration over contract negotiation". What this suggests is that individuals, teams and organizations need to work closely together with clients to achieve a common goal instead of spending most of their effort on securing the deal. Research indicates that many organizations and customers within Agile environments do not abide by this principle. Some of the challenges faced by the lack of collaboration are Agile teams being overly committed, loss of business and productivity, products and user requirements not aligning, and poor feedback mechanisms [37].

H13: Adopters have more collaboration with their clients than non-adopters. Collaboration is positively correlated to adoption.

Management Support: Management support allow organizations to look at innovation adoption from a positive perspective, and this creates a conducive environment for innovativeness [26]. Two findings that are of interest for this study is firstly, management that penalizes employees for mistakes made does not encourage innovativeness, and secondly, management support has a direct effect on the adoption of innovation [31].

H14: Therefore, management support is hypothesized to be a crucial contributor to the adoption of Scrum. Management support will be higher for adopters.

Organizational Culture: Organizational culture which is so eloquently defined by E.H. Schein (1990) is quoted as saying "a pattern of basic assumptions invented, discovered or developed by a given group as it learns to cope with its problems of external adaptation and integration that has worked well enough to be considered valid and, therefore, is to be taught to new members as the correct way to perceive, think, and feel in relation to those problems" [26].

H15: A supportive company culture is positively related to Scrum adoption. Individuals who adopt Scrum will be in firms with a supportive organizational culture.

Organizational Structure: The organization structure is a system with defined activities which governs how individuals within roles, and procedures are coordinated to achieve the goals and objectives of the organization. Evidence from previous studies indicates that organizations that allow for an open and integrated environment with a less hierarchical structure improves the innovation adoption rates [31]. Whilst previous studies have broken up organizational structure into the three components of centralization, formalization and integration, the authors however kept it as a single variable for reasons of simplicity.

H16: Organizational structure contributes negatively towards Scrum adoption. Lack of structure is hypothesized to be higher in organizations who adopt Scrum.

D. Technology Factors

The fourth set of variables relates to the DOI theory and the individual's perception of the Scrum methodology as an innovation.

Relative Advantage: Relative advantage is measured (within the context of this study) as the degree to which Scrum has made a positive contribution to the existing conditions of the individual and organization [31].

H17: There is a linear relationship between perceived relative advantage and Scrum adoption. Adopters will perceive higher relative advantage in the methodology.

Complexity: Complexity is the degree of difficulty experienced by individuals and organizations when adopting Scrum as an innovation [31].

H18: There is a linear relationship between perceived complexity and the adoption of Scrum. Non-adopters will perceive a higher degree of complexity in Scrum than adopters.

Compatibility: The compatibility of Scrum against the existing values of the company and the individuals whom it employs provide an indication to the likelihood of individuals adopting or rejecting it [31].

H19: There is more compatibility with adopters of Scrum than companies and individuals that rejects it. The higher the compatibility the more likely there is the potential of adoption.

The following section is dedicated to proposing a practical application model to detect the presence of Scrum adoption challenges encountered by individuals within software organizations. The aforementioned hypotheses are tested with the use of this model.

IV. Scrum Adoption Challenges Detection Model (SACDM)

While DOI as a theoretical model covers both the individual and organizational aspects of IT adoption studies [31], it is not enough though for complex methodologies within Agile, such as Scrum.

According to Chau and Tam [30], diffusion variables are not sufficient enough as a predictor of complex organizational innovation adoption, as the independent and control variables it provides might be of limitation. Bayer and Melone [50] provides a few failures of DOI due to its limitations, two of the failures being the lack of theoretical justification for the five adopter categories without sufficient empirical support for the classifications used, and not taking the interactions between various social systems into account.

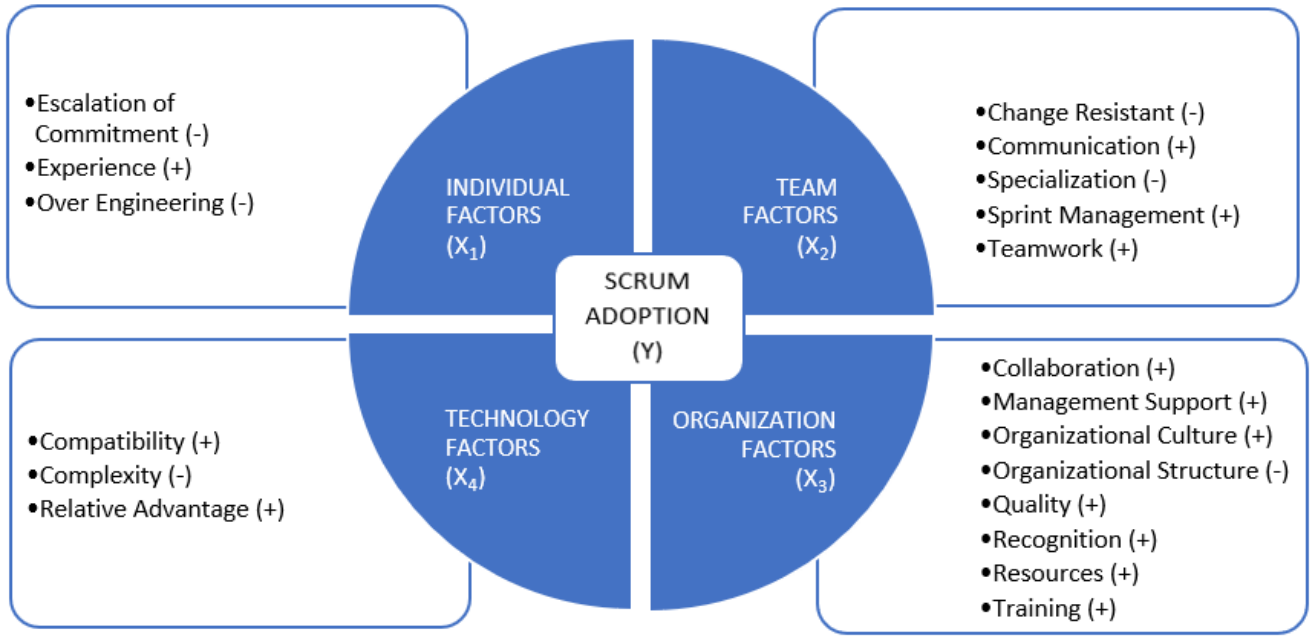
Because the Scrum methodology is a social phenomenon with strong emphasis on project management, it is important that we develop a model that include behavioral aspects to it, which unfortunately, haven't received much attention by previous IS adoption studies [51]. As Chan and Thong [26] so eloquently puts it, "There is an urgent need to conduct a critical review of the extant literature to develop a conceptual framework (CF) for Agile methodologies acceptance."

We used the idea of Senapathi et al. [44], who developed a CF based on a synthesis of past research in DOI, agile implementation, and IS implementation literature. Their five factor groups are agile innovation, organizational, sociological, team, and technological factors, which have been adopted from agile, XP, DOI, and IS frameworks and literature.

With a similar approach this study uses a CF which is a synthesis of research composed of theoretical models, DOI, Agile adoption, Scrum adoption, SDM adoption, and IS innovation literature.

As indicated in the introduction of this paper, a model will assist to detect Scrum adoption challenges within software organizations. The detection of the challenges can be generalized with the help of a quantitative survey research design. In future studies the model can contribute towards predictive analysis of Scrum adoption. The custom model is adapted from the study by Sultan and Chan [31], which looked at the adoption of OT in software companies. The authors propose a Scrum adoption challenges detection model, incorporating the DOI theoretical model, by extending the theoretical model to include constructs such as organizational factors, team factors and individual factors (SACDM) as displayed in Fig. 1.

Fig. 1 displays the custom model with the four constructs comprising of nineteen independent variables, and one dependent variable. The model for this research indicates the variables which are hypothesized to have an influence in adoption of a new methodology such as Scrum by individuals, and the proposed directionality of these relationships.



Independent variables are depicted as X_1, X_2, X_3 and X_4.

Dependent variable is Y with $Y = f(X_1.X_2.X_3.X_4)$.

When Y = 1, the individual within a software organization is an adopter of Scrum.

When Y = 0, the individual within a software organization is a non-adopter of Scrum.

Note: The hypothesized relationships are shown by the symbols in parenthesis.

Fig. 1. Scrum Adoption Challenges Detection Model.

The final list of independent variables is displayed below, in ascending order.

- Change Resistant
- Collaboration
- Communication
- Compatibility
- Complexity
- Escalation of Commitment
- Experience
- Management Support
- Organizational Culture
- Organizational Structure
- Over Engineering
- Quality
- Recognition
- Relative Advantage
- Resources
- Specialization
- Sprint Management
- Teamwork
- Training

This model will be used to differentiate adopters from non-adopters of Scrum, which is important to understand which constructs and variables significantly contributes towards the acceptance or rejection of Scrum. The dependent variable in this study (Y) is the adoption of Scrum. This will be related to the independent variables included in the four sets of variables shown in Fig. 1: (X_1) individual factors; (X_2) team factors; (X_3) organizational factors; and (X_4) technology factors.

This paper concludes by providing a summary of the SACDM and the potential advantages it may have for software organizations looking to improve on their project management operations.

V. Conclusion

Scrum is the most widely used SDM at present, providing many organizations with a simple to understand methodology to complete project management tasks. While the advantages to using this methodology are easily noticed by adopters, the challenges during the adoption stage are currently not quantitatively detected. Most research on the

adoption challenges primarily focused on qualitative measures for detection with case studies being the most implemented strategy. The inability to easily detect these adoption challenges can lead to teams and individuals within software organizations not using Scrum correctly or not adopting Scrum altogether, which could potentially limit the successful outcomes of a project.

It is proposed that a practically applied Scrum adoption challenges detection model such as SACDM, will aid in the awareness of the challenges faced by software organizations, and thus potentially limit the negative effects these adoption challenges might have on the individuals and organizations using Scrum. The extant Scrum adoption challenges were acquired through a narrative review of Scrum adoption challenges, both within the global and SA context. The SACDM was developed to detect Scrum adoption challenges with the objective of equipping adopters with the knowledge and awareness to overcome them.

Future research will aim to improve the SACDM, by designing an automated Scrum adoption challenges self-evaluation questionnaire. This questionnaire will allow the authors to gather and analyze the response data, which will be used to create a generalized result-set for the benefit of potential adopters to improve their awareness of Scrum adoption challenges and the correlation to Scrum adoption. The long-term vision of the SACDM is to allow individuals and organizations to predict Scrum adoption with the help of a research database and algorithms used to perform Scrum adoption predictive analysis.

REFERENCES

[1] Leffingwell, D. 2011. Agile software requirements: lean requirements practices for teams, programs, and the enterprise. Boston: Pearson Education, Inc.

[2] Schwaber, K. & Sutherland, J. 2011. The Scrum Guide. Scrum.org, October, 2: 17. http://www.Scrumalliance.org/.

[3] Pressman, R.S. 2005. Software Engineering. A Practitioner's Approach. 6th ed. New York, USA: McGraw-Hill.

[4] Allisy-Roberts, P., Ambrosi, P., Bartlett, D.T., Coursey, B.M., DeWerd, L.A., Fantuzzi, E. & McDonald, J.C. 2017. The 11th Annual State of Agile Report. Journal of the ICRU, 6(2): 7–8. https://academic.oup.com/jicru/article-lookup/doi/10.1093/jicru/ndl025.

[5] Rogers, E.M. 2003. Diffusion of Innovations, 5th Edition. Free Press. https://books.google.co.za/books?id=9U1K5LjUOwEC.

[6] Chan, K.Y. & Thong, J.Y.L. 2007. An Integrated Framework of Individual Acceptance of Agile Methodologies. PACIS 2007 Proceedings: 154.

[7] Noruwana, N. & Tanner, M. 2012. Understanding the structured processes followed by organisations prior to engaging in Agile processes: A South African Perspective. SACJ, (48): 8.

[8] Vijayasarathy, L. & Turk, D. 2012. Drivers of Agile software development use: Dialectic interplay between benefits and hindrances. Information and Software Technology, 54(2): 137–148.

[9] Anderson, D.J., Concas, G., Lunesu, M.I., Marchesi, M. & Zhang, H. 2012. A Comparative Study of Scrum and Kanban Approaches on a Real Case Study Using Simulation. In C. Wohlin, ed. Agile Processes in Software Engineering and Extreme Programming. Malmö: Springer Berlin Heidelberg: 123–137.

[10] Blankenship, J., Bussa, M. & Millett, S. 2011. Pro Agile .NET Development with Scrum. Berkeley, CA: Apress. http://link.springer.com/10.1007/978-1-4302-3534-7.

[11] Tanner, M., Khalane, T. 2013. Software Quality Assurance in Scrum: The need for concrete guidance on SQA strategies in meeting user expectations. IEEE ICAST 2013: 6.

[12] Stray, V.G., Moe, N.B. & Dingsøyr, T. 2011. Challenges to Teamwork: A Multiple Case Study of Two Agile Teams. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 146–161.

[13] Asnawi, A.L., Gravell, A.M. & Wills, G.B. 2011. Empirical Investigation on Agile Methods Usage: Issues Identified from Early Adopters in Malaysia. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 192–207.

[14] Santos, R., Flentge, F., Begin, M.-E. & Navarro, V. 2011. Agile Technical Management of Industrial Contracts: Scrum Development of Ground Segment Software at the European Space Agency. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 290–305.

[15] Fægri, T.E. 2010. Adoption of Team Estimation in a Specialist Organizational Environment. In A. Sillitti, A. Martin, X. Wang, & E. Whitworth, eds. Agile Processes in Software Engineering and Extreme Programming. Trondheim: Springer Berlin Heidelberg: 28–42.

[16] Marchenko, A. & Abrahamsson, P. 2008. Scrum in a Multiproject Environment: An Ethnographically-Inspired Case Study on the Adoption Challenges. In Agile 2008 Conference. IEEE: 15–26. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4599449.

[17] Overhage, S., Schlauderer, S., Birkmeier, D. & Miller, J. 2011. What Makes IT Personnel Adopt Scrum? A Framework of Drivers and Inhibitors to Developer Acceptance. In 2011 44th Hawaii International Conference on System Sciences. IEEE: 1–10. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5718964.

[18] Heikkila, V.T., Paasivaara, M. & Lassenius, C. 2013. ScrumBut, But Does it Matter? A Mixed-Method Study of the Planning Process of a Multi-team Scrum Organization. In 2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement. IEEE: 85–94. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6681341.

[19] Dong, L. 2008. Exploring the impact of top management support of enterprise systems implementations outcomes: Two cases. Business Process Management Journal, 14(2): 204–218.

[20] Dong, L., Neufeld, D. & Higgins, C. 2009. Top management support of enterprise systems implementations. Journal of Information technology, 24(1): 55–80.

[21] Hardgrave, B.C. & Johnson, R.A. 2003. Toward an information systems development acceptance model: the case of object-oriented systems development. IEEE Transactions on Engineering Management, 50(3): 322–336.

[22] Mnkandla, E., Dwolatzky, B. 2004. A survey of Agile methodologies. THE TRANSACTIONS OF THE SA INSTITUTE OF ELECTRICAL ENGINEERS, 3(December): 236–247.

[23] Du Toit, R. 2013. Enterprise Agile Adoption with Parental Guidance - PG. http://Agileafrica.jcse.org.za/sites/default/files/Riaan-du-Doit-Agile-Adoption-with-Parental-Guidance.pdf 12 August 2014.

[24] Tanner, M. & Mackinnon, A. 2013. Sources of Disturbances Experienced During a Scrum Sprint. ICIME2013.

[25] Tanner, M., Wallace, C. 2012. TOWARDS AN UNDERSTANDING OF THE CONTEXTUAL INFLUENCES ON DISTRIBUTED AGILE SOFTWARE DEVELOPMENT : A THEORY OF PACTICE PERSPECTIVE. In European Conference on Information Systems (ECIS). Association for Information Systems AIS Electronic Library (AISeL): 13.

[26] Chan, F.K.Y. & Thong, J.Y.L. 2009. Acceptance of Agile methodologies: A critical review and conceptual framework. Decision support systems, 46(4): 803–814.

[27] Mohan, K. & Ahlemann, F. 2013. Understanding acceptance of information system development and management methodologies by actual users: A review and assessment of existing literature. International Journal of Information Management, 33(5): 831–839.

[28] Hardgrave, B.C., Davis, F.D. & Riemenschneider, C.K. 2003. Investigating determinants of software developers' intentions to follow methodologies. Journal of Management Information Systems, 20(1): 123–151.

[29] Eccles, M., Smith, J., Tanner, M., Van Belle, J.-P. & van der Watt, S. 2010. The Impact of Collocation on the Effectiveness of Agile IS Development Teams. Communications of the IBIMA, 2010: 1–11.

[30] Chau, P.Y.K. & Tam, K.Y. 1997. Factors affecting the adoption of open systems: an exploratory study. MIS quarterly: 1–24.

[31] Sultan, F. & Chan, L. 2000. The adoption of new technology: the case of object-oriented computing in software companies. IEEE transactions on Engineering Management, 47(1): 106–126.

[32] Stray, V.G., Moe, N.B. & Dybå, T. 2012. Escalation of Commitment: A Longitudinal Case Study of Daily Meetings. In C. Wohlin, ed. Lecture Notes in Business Information Processing. Malmö: Springer Berlin Heidelberg: 153–167. http://link.springer.com/10.1007/978-3-642-30350-0_11.

[33] Brooks, R.E. 1980. Studying programmer behavior experimentally: The problems of proper methodology. Communications of the ACM, 23(4): 207–213.

[34] Strebel, P. 1996. Why do employees resist change? Harvard business review, 74(3): 86.

[35] Cohn, M. & Ford, D. 2003. Introducing an Agile process to an organization [software development]. Computer, 36(6): 74–78.

[36] Bishop, J. 1987. The recognition and reward of employee performance. Journal of Labor Economics, 5(4, Part 2): S36–S56.

[37] Hoda, R., Noble, J. & Marshall, S. 2011b. The impact of inadequate customer collaboration on self-organizing Agile teams. Information and Software Technology, 53(5): 521–534.

[38] Kapitsaki, G.M. & Christou, M. 2014. Where Is Scrum in the Current Agile World? In Proceedings of the 9th International Conference on Evaluation of Novel Approaches to Software Engineering. SCITEPRESS – Science and Technology Publications: 101–108. http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004867701010108.

[39] Bjarnason, E. & Regnell, B. 2012. Evidence-Based Timelines for Agile Project Retrospectives – A Method Proposal. In C. Wohlin, ed. Lecture Notes in Business Information Processing. Malmö: Springer Berlin Heidelberg: 177–184. http://link.springer.com/10.1007/978-3-642-30350-0_13.

[40] Irrazabal, E., Vásquez, F., Díaz, R. & Garzás, J. 2011. Applying ISO/IEC 12207:2008 with SCRUM and Agile Methods. In R. V. O'Connor, T. Rout, F. McCaffery, & A. Dorling, eds. Software Process Improvement and Capability Determination. Dublin: Springer Berlin Heidelberg: 169–180.

[41] Hoda, R., Noble, J. & Marshall, S. 2011a. Supporting Self-organizing Agile Teams. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 73–87. http://link.springer.com/10.1007/978-3-642-20677-1_6.

[42] Hoda, R., Noble, J. & Marshall, S. 2010. Agile Undercover: When Customers Don't Collaborate. In A. Sillitti, A. Martin, X. Wang, & E. Whitworth, eds. Agile Processes in Software Engineering and Extreme Programming. Trondheim: Springer Berlin Heidelberg: 73–87.

[43] Dorairaj, S., Noble, J. & Malik, P. 2011. Effective Communication in Distributed Agile Software Development Teams. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 102–116.

[44] Senapathi, M., Middleton, P. & Evans, G. 2011. Factors Affecting Effectiveness of Agile Usage – Insights from the BBC Worldwide Case Study. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 132–145.

[45] Ressin, M., Abdelnour-Nocera, J. & Smith, A. 2011a. Defects and Agility: Localization Issues in Agile Development Projects. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 290–305.

[46] Ressin, M., Abdelnour-Nocera, J. & Smith, A. 2011b. Lost in Agility? Approaching Software Localization in Agile Software Development. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 320–321.

[47] Santos, V. & Goldman, A. 2011. An Approach on Applying Organizational Learning in Agile Software Organizations. In A. Sillitti, O. Hazzan, E. Bache, & X. Albaladejo, eds. Agile Processes in Software Engineering and Extreme Programming. Madrid: Springer Berlin Heidelberg: 324–325.

[48] Kim, E. & Ryoo, S. 2012. Agile Adoption Story from NHN. In 2012 IEEE 36th Annual Computer Software and Applications Conference. IEEE: 476–481. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6340199.

[49] Ihme, T. 2013. Scrum adoption and architectural extensions in developing new service applications of large financial IT systems. Journal of the Brazilian Computer Society, 19(3): 257–274. http://link.springer.com/10.1007/s13173-012-0096-0.

[50] Bayer, J. & Melone, N. 1989. A critique of diffusion theory as a managerial framework for understanding adoption of software engineering innovations. Journal of Systems and Software, 9(2): 161–166.

[51] Jeyaraj, A. & Sabherwal, R. 2008. Adoption of information systems innovations by individuals: A study of processes involving contextual, adopter, and influencer actions. Information and Organization, 18(3): 205–234.

[52] Przybyłek, A. & Zakrzewski, M. 2018. Adopting Collaborative Games into Agile Requirements Engineering. In Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2018), 54-64.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 959–968
DOI: 10.15439/2018F85
ISSN 2300-5963 ACSIS, Vol. 15

Assessing Effectiveness of Recommendations to Requirements-Related Problems through Interviews with Experts

Aleksander Jarzębowicz, Wojciech Ślesiński
Department of Software Engineering, Faculty of Electronics, Telecommunications and Informatics,
Gdańsk University of Technology, Narutowicza 11/12, 80-233, Gdańsk, Poland
Email: olek@eti.pg.edu.pl, slesinski.wojciech@gmail.com

***Abstract*—Requirements Engineering and Business Analysis are known as very important to software project outcome but also difficult activities, coping with many problems and challenges. The work reported in this paper was preceded by a survey which revealed most common requirements-related problems in Polish IT industry. We addressed ten most frequently reported problems by reviewing the literature for recommendations how to cope with those problems. The resulting set of recommendations is included in the paper. Next, we conducted interviews with three experienced IT analysts asking them to assess effectiveness of particular recommendations, based on their experience. The results show significant differences in assessments and indicate that effectiveness is dependent on contextual factors to a large extent. Our conclusion is that a follow-up work is required to document more recommendations and to annotate them with guidelines about applicability, intended context of use and possible pitfalls.**

I. Introduction

REQUIREMENTS are an essential element of virtually any software project. Regardless of which term is used to describe requirements-related activities – Requirements Engineering (RE) or a more recent, broader term of Business Analysis (BA) - delivery of an effective IT solution depends on identifying and managing business goals and stakeholders' needs. Negligence and flaws in RE/BA processes too often result in unpleasant consequences affecting the whole software project, including its failure and cancellation [1]–[3].

RE/BA is known as a difficult software project area [4]–[6] and as such is a subject of surveys aimed at identification of the most frequent and/or severe problems encountered by practitioners dealing with requirements e.g. [7]–[9]. Also, knowledge about such problems is disseminated by reporting experience [10]. To a large extent, RE/BA problems are therefore known and described in literature since a long time [11]. Yet, they are still present in industrial practice and recognized as causes of project failures even by very recent studies [3]. It raises a question what counter-measures exist to address such problems and how effective they are.

We intended to answer these questions in the context of IT industry in Poland. The first step was to identify relevant RE/BA problems to be addressed. Despite the fact that a number of surveys were conducted in several countries [7]–[9], it is hard to assume the results would be the same for e.g. different countries or different software development approaches – in fact differences are reported [9] and for that reason surveys on RE/BA problems are replicated in different settings [3].

In 2017 we conducted a survey among IT analysts from Poland to identify RE/BA problems which were most frequently encountered in their professional experience [12]. It provided us with a starting point to the research described in this paper, guided by the following research questions:

- RQ1: What recommendations are known to remedy most frequent requirements-related problems of Polish IT industry?
- RQ2: What is the effectiveness of applying such recommendations in practice?

To answer these questions, first we searched for recommendations by reviewing books on RE/BA and other sources reporting on industrial practice. For each of the top problems from the survey, we found several recommendations about how to deal with them and we documented them. Next, we interviewed 3 experienced analysts asking them to select recommendations they had applied in response to particular problems and to assess whether such recommendations turned out to be effective or not according to their experience.

The main contribution of this paper is the set of recommendations concerning 10 requirements-related problems, provided with assessments of their effectiveness from interviewed experts.

The remainder of the paper is structured as follows. In Section II we review the related work. Section III briefly describes the survey and its results, which provide input to the research study reported here. The study is presented in Sections IV-VI which cover: the search for recommendations and its results (IV), assessment of recommendations effectiveness through interviews with experts (V) and discussion of results and their validity (VI). The paper is concluded in Section VII.

II. RELATED WORK

In our study we address requirements-related problems by interviewing RE/BA experts. Interviews are dedicated to the assessment of the effectiveness of recommendations on how to mitigate these problems. Below we outline other works situated in the RE/BA domain, concerning recommendations to problems or good practices to be followed, identified on the basis of industrial experiences.

Several studies on addressing known RE/BA problems were reported. El Emam and Madhavji [13] interviewed practitioners, which resulted in identifying seven key issues of greatest concern in RE practice at the time and issuing recommendations for improving RE processes w.r.t. these issues. Bjarnason et al. [14] reported a case study of large software development company transitioning towards agile processes. Its employees were asked whether particular agile RE practices mitigate known problems of traditional RE and what new challenges those practices can introduce. De Oliveira Neto et al. [15] looked into the practices influencing both RE and system testing. They used knowledge gathered from interviews with practitioners to map the practices to the challenges encountered in large-scale agile development. Alsaqaf et al. [16] investigated to what extent and how are the non-functional requirements (NFR) included in agile large-scale distributed development projects. They interviewed practitioners to learn about existing practices as well as associated challenges with the ultimate goal to develop a set of good practices concerning NFRs in that kind of projects.

Recommendations and good practices can also be identified from experience of practitioners, without mapping them to the problems. Hickey and Davis [17] interviewed 9 well-known RE experts to determine which requirements elicitation techniques they would select in various situational characteristics of software projects. As a result, recommendations about the applicability of particular elicitation techniques were derived. Cao and Ramesh [18] focused on agile RE practices and used interviews, observations and documentation reviews to learn about implementations of 7 agile RE practices and associated benefits and challenges. Paul and Tan [19] conducted interviews with expert analysts to gather their opinions about the role of business analyst, his/her contribution to software project success and essential skills.

Our aim was to address particular problems identified in our survey conducted in Poland [12]. Problems, challenges and practices described in the abovementioned sources differed from our case, so no results could be used directly and a dedicated study had to be conducted.

TABLE I.
TOP REQUIREMENTS-RELATED PROBLEMS ACCORDING TO SURVEY RESULTS

ID	Problem	Description	Mean value
P1	Unrealistic expectations of stakeholders	A stakeholder has unrealistic expectations (e.g. large functionality to be delivered in a very short time or at low cost)	2,95
P2	"Obvious" requirements not communicated	Some requirements are so obvious to stakeholders that they do not even mention them	2,80
P3	Scope creep	Constant changes to project scope, modified or expanded requirements (scope creep)	2,80
P4	Too short time for analysis	Too short time for analysis is planned in project schedule	2,75
P5	Stakeholders' low availability	A stakeholder has little time available to commit to the project	2,73
P6	Solutions issued instead of requirements	Stakeholders describe solutions (e.g. detailed UI design) instead of requirements	2,62
P7	Requirements in the form of change requests only	Stakeholders are not able to express their requirements, they express them later as change requests to working software	2,56
P8	Conflicting requirements	Different stakeholders have mutually contradicting (conflicting) requirements	2,53
P9	Stakeholders ignore business goals	Stakeholders focus on requirements only, they do not define business goals or do not consider them important	2,49
P10	Requirements exceeding project scope	Stakeholders define requirements which are clearly outside project's scope established earlier	2,44

III. SURVEY AND ITS RESULTS

This section briefly describes the survey and the top problems reported by respondents, which provide input to further research presented in the next sections.

The survey took place in Spring of 2017. We prepared a web-based questionnaire (in Polish), published it and invited respondents through websites and social network groups for IT analysts. We received 55 answers. As only analysts were targeted, we included a question about respondent's experience in RE/BA. The distribution of answers was as follows:

- 3 (6%) – less than 1 year;
- 9 (16%) – over 1 but less than 2 years;
- 21 (38%) – over 2 but less than 5 years;

- 14 (25%) – over 5 but less than 10 years;
- 8 (15%) – over 10 years.

The questionnaire included 64 RE/BA problems gathered from the review of literature and the direct communication with a number of analysts [12]. For each problem, the survey respondent was supposed to assess how frequent such problem had been encountered in his/her work experience. The following answers were available: never (0); rarely (1); sometimes (2); often (3); always (4). The numbers in parentheses indicate numerical values that were used when processing and analyzing the survey results. We used them to calculate metrics, which represent frequency of occurrence of each problem and provide a basis for a ranking. Despite the fact that ordinal scale was used to represent answers, we calculate means values, not medians, as for a 5-point scale it is unlikely to spot differences comparing medians.

Table I gives a summary of survey results – 10 problems assessed as most frequent by survey respondents. For each problem, the table includes its artificial ID (to be used as reference in the remainder of the paper), short name of a problem, a longer description (used in the survey questionnaire) and the calculated mean value of answers, which is used as a metric representing frequency.

IV. RECOMMENDATIONS

The search for recommendations addressing the problems listed in Table I included books dedicated to RE: [20]–[23]. We consider them representative, as both internationally recognized items and positions limited to Polish readers are included. Most of them were published in recent years, except the book by Leffingwell and Widrig, which we included due to its influence (number of citations). Additionally, the training materials issued by RE/BA professional associations as well as the on-line courses were reviewed w.r.t. the potential solutions to the abovementioned problems.

Below, we present recommendations we were able to find, divided into sub-sections dedicated to particular problems. Descriptions of the recommendations were shortened and unified when multiple sources mentioned a similar way of dealing with a given problem. Recommendations are assigned the identifiers used in the remainder of the paper. The identifier Rx.y denotes a recommendation number y to a problem number x.

A. P1 - Unrealistic expectations of stakeholders

Two major sources of this problem can be identified and, thus, two groups of recommendations are distinguished.

Unrealistic expectations can stem from stakeholders' attitude to issue all requirements they can think of, even those unfounded by business needs or concerning very distant (and uncertain) future. In such case, expectations can be toned down by reducing the number and scope of requirements by the following actions:

- R1.1: Identify business goals as a reference point for requirements.
- R1.2: Identify what does the customer actually expects after the product is deployed (it may not be explicitly articulated without guiding questions).
- R1.3: Verify quality of issued requirements (rationale, unambiguity, feasibility) and do not proceed with requirements of low quality.
- R1.4: Conduct requirements analysis and verify how particular requirements support business goals.

Another possibility is that requirements are consistent with business needs, but their scope is not adjusted to the project constraints. The point of view of at least some of the customers can be to get as much as possible, as fast as possible and at minimal cost, which is a clear violation of the "iron triangle" [24] constraints and is not feasible in real-life projects. To mitigate such situation it is recommended to:

- R1.5: Precisely define product's scope agreed between the supplier and the customer.
- R1.6: Develop a realistic project schedule and manage it during the whole development process.
- R1.7: Estimate development costs (including all relevant categories) and monitor expenses to keep the project within its budget.
- R1.8: Continuously monitor requirements' statuses to be aware which ones are satisfied by the current version of the system under development.

B. P2 - "Obvious" requirements not communicated

Stakeholders can assume that some requirements are so obvious that they do not even have to mention them. Unfortunately, it leads to incomplete requirements. Especially non-functional requirements can be omitted this way. Missing requirements are harder to detect than e.g. conflicting ones and thus they can be identified late e.g. during acceptance tests. To prevent this:

- R2.1: Research the literature on the problem domain as well as the similar existing software systems (even before eliciting requirements from human stakeholders).
- R2.2: Make sure that all stakeholders and points of view relevant to the developed system are identified (using e.g. stakeholder map or onion model techniques).
- R2.3: Involve technical experts (on e.g. security, usability) into requirements elicitation and analysis processes.
- R2.4: Use appropriate requirements elicitation techniques, preferably a combination of several techniques including group work (e.g. workshops, focus groups).
- R2.5: When specifying requirements, use Software/ System Requirements Specification templates, which include the sections covering various categories of requirements, goals and constraints.

- R2.6: Apply requirements analysis techniques (e.g. checklists, CRUD tables) aimed at the identification of inconsistent, incomplete or missing requirements.

C. P3 - Scope creep

Requirement changes are to be expected in virtually any project, but continuous and uncontrolled changes will probably lead to scope creep. Such situation is considered harmful, especially when such changes are avoidable (e.g. result from communication flaws). Moreover, applying a change usually requires additional resources, but it is often not recognized by the customer. The following recommendations address the scope creep problem:

- R3.1: Include a "buffer" – some surplus when estimating project resources (budget and schedule) – it allows (to some extent) adjusting to changes. It is especially advised when no prior experience can be used as a reliable estimation base.
- R3.2: Define project scope from the beginning. This task must involve key stakeholders and result in defining, agreeing, confirming and documenting high-level requirements: business goals, product vision, product scope and main constraints. Requirements which are known but will not be implemented within the current project should also be documented and assigned appropriate statuses.
- R3.3: Change request issued by the customer should be processed by the change control process. Its first step is the impact assessment. Apart from analyzing influence on other requirements, it should include: verification of compliance to business goals and estimation of required resources.
- R3.4: Use requirements priorities assigned by customer representatives to distinguish essential and secondary requirements.
- R3.5: Include requirements sign-off activity in a project (i.e. the customer should confirm in writing that documented requirements are valid).
- R3.6: Approved scope changes should be immediately communicated to all stakeholders.
- R3.7: Approved scope changes should result in adjusting the project's budget and schedule.
- R3.8: Apply prototyping technique to prevent changes stemming from misunderstandings between stakeholders and developers (as mockups and other throw-away prototypes are relatively cheap to develop).
- R3.9: If impact analysis reveals change proposal to be harmful (e.g. impossible to implement within given constraints, not consistent with the business goals) – confront the customer with it.

D. P4 - Too short time for analysis

Reduced time available for analysis can be just a planning flaw, but often it is rather a result of the customer's pressure to produce software as soon as possible and reduce or skip all activities not directly resulting in code development. The customer can e.g. refuse to pay for RE/BA activities or insist on planning them very short and/or only at the very beginning of the project. It is very likely that such attitude would lead to misunderstandings, rework, delays, and increased costs, so it should be prevented by e.g.:

- R4.1: The most essential thing is to raise customer's awareness. The customer often has a limited knowledge about software project organization and does not understand the importance of the RE/BA activities. Information about benefits of well conducted RE/BA activities as well as risks caused by reducing them should be provided to the customer.
- R4.2: It is important to build an atmosphere of trust in customer-supplier cooperation. The customer should be convinced that activities and techniques used in RE/BA are selected because of their contribution to the project success and not in order to increase costs.
- R4.3: Negative examples can be used to convince the customer e.g. cases where invalid requirements led to significant rework effort and related costs.

E. P5 - Stakeholders' low availability

A stakeholder is likely to be busy with his/her everyday work duties and thus unable to contribute much time to the project e.g. to be interviewed by the analyst or to validate the specified requirements. Stakeholders are however indispensable for the acquisition of the domain knowledge as well as the elicitation and validation of the requirements. In order to win their commitment and to minimize their additional effort, the following practices can be considered:

- R5.1: At the beginning of the project present the customer with the RE/BA plan. The plan should define the activities and their phases as well as clearly indicate what kind of stakeholders' involvement is required and when.
- R5.2: When the representatives of key stakeholders are already identified, inform the customer's executives about the need to add tasks related to the participation in RE/BA to the representatives' responsibilities and schedules.
- R5.3: When the information from human stakeholder is to be obtained, a direct communication or video-conference is preferred, as other communication means would probably prolong this process. An analyst should however strive to minimize disruptions to the customer organization's processes and to the stakeholders' tasks by appropriate planning of the meetings.
- R5.4: Each meeting between an analyst and the stakeholder(s) should have clearly assigned participants and goal (when it is achieved, the meeting ends). Also the maximum duration of the meeting should be planned ahead.

- R5.5: Number of the meeting participants should be limited because otherwise the discussion is difficult to control (because of subgroups emerging, some participants getting frustrated with topics they do not comprehend or find not interesting to them etc.).

F. P6 - Solutions issued instead of requirements

RE/BA should be focused on identifying requirements which are then processed to design the best solutions (within given constraints) that fulfil them. It is however quite typical that what a stakeholder describes is neither a requirement nor a constraint, but a particular solution (which may not be optimal at all, considering the bigger picture). The following recommendations were identified to deal with such problem:

- R6.1: When information obtained from a stakeholder turns out to be a solution instead of a requirement, it should not be ignored as it is a valuable lead for further inquiry. An analyst should note it down, assign it to a separate category (other than requirements) and use in further interviews as a means to steer the discussion.
- R6.2: A particular solution described by a stakeholder may indicate that he/she already knows a similar software (perhaps a legacy or competitor system). An analyst should study such system and review its features with stakeholders, trying to derive requirements. Missing requirements can be identified as a side effect.
- R6.3: Stakeholders often prefer to discuss solutions because they perceive it as something more concrete and easier to comprehend. It is important for an analyst to have a sufficient knowledge about problem domain (acquired earlier) and use it when interviewing a stakeholder to generalize discussed concept and express it as a requirement.
- R6.4: A feasible way to discover an actual requirement behind a design solution is to ask „why" questions i.e. request a stakeholder to provide the rationale behind that particular solution he/she describes.
- R6.5: To derive a requirement from a design solution, an analyst has to paraphrase and generalize the input received from a stakeholder. Applying a combination of requirement elicitation techniques is beneficial, because more diversified information can be obtained this way and provide a better basis for derivation of requirements.

G. P7 - Requirements in the form of change requests only

Requirements elicitation is usually a difficult task as stakeholders can have problems with articulating their needs or even realizing them. It leads to a situation when stakeholders "don't know what they want, but know what they don't want when presented with it" and issue change requests to the published release of a developed system. To prevent it, the following practices can be considered:

- R7.1: When human stakeholders turn out to be uncooperative, requirements can be elicited (to some extent) from other sources through e.g. document analysis - reviewing organizational charts, procedures, existing process models, memos etc.
- R7.2: The main scope of the developed system's functionality and other essential features are probably included in the contract document, which can be used as a starting point for requirements elicitation.
- R7.3: If the system under development is to support existing or modified business processes in the customer organization, observation technique can be used to allow the analyst to gather the information firsthand by witnessing the work activities.
- R7.4: Existing IT systems used in the customer's organization can be a valuable source of requirements. Moreover, discussing obsolete, missing and suboptimal functions of the existing system with the stakeholders can be much more effective than just asking them about their needs.
- R7.5: After identifying initial requirements, an analyst should facilitate a workshop to present and discuss them with a group of stakeholders. Despite being rather generic and not validated, such requirements can constitute the basis for discussions. It is essential to present the requirements in a communicative way e.g. by using prototypes. A facilitated discussion during which the stakeholders can refer to the initial requirements can enable them to communicate requirements they were unable to articulate earlier.

H. P8 - Conflicting requirements

Requirements elicitation is followed by requirements analysis which is likely to uncover issues like ambiguous, incomplete and, last but not least, conflicting requirements. Conflicting requirements may be attributed to mistakes, but more likely they are the result of different points of view, needs and expectations of various stakeholders and, as such, should be properly managed:

- R8.1: It is worth to identify an authorized customer's representative, who would be capable of resolving conflicts in case stakeholders cannot do it themselves.
- R8.2: When conflicting requirements occur, the first step should be to verify their consistency with the business goals and project scope. The requirements that fail such a test can be revealed as unfounded and, as a result, the conflict may be solved.
- R8.3: An analyst needs to acquire domain knowledge to be able to determine the degree to which particular requirements contribute to the achievement of the business goals and to the project success. It can be used as a criterion to choose one option from the conflicting ones.
- R8.4: When conflicting requirements occur, this matter should be discussed at a meeting with all involved

stakeholders. The analyst should moderate the discussion and aim at explicitly presenting the trade-offs, suggesting ways of resolving conflicts and reaching the consensus.

- R8.5: Two or more conflicting requirements (concerning e.g. a function) can be replaced with a single one (more generic or more complex) that addresses the goals/rationales of each of them.
- R8.6: If substantial effort has been made and still stakeholders cannot (or are unwilling to) reach consensus, such issue can be delegated to a decisive entity e.g. a project sponsor or the senior management of the customer organization.

I. P9 - Stakeholders ignore business goals

Business goals defined by authorized customer representatives are the very reason behind customer's decision to acquire an IT system and start a software project. Also a primary indicator of the project success is whether the developed software allowed to achieve the business goals. If business goals are not precisely specified or even completely ignored, then there is no point of reference to determine which requirements contribute to the business benefits of the customer organization. It is thus advised that:

- R9.1: An analyst should consult key representatives of customer organization and insist on defining business goals before eliciting lower level requirements. Business goals (specified in precise and unambiguous way) should then be disseminated among all parties involved, both from supplier and customer sides.
- R9.2: Business goals should be maintained during the whole project schedule. If business goals change (due to e.g. market situation), all parties involved in software project should be notified and requirements need to be reviewed w.r.t. the new goals.
- R9.3: Even if substantial effort was made to identify business goals, it is possible that some goals were omitted. When dealing with a requirement that cannot be associated with any goal, a possibility that the set of goals is incomplete should be considered. It may lead to defining an additional goal.

J. P10 - Requirements exceeding project's scope

When project scope is agreed between the customer and the supplier, usually the associated project constraints (budget, schedule) are defined in accordance with the scope. A situation that stakeholders define requirements exceeding this scope is dangerous, because it will likely also result in project not meeting its constraints. It may however be caused by invalid scope assumed at the beginning of the project. To avoid the problem, the following actions can be taken:

- R10.1: It is important to ensure that the definition of the project scope involves several stakeholders, selected on the basis of their domain knowledge and authority. It should prevent invalid scope definition.
- R10.2: Change management process needs to be defined and followed. Such process should treat the cases of requirements exceeding the scope as changes to the project scope. Impact analysis should be conducted for such changes and include the analysis of the change's influence on schedule and budget.
- R10.3: The agreed project scope should be documented and formally approved by an authorized customer representatives. An analyst can refer to such a document when particular stakeholders insist on adding requirements which surpass that scope.

V. Interviews with experts

After collecting the recommendations, the next step was to assess their effectiveness. This assessment was supposed to be based on real-life experiences, not speculations. We decided to interview experts for this purpose. We selected 3 analysts whom we considered to be experts in RE/BA domain, with regard to their knowledge (confirmed by trainings completed and certificates obtained) and professional experience (work history including several companies and projects from various business domains). Their brief characteristics are given below:

- Expert A – 6 years of experience as an analyst (11 years in IT in total). Software projects from the following business domains: transportation, logistics, medical, financial. PhD in computer science, REQB FL certificate holder.
- Expert B – 8 years of experience as an analyst (10 years in IT in total). Software projects from the following business domains: public administration, telecommunications, transportation. IIBA CBAP certificate holder. Active participation in RE/BA professional association (local chapter board member).
- Expert C – 15 years of experience as an analyst. Software projects from the following business domains: insurance, public administration, government, electronics and telemetry. REQB FL certificate holder. Job history including middle- and high-level management positions and the role of coach in several commercial RE/BA courses.

Each expert's task was to review the presented recommendations and provide his/her feedback. For each problem, an expert had to answer the following questions:

- Have you encountered this problem in your work experience as an analyst?
- Which of the listed recommendations have you tried to address this problem?
- Which among the applied recommendations turned out to be effective?
- Which among the applied recommendations failed to mitigate the problem?

TABLE II.
ASSESSMENT RESULTS FOR RECOMMENDATIONS TO PROBLEMS P1-P5

Problem	Recommen-dation	A	B	C	Remarks
P1	R1.1	Y	U	Y	A: The recommendations proved effective but only when used altogether and not selectively e.g. the identification of the business goals only. B: The communication and education of the customer is the key (none of the recommendations would succeed if used without it). It is also useful to define the Minimum Viable Product and other versions with enhanced functionality, estimate the cost of each version. C: If possible, try to simplify the solution (e.g. scope of functionality) in a way not compromising the business goals.
	R1.2	Y	U	Y	
	R1.3	Y	U	X	
	R1.4	Y	U	U	
	R1.5	Y	U	Y	
	R1.6	Y	U	N	
	R1.7	U	U	N	
	R1.8	U	U	X	
P2	R2.1	U	Y	X	A: Interviews (unstructured, including apparently obvious matters) and workshops are the most effective elicitation techniques to address this problem. B: R2.1 is a reasonable recommendation in general, but an analyst should be aware that sometimes it may fail (a specific customer deviating from standard processes known in problem domain, an existing system which is poorly tailored to the needs). B: R2.5 requires good understanding of the template contents - if a template is used with an attitude to leave no section empty and there is no focus on quality of the content, then nothing good comes out of it. B: An additional way to address this problem is the prototyping which often reveals missing "obvious" requirements. C: It is the requirements validation rather than the requirements analysis (R2.4) that can reveal hidden requirements. C: Additional recommendations: 1. Requirements validation through a dedicated meeting; 2. Business process modelling with mapping between the processes and the requirements.
	R2.2	U	Y	Y	
	R2.3	Y	U	Y	
	R2.4	Y	Y	Y	
	R2.5	U	U	X	
	R2.6	U	Y	N	
P3	R3.1	U	U	N	A: Minimum Viable Product concept is also useful here. B: Presented recommendations are good ideas, but none of them guarantees the scope creep prevention e.g. a buffer (R3.1) may be insufficient, some customer representatives do not feel obliged by sign-off and reject earlier agreements etc. B: Other recommendations could be: 1. accept small changes (up to a specific effort estimation) and move bigger ones to future releases; 2. develop iteratively in time & material mode; 3. acquire knowledge about problem domain and customer organization (structure, processes) to get better understanding of the possible scope definitions and avoid surprises.
	R3.2	Y	U	Y	
	R3.3	Y	U	Y	
	R3.4	Y	U	X	
	R3.5	U	U	N	
	R3.6	U	U	X	
	R3.7	U	U	X	
	R3.8	U	U	Y	
	R3.9	U	U	Y	
P4	R4.1	U	U	Y	B: All 3 recommendations are good ideas but some customers still just say "no" and refuse to listen to any arguments B: Other ideas: 1. make project phases/activities non-negotiable, the project should be "sold" to the customer only as a whole; 2. develop iteratively and demonstrate the results (analysis will be divided into many smaller tasks spread over time and the customer will get partial result earlier); 3. try to negotiate stakeholders' commitment e.g. analysis will be shorter if a given stakeholder joins the development team for a week. C: It is crucial to develop the RE/BA plan (consistent with overall project plan and customer's expectations) which should specify who is to be involved from customer's side, in which activities and why such involvement is necessary.
	R4.2	Y	U	Y	
	R4.3	U	U	N	
P5	R5.1	N	U	Y	A: Stakeholders' tasks and expected input have to be precisely defined. B: Again, recommendations are worth trying, but if the low availability is not due to the lack of awareness but e.g. heavy workload then none of suggested actions will change that. B: Additional recommendations: 1. Escalation - ensure stakeholders' availability by talking to their superiors (executives, project sponsor etc.); 2. Include the specific clauses about maximum response time and minimal effort required from customer's side in the project contract.
	R5.2	U	U	X	
	R5.3	Y	U	Y	
	R5.4	Y	U	Y	
	R5.5	U	U	Y	

The summarized results of the assessments are presented in Tables II and III. Both tables have the same columns: problem ID; associated recommendations IDs; assessments by experts A, B and C; additional remarks made by experts.

The following symbols are used in the table to denote the assessment results:

- X – Recommendation was not used by an expert to mitigate a given problem;

- Y – (Yes), recommendation was used and proved effective;
- N – (No), recommendation was used and failed to mitigate the problem;
- U – (Uncertain), recommendation was used but it is not possible to determine its exact contribution to the problem-solving or the outcome of applying the recommendation differed from project to project;

In addition, the colors are used to distinguish the assessment results. Only expert A denied experiencing some of the enumerated problems (P7, P9 and P10), thus X values are used for all associated recommendations.

TABLE III.
ASSESSMENT RESULTS FOR RECOMMENDATIONS TO PROBLEMS P6-P10

Problem	Recommen-dation	A	B	C	Remarks
P6	R6.1	U	U	N	A: It is not recommended to review an existing system with a stakeholder, because it results in closing his/her mind to the options other than present in that system. B: Some people are stubborn and resistant to any reasoning. They will most likely insist on a particular solution even when presented with strong arguments against it. C: Education of stakeholders, so they know what a requirement is and how to express it - such issue should be brought to the main customer's representative.
	R6.2	N	U	Y	
	R6.3	U	U	Y	
	R6.4	U	U	N	
	R6.5	Y	U	N	
P7	R7.1	X	U	Y	B: R7.2 does not make much sense - if stakeholders cannot articulate their needs, how to determine the scope/features to be written in the contract? B: Alternative sources of requirements (documents, observation, existing systems) may not provide all the necessary information or be unavailable (e.g. no existing IT system used, new business processes not implemented yet and impossible to observe). B: Prototyping or incremental development can be effective as the cost of changes is lower. C: Additional recommendations: 1. Business process modelling with mapping between the processes and the requirements; 2. Specify the stakeholders' points of view.
	R7.2	X	N	N	
	R7.3	X	U	N	
	R7.4	X	U	Y	
	R7.5	X	Y	Y	
P8	R8.1	X	U	X	B: Relying on an authorized representative is a good idea, provided that such person is willing to take responsibility and make difficult decisions. B: Some stakeholders may insist on "their" requirements regardless of the consistency with the business goals. B: Additional recommendation: use multi-criteria assessment (including priority, difficulty, cost, conformance to standards etc.) and apply it to the conflicting requirements to choose the optimal one (w.r.t. those criteria).
	R8.2	X	U	Y	
	R8.3	X	U	N	
	R8.4	Y	Y	Y	
	R8.5	Y	U	X	
	R8.6	Y	U	Y	
P9	R9.1	X	U	Y	B: If stakeholders refuse to define the goals, no one can force them or do it on their behalf. B: Additional recommendations: 1. Educate the stakeholders about the importance of business goals and the risk of ignoring them; 2. Try to define business goals on the basis of elicited requirements and ask the stakeholders to validate them (but be aware of a risk that they will confirm the goals without actually considering them). C: Specify the stakeholders' points of view.
	R9.2	X	U	Y	
	R9.3	X	U	N	
P10	R10.1	X	U	N	B: Additional recommendation: A workshop during which the stakeholders associate their requirements to business goals/high-level requirements. C: Specify the stakeholders' points of view. Constantly monitor the scope.
	R10.2	X	Y	Y	
	R10.3	X	U	Y	

VI. DISCUSSION

A. Discussion of results

Recommendations found cover all the problems we intended to address and there was no single case that a given problem was not recognized by literature or without any suggestions how to cope with it. The list of recommendations we assembled is however far from being complete, as analysts we interviewed reported several additional remedies to these problems from their experience. Keeping in mind that it was a very small scale study (3 interviewees), we should expect more recommendations if a larger group of analysts were involved.

The recommendations differ according to their nature: some are just choices of a particular technique (for e.g. requirements elicitation), some are rather related to the processes and organizational issues (e.g. planning/ monitoring activities, participation of particular people), while other concern cooperation and relationships (e.g. educating stakeholders, atmosphere of trust). Quite often, multiple recommendations benefit from being used together.

Assessments of the recommendations' effectiveness collected from the interviews are inconclusive at best.

Knowing the complexity of the RE/BA processes (and the software development in general) as well as the differences between the organizations, projects and business domains, we did not expect the same answers from each interviewee. However, differences in their assessments are greater than expected - there are literally two cases in which all 3 experts made the same assessment. To some extent it can be attributed to differences in experts' perception/attitude. It is clearly seen that expert B avoided giving definite answers (very few Y or N, mostly U assessments). This person was however the most active in providing additional remarks about e.g. factors influencing recommendations' effectiveness or situations that they would not work. The remaining two experts were more willing to summarize their experiences in a yes/no answer.

We considered this interview study as an initial validation of the recommendations found. Results of the study clearly indicate that it is better to be careful with using recommendations, as the outcome can differ in various domains, projects and other settings. Thus, such a list as the one assembled by us has a limited use, as it is neither complete nor provides enough guidelines about the context a given recommendation should or should not be used. We see two possible improvements. The first is to extend the list with additional recommendations identified through the literature reviews, interviews, surveys etc. Such list can grow and become more difficult to use, but it seems necessary as apparently no problem (at least not those listed in Table I) can be mitigated by a single action with a 100% effectiveness. The second improvement is to annotate the recommendations with guidelines about their applicability, limitations, trade-offs etc. (perhaps in a similar way as for the design patterns).

Another issue worth addressing in future is to consider problems and recommendations in a wider context. RE/BA does not exist in isolation but is a part of an overall software development process and strongly influences e.g.: software architecture [25] or testing (also verification & validation in general) [26]. When developing and presenting recommendations, it would be worthwhile to take into account contextual factors and consequences to software project activities and artefacts other than just those directly related to RE/BA.

B. Discussion of validity

We are aware of several limitations of our study. The number of interviewees is the most important one, as small scale studies cannot be considered as very convincing. A larger group of experts would make results more valid.

Another issue is the competence and representativeness of the interviewees. Knowledge and experience were applied as criteria for interviewee selection. Moreover, the diversity of projects and business domains each of them worked for is an argument for their credibility. We did not analyze their representativeness among the general population of analysts though, which is a possible validity threat, especially if results were to be generalized as applicable worldwide.

The quality of input data should be considered. In this case, such input was the list of recommendations presented to interviewees. Although an effort was made to assemble it on the basis of multiple sources, it certainly was not exhaustive, but that was not necessary – interviewees were only supposed to assess each of the presented recommendations and optionally report others known to them.

A common validity threat to interview-based research is the honesty of interviewees. We mitigated this threat by arranging interviews with volunteers only and assuring that they would remain anonymous. Subjectivity is also an inevitable aspect of interviews – despite our instructions to give assessments on experience basis, no strict, quantitative criteria were defined and interviewees could differ in their judgements. For example, a recommendation which was successfully applied in many cases but proved unsuccessful once could be classified as effective (Y) by one expert and uncertain (U) by another.

Finally, researchers can be the source of validity threats too by their possible bias regarding e.g. conviction about the effectiveness of particular recommendations. We addressed the first issue by providing interviewees problems and recommendations in a written form, without any additional suggestions and by interpreting the outcome separately by 2 researchers and then comparing their opinions.

VII. CONCLUSION

In this paper we reported a research study aimed at identifying and assessing known recommendations that address top requirements-related problems in the IT industry in Poland. Answers to the research questions formulated in the introductory section were obtained through reviewing relevant sources (RQ1) and interviewing experienced analysts (RQ2). These answers were presented in Section IV – Recommendations (RQ1) and in Tables II-III (RQ2).

The insight gathered from the study can be briefly summarized as follows. There are many recommendations available and described in various sources. They can be considered accurate in general – there was not a case of our experts rejecting a recommendation unanimously as inapplicable or invalid (maybe with the exception of R7.2). Recommendations are, however, not the "silver bullets" that can be used in any context with a guarantee of success. Experience-based assessments made by a small number of experts proved to be very diverse, which, on the basis of the interviews, we attribute to the contextual differences (domain, customer and supplier profiles, project organization etc.). Such applicability context is not necessarily specified as part of the documented recommendations.

The possible future research include: identifying a larger set of recommendations (by e.g. literature reviews, surveys, interviews), specifying their applicability context and other constraints (if such issues are not provided together with a

recommendation itself, they can be established by surveying/ interviewing practitioners), and making additional assessments of recommendations' effectiveness (in a more specific context, by a larger group of experts). The possible outcome is a sort of "catalogue" of recommendations that practitioners could review and select recommendations to apply in their project as a response to encountered problems.

ACKNOWLEDGMENT

We are grateful to the interviewed experts for their participation and sharing their experiences.

REFERENCES

[1] R. N. Charette, "Why Software Fails", *IEEE Spectrum*, vol. 42, no. 9, pp. 42–49, 2005, https://doi.org/10.1109/mspec.2005.1502528

[2] J. McManus and T. Wood-Harper, "Understanding the Sources of Information Systems Project Failure - A study in IS project failure", *Manag. Serv.*, vol. 51, no. 3, pp. 38–43, 2007.

[3] D. Mendez Fernández et al., "Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice", *Empir. Softw. Eng.*, vol. 22, no. 5, pp. 2298–2338, 2017, https://doi.org/10.1007/s10664-016-9451-7

[4] B. H. C. Cheng and J. M. Atlee, "Research Directions in Requirements Engineering", *Proceeding FOSE '07 2007 Futur. Softw. Eng.*, pp. 285–303, 2007.https://doi.org/10.1109/fose.2007.17

[5] A. Przybyłek, "A Business-oriented Approach to Requirements Elicitation", in *9th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'14)*, Lisbon, Portugal, pp. 152-163, 2014. https://doi.org/10.5220/0004887701520163

[6] A. Przybyłek and M. Zakrzewski, "Adopting Collaborative Games into Agile Requirements Engineering", *in 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'18)*, Funchal, Madeira, Portugal, pp. 54-64, 2018, https://doi.org/10.5220/0006681900540064

[7] T. Hall, S. Beecham and A. Rainer, "Requirements problems in twelve software companies: an empirical analysis", *IEE Proc. - Softw.*, vol. 149, no. 5, p. 153, 2002, https://doi.org/10.1049/ip-sen:20020694.

[8] N. K. Sethia and A. S. Pillai, "A study on the software requirements elicitation issues - its causes and effects", *2013 Third World Congr. Inf. Commun. Technol. (WICT 2013)*, pp. 245–252, 2013, https://doi.org/10.1109/wict.2013.7113143

[9] D. Mendez Fernandez et al., "Naming the Pain in Requirements Engineering: Comparing Practices in Brazil and Germany", *IEEE Softw.*, vol. 32, no. 5, pp. 16–23, 2015, https://doi.org/10.1109/ms.2015.122

[10] D. Firesmith, "Common requirements problems, their negative consequences, and the industry best practices to help solve them", *J. Object Technol.*, vol. 6, no. 1, pp. 17–33, 2007, https://doi.org/10.5381/jot.2007.6.1.c2

[11] B. Davey and K. Parker, "Requirements elicitation problems: a literature analysis", *Issues Informing Sci. Inf. Technol.*, vol. 12, pp. 71–82, 2015, https://doi.org/10.28945/2137

[12] A. Jarzębowicz and W. Ślesiński, "What is Troubling IT Analysts? A Survey Report from Poland on Requirements-related Problems", in *Proc. of 20th KKIO Software Engineering Conference*, Advances in Intelligent Systems and Computing vol. 830, pp. 3-19, Springer, 2018, https://doi.org/10.1007/978-3-319-99617-2_1

[13] K. el Emam and N. H. Madhavji, "A field study of requirements engineering practices in information systems development", *International Conference on Requirements Enineering*, pp. 68–80, 1995, https://doi.org/10.1109/isre.1995.512547

[14] E. Bjarnason, K. Wnuk, and B. Regnell, "A case study on benefits and side-effects of agile practices in large-scale requirements engineering", *Proc. 1st Work. Agil. Requir. Eng. - AREW '11*, pp. 1–5, 2011, https://doi.org/10.1145/2068783.2068786

[15] F. G. De Oliveira Neto, J. Horkoff, E. Knauss, R. Kasauli, and G. Liebel, "Challenges of aligning requirements engineering and system testing in large-scale agile: A multiple case study" *Proc. 2017 IEEE 25th Int. Requir. Eng. Conf. Work. REW 2017*, p,p. 315–322, 2017, https://doi.org/10.1109/rew.2017.33

[16] W. Alsaqaf, M. Daneva, and R. Wieringa, "Quality requirements challenges in the context of large- scale distributed Agile : An empirical study", in *Proc. of 24th Requirements Engineering: Foundation for Software Quality Conference*, pp. 139-154, 2018, https://doi.org/10.1007/978-3-319-77243-1_9

[17] A. M. Hickey and A. M. Davis, "Elicitation technique selection: How do experts do it?", in *Proceedings of the IEEE International Conference on Requirements Engineering*, 2003, pp. 169–178. https://doi.org/10.1109/icre.2003.1232748

[18] L. Cao and B. Ramesh, "Agile requirements engineering practices: An empirical study", IEEE Softw., vol. 25, no. 1, pp. 60–67, 2008, https://doi.org/10.1109/ms.2008.1

[19] D. Paul and L. Y. Tan, "An Investigation Of The Role Of Business Analyst In IS Development", ECIS 2015 Proc., pp. 1–14, 2015.

[20] K. Wiegers and J. Beatty, "*Software Requirements"*, 3rd ed. Microsoft Press, 2013, ISBN: 978-0735679665.

[21] D. Leffingwell and D. Widrig, *Managing Software Requirements*, Pearson Education, 2003, ISBN:032112247X

[22] B. Chrabski and K. Zmitrowicz, *Requirements Engineering in Practice* (in Polish: Inżynieria Wymagań w Praktyce), Wydawnictwo Naukowe PWN, 2015, ISBN: 9788301180188

[23] M. Bartyzel, *Tailored software - how to speak to customers who don't know what they want (in Polish: Oprogramowanie szyte na miarę. Jak rozmawiać z klientem, który nie wie, czego chce)*. Wydawnictwo Helion, 2012, ISBN: 978-83-246-3932-8

[24] E. Bernroider and M. Ivanov, "IT project management control and the Control Objectives for IT and related Technology (CobiT) framework", *Int. J. Proj. Manag.*, vol. 29 no. 3, pp. 325-336, 2011, https://doi.org/10.1016/j.ijproman.2010.03.002

[25] J. Cleland-Huang, R. S. Hanmer, S. Supakkul, and M. Mirakhorli, "The Twin Peaks of Requirements and Architecture", *IEEE Softw.*, vol. 30, no. 2, pp.24-29, 2013, https://doi.org/10.1109/MS.2013.39

[26] E. Bjarnason et al., "Challenges and practices in aligning requirements with verification and validation: a case study of six companies", *Empir. Softw. Eng.*, vol. 19, no. 6, pp. 1809-1855, 2014, https://doi.org/10.1007/s10664-013-9263-y

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 969–973
DOI: 10.15439/2018F53
ISSN 2300-5963 ACSIS, Vol. 15

Agile to Lean Software Development Transformation: a Systematic Literature Review

Filip Kišš and Bruno Rossi
Faculty of Informatics
Masaryk University, Brno, Czech Republic
Email: {390917,brossi}@mail.muni.cz

Abstract*—*Context: **Lean development has been often proposed as an adaptation to agile for scaling-up to larger contexts.** ***Goals:*** **we wanted to better understand the "agile-to-lean" transformation, in terms of: i) reported benefits, ii) challenges faced, iii) metrics used.** ***Method:*** **we performed a Systematic Literature Review (SLR) about "agile-to-lean" transformations.** ***Results:*** **reduced lead time, improved flow, continuous improvement, and improved defect fix rate were the main reported benefits. Adaptation to lean thinking, teaching the lean mindset, identification of the concept of waste, and scaling flexibility were the main challenges. Lead time was the most reported metric.**

I. Introduction

NOWADAYS, many software organizations use agile methodologies for their software development processes, finding benefits for process improvement [1]–[3]. Lean software development [4] has been used to optimize development processes, mainly due to the concept of waste reduction involved in the optimization of all activities producing inefficiencies [4]. This view is complementary to agile principles, more focused on all activities that create value for the customer.

The term lean software development originated from the work of Mary and Tom Poppendieck [4]. Lean software development can be characterized as a combination of lean manufacturing with the lean IT principles and their application into software development [5]. This approach is driven by a series of seven principles: eliminate waste, decide as late as possible, amplify learning, deliver as fast as possible, empower the team, build integrity in, and see the whole [4], [5].

Many authors suggest that lean thinking can be used as guiding principles to implement and adopt agile development practices [4]. Wang [1] has identified and analyzed various combinations of lean and agile as reported by previous research: lean principles can either be used to adapt existing agile practices or to scale-up the agile software development practices [1]. However, there is no universal type of "agile-lean" combination that can be used for every situation [1].

The goal of this paper is to identify benefits, challenges, and metrics used in *"agile-to-lean"* transformations. Identification of such factors can allow to better understand *"agile-to-lean"* transformations, providing more evidence about how such transformation is happening.

II. SLR

A Systematic Literature Review (SLR) is a process which summarizes, organizes, and documents previous research of a field in a systematic way [6]. To conduct the review on "agile-to-lean" transformations, we followed the SLR guidelines by Kitchenham and Charters [6].

A. Needs for an SLR

"Agile-to-lean" is a less explored research area compared to "waterfall-to-agile" (e.g., Middleton [7]). There are other SLRs performed on lean methodologies but they are focused either at the business level [8], or at the level of metrics used in lean / agile software projects within industry [9]. One literature review on the "agile-to-lean" transformation [10], focused on categorizing and comparing the transformations including 30 experience reports. Compared to the study by Wang et al. [10], our study is more focused on benefits, challenges and metrics.

B. Research questions

RQ1. *What are the **benefits** that have been reported after the adoption of lean principles in the context of an ongoing agile development process?*

RQ2. *What are the **challenges** that have been reported after the adoption of lean principles in the context of an ongoing agile development process?*

RQ3. *Which **metrics** have been used to measure the "agile-to-lean" transformation?*

C. Search strategy & study selection

We have collected research papers available online in three digital repositories: Web of Science (WoS), IEEExplore, ACM Digital Library (DL) (on 2017-06-10, Fig. 1). We used *"lean software development OR agile transformation OR lean transformation"* as search string, as we preferred to start with a more general query and filter out results later.

The first stage of search strategy (automated search in online databases) included only studies written in English. The EndNote reference manager software was used for excluding duplicates and narrowing down the initial search results from 1,787 to 856 research papers (Fig. 1). We have included journals and conference papers published since 2003, after the work of Mary and Tom Poppendieck [4].

In the second stage (filtering based on title and abstract), multiple search criteria have been applied such as the exclusion of papers that involved lean manufacturing or any other subject areas outside software engineering. A total of 131 papers have passed the second stage of the search strategy (Fig. 1).

Stage three (full-text filtering) was performed manually going through the remaining entries. 18 papers were included based on full text reading. Quality criteria (section II-D) were applied to find papers involving organizational transformation from agile to lean. After quality assessment, a total of 8 papers were included in the final SLR list (Fig. 1).

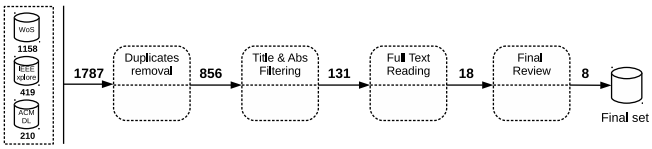


Fig. 1. SLR papers selection process

D. Study quality assessment criteria

After 18 papers were selected (Table I), we have conducted a quality assessment process to select the papers fitting the exact purpose of this research:

- C1. *Is the agile methodology at the initial state of the transformation?*
- C2. *Is the reported company making a transformation towards lean software development practices?*
- C3. *Does the paper provide metrics used for the transformation and states clear outcomes?*

Papers were given points from 0 (meaning *"no"*), 0.5 (*"partly"*) to 1 (*"yes"*). All papers reaching the score marked as ≥ 2.5 for $sum(C1, C2, C3)$ (Table I) were included for conducting the literature review (Table II).

TABLE I
QUALITY ASSESSMENT OF THE SELECTED PAPERS

SLR	Article	C1	C2	C3	Score
S1	Hayata et al. [11]	yes	no	no	1.0
S2	Jakobsen and Poppendieck [12]	yes	yes	yes	3.0
S3	Kuusela and Koivuluoma [13]	yes	partly	no	1.5
S4	Middleton and Joyce [2]	yes	yes	yes	3.0
S5	Misaghi and Bosnic [14]	yes	yes	yes	3.0
S6	Paasivaara et al. [15]	no	yes	yes	2.0
S7	Perera and Fernando [16]	yes	yes	yes	3.0
S8	Petersen and Wohlin [17]	yes	yes	yes	3.0
S9	Rodríguez et al. [18]	no	yes	yes	2.0
S10	Rodríguez et al.[19]	yes	yes	yes	3.0
S11	Samanta et al. [20]	-	yes	yes	2.0
S12	Schnitter and Mackert [21]	yes	partly	partly	2.0
S13	Sjøberg et al. [22]	yes	no	partly	1.5
S14	Swaminathan and Jain [23]	yes	yes	yes	3.0
S15	Trimble and Webster [24]	no	yes	no	1.0
S16	Vilkki [25]	no	yes	no	1.0
S17	Viswanath [26]	no	yes	yes	2.0
S18	Walter et al. [27]	yes	yes	partly	2.5

III. SLR RESULTS

A. Benefits (RQ1)

To answer our first research question (RQ1, Fig. 2), we have reviewed and categorized the reported benefits.

1) Reduced lead time: The most common benefit that has been reported is reduced lead time as it was found in six studies [2], [16], [17], [19], [23], [27]. Middleton and Joyce [2] describe lead time as the total time recorded from a customer request to the completed work delivery. Reducing

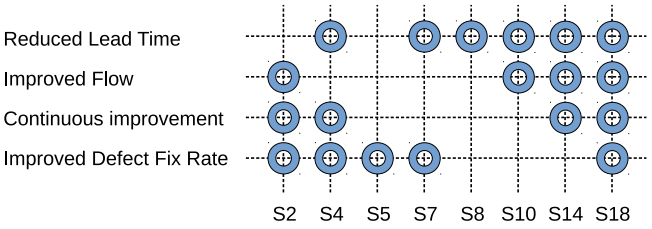


Fig. 2. Benefits mapped to papers. O = benefit reported in the paper

lead times contributes to flow improvement so that activities are organized continuously, enabling smooth deliveries to the customer. Middleton and Joyce [2] also state that they have experienced 47% less variance and on average 37% shorter lead time to deliver software, which is a significant improvement. Moreover, Walter et al. [27] have recorded an enhancement of approximately 70% which can dramatically increase the time-to-market responsiveness. Early feedback from the customer implies frequent integration, an important factor to improve the software product quality.

2) Improved flow: Improved flow was reported in four cases [12], [19], [23], [27]. Walter et al. [27] have discovered that the secret to improve the flow is to control Work-in-Progress (WiP) items. By reducing the number of simultaneous tasks they have reached lean flow state with constant throughput. Flow can be seen as the number of WiP items, so that lowering the number of such items can speed-up the whole process and more features can be implemented and eventually delivered. Therefore, to speed-up the flow, it is essential to remove waste in the inventory to avoid piling up user stories. Rodríguez et al. [19], Middleton and Joyce [2] state that limiting WiP is an important element for achieving flow. Limited WiP leads to more organized activities so that the improved flow can lead to smooth deliveries [19]. Additionally, continuous integration and test automation significantly supported the flow by frequent and smaller builds [19]. According to Swaminathan and Jain [23], tracking and acting on the visual indicators provided by the cumulative flow diagram helped to maintain a uniform flow. Moreover, it also helped to identify and remove bottlenecks and improve the efficiency of the process. This increase in efficiency also enforced rapid development and continuous improvement to software delivery [23]. Jakobsen and Poppendieck [12] have improved the flow of story implementation from 30% to 60%. The flow was improved in various areas such as test, development, project start-up and customer activities related to contracting and ongoing clarifications, major benefits for the company.

3) Continuous improvement: Four studies [2], [12], [23], [27] experienced continuous improvement as a benefit of their transformation. In Swaminathan and Jain [23], the story rate per iteration was used as a metric to measure continuous improvement, which was proved by the evidence of cumulating story points. On the other hand, the basis for the continuous improvement in Walter et al. [27] was established by letting each team self-organize and set their own WiP size.

Continuous improvement carried out on a daily basis showed a significant increase in the software delivery pre-

TABLE II
PRIMARY STUDIES LINKED TO THE REFERENCES WITH COUNTRY AND DOMAIN OF THE CASE STUDY/EXPERIMENT

Paper ID	Reference	Domain	Type	Year	Country
S2	Jakobsen and Poppendieck [12]	Software company: complex and critical IT solutions	case study	2005	Denmark
S4	Middleton and Joyce [2]	Webmedia department software processes	case study	2009	UK
S5	Misaghi and Bosnic [14]	Software company: leading supplier of systems for the supply chain	case study	2011	Brazil
S7	Perera and Fernando [16]	Students groups during len-to-agile transformation	experiment	2007	Sri Lanka
S8	Petersen and Wohlin [17]	Large provider of ICT to service providers	case study	2009	Sweden
S10	Rodríguez et al. [19]	Wireless Embedded Systems	case study	2010	Finland
S14	Swaminathan and Jain [23]	Multinational IT consulting organization	case study	2012	India
S18	Walter et al. [27]	Software development for big telecommunication companies	case study	2015	Brazil

dictability [2]. Moreover, data collected over a twelve months period showed significant improvement as the time taken to resolve issues was reduced by 81%.

4) Improved defect fix rate: Adopting lean principles to an ongoing agile process has reflected in improved defect fix rate in five papers [2], [12], [14], [16], [27]. Lean promotes finding and fixing defects early, so there is a better and control over quality from the beginning. Therefore, continuous integration tools are often widely used in lean software development paradigm. Middleton and Joyce [2] not only managed to fix issues in a shorter period, but also experienced a lower amount of bugs. As the bug rate decreased, the team had reportedly more time for completing customer stories. Better product quality was reported by Misaghi and Bosnic [14] with the reduction of time spent on bug fixing. Similar to the study by Middleton and Joyce [2], fewer errors were released because of the higher amount of time for new features and improvements [14]. Another success story of enhanced software quality has been achieved with 50% bug reduction as reported by Walter et al. [27]. Two experimental groups for measuring defect rate had been observed during the experiment by Perera and Fernando [16]. One group adopted the combination of lean and agile, the other one just pure agile methodology. The pure agile group experienced a lower amount of defects at the beginning. However, this situation swapped at later stages and the lean-agile sample had a minimal defects rate [16]. A higher number of defects for the agile group was probably due to unfixed hidden defects at the early stages.

B. Challenges (RQ2)

In general, it was difficult to extract information about the challenges from the papers (RQ2, Fig. 3), as controversial/negative aspects might be omitted from papers.

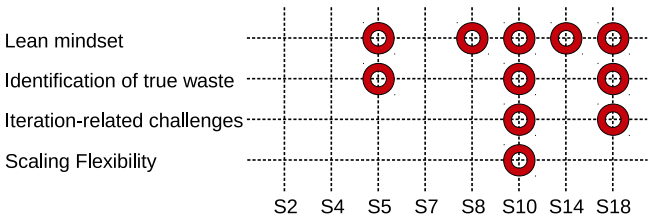


Fig. 3. Challenges mapped to papers. O = metric reported in the paper

An ”agile-to-lean” transformation brings a series of challenges. Maintaining process visibility, managing sustainability, and communication among teams are seen often as key challenges [5]. The improvement of the testing process is also reported as a key challenge, as the driving principle is perfectionism and identification of root-causes for software defects [5]. As such, placing lean on top of agile brings even more importance to the testing process. Acquiring the proper mindset can be seen as a challenge in lean adoption [14], [20], as lean requires a different mindset compared to the application of pure agile practices. Another challenge of ”agile-to-lean” is to integrate the concepts of waste minimization and quality improvements that are part of the lean philosophy, bringing to a more complete development and management process [16]. Lean on top of agile brings more the focus on the end-to-end value flow of the whole development process, thus putting lots of emphasis on different tools and their support, like value stream mapping or Kanban [17]. Constant management of flow is also a relevant issue [27], together with building a lean mindset towards defects reduction [23]. Scaling flexibility, business management involvement and waste reduction were found as challenges, with scaling flexibility problematic due to the management of the whole value streaming, making flexibility more difficult to reach [19].

1) Adapting to the Lean Mindset: In general, resistance to change is a common problem when trying to adopt new ways of working. This was the case in the reviewed studies [14], [17], [19], [23], [27]. Therefore, getting the commitment to this new paradigm is required from the longtime perspective as it is a continuous process that needs to be enforced. Misaghi and Bosnic [14] state that defining the criteria to implement the *”lean mindset”* into the organization is a main challenge. On the other hand, even though there are some tough challenges at the beginning to get the team to think end-to-end and work in a new way, once the team starts to see the added value, such way becomes naturally accepted within the team [23].

2) Identification of ”true waste”: Waste reduction is the key principle to maintain when working with a lean mindset. Rodríguez et al. [19], Walter et al. [27] found that it can be hard to identify *”what a true waste is”* within the organization and it may be even more challenging to eliminate such waste. Although setting-up teams and establishing self-organization within teams have not been hard to achieve, scaling flexibility and involvement of business management tasks were much more challenging in the lean way of working [14]. Aspects like people multitasking, which may seem at first appropriate to be more efficient, can also be seen as a waste as they might be the major cause slowing down productivity [14].

3) Iteration-related challenges: Studies have faced challenges also with coaching, estimates, pair programming, all during the run of development iterations [27]. In Walter et al. [27], team coaches had difficulties to ensure that tasks were delivered on time and with quality. For this reason, the coaching process adopted some more visible indicators (physical flags). Formation of pair programming couples and iteration estimations were adapted to the needs of the process [27]. In Rodríguez et al. [19], teams complained about too long feedback loops, caused by the involvement of business management in the whole value stream mapping process.

4) Scaling Flexibility: Scaling flexibility was defined as the easiness of performing changes during the software development process and was found as one of the key challenges in Rodríguez et al. [19]. The issue in "agile-to-lean" is that flexibility needs to pass through the whole value stream, making the process more complex than in pure agile contexts [19].

C. Metrics (RQ3)

To answer our third research question (RQ3, Fig. 4), we have reviewed and categorized findings of metrics that have been used to measure the lean transformation.

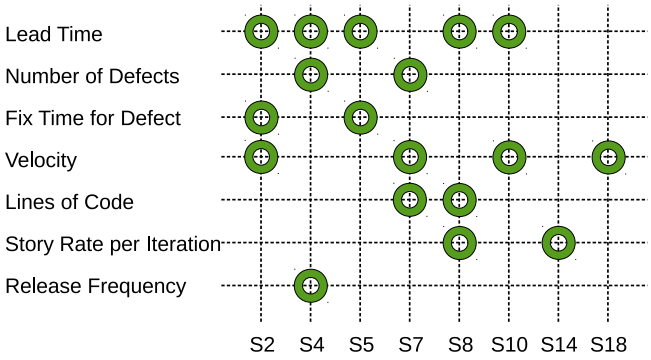


Fig. 4. Metrics mapped to papers. O = metric reported in the paper

1) Lead time: Five studies [2], [12], [14], [17], [19] have used lead time as one of the metrics to measure the progress of their organizational transformation. As we have described earlier (Section III-A about the benefits), lead time represents the duration from the customer request until the time the product is shipped to the customer. Lead time is usually measured in working days and it is stopped when user acceptance testing is complete and the product is ready for release [2]. Petersen and Wohlin [17] achieved higher responsiveness to customer needs by means of reduced lead time, managing to decrease time for delivery. Petersen and Wohlin [17] claim that this metric is highly important as the customer often needs frequent changes. The ability to respond to these changes quickly is a powerful competitive advantage. Along with this metric, Middleton and Joyce [2] reported using also cycle time, that can be considered as a sub-value of lead time. Cycle time is the time from actual initiation of the feature development (start of the working on the item) until the work on the feature is completed.

2) Number of defects: The number of defects per given time period was used as a metric in two studies [2], [16]. The main focus of Perera and Fernando [16] was on minimizing the number of defects. However, this study points out that higher number of defects at the early stages is expected, as described in section III-A4. Therefore, number of defects need to be measured with a long time perspective in mind. Middleton and Joyce [2] were measuring the number of defects per week for a twelve-month time period. The mean numbers of bugs open each week of their issue tracking system declined [2].

3) Fix time for defect: Similarly to the previous metric, two other studies [12], [14] examined defects from the fixing time perspective. Misaghi and Bosnic [14] have observed that, as the time spent on developing new features increased, the time spent on bug fixing decreased. The overall quality of the product improved as the releases contained fewer defects [14]. Measurements have been observed for the period of one year, that shows the positive long-term effect of lean.

4) Velocity: We have found this metric in four studies [12], [16], [19], [27]. Velocity can be measured by dividing the expected time of task completion by the actual time the task has been closed. Walter et al. [27] have improved their velocity by categorizing the time estimates using different sizes (extra-small, small, medium, big, extra-big). The number of hours was estimated based on their historical values and added to each category [27]. In Jakobsen and Poppendieck [12], velocity was measured as a sub-flow for story implementation, ensuring that the stories were developed in a smooth flow, eliminating the waste associated with context shifts and handovers. To verify whether the project is achieving the goals related to its schedule, expected work and actual work levels were also used by Perera and Fernando [16]. However, this time the metric was slightly modified by dividing the divergence between actual and expected work level with the expected work level [16].

5) Lines of code: Two studies [16], [17] have been measuring lean performance by evaluating the number of Lines of Code (LoC) developed. The outcome of the study by Perera and Fernando [16] was that the hybrid lean-agile approach produced more LoCs. Perera and Fernando [16] evaluated this metric from various perspectives such as new LoCs, removed LoCs and changed LoCs. On the other hand, Petersen and Wohlin [17] used this metric in a slightly different way, by measuring value efficiency: the difference between the value of output and value of input within a given time-frame [17].

6) Story rate per iteration: Customer requirements are captured in the form of user stories, which are afterwards estimated and prioritized [28]. For visualizing the long-term effect with this metric, data have been displayed mostly in story flow diagrams and cumulative story flow diagrams to measure continuous improvement from the longtime perspective, which also helped to identify the piling up of inventory [23]. Depending on the time dedicated to development, each story was given a story point value discussed and sorted out by the developers of the team. Usually, developers were estimating the story size as small and even, to better structure

their work. Swaminathan and Jain [23] measured story rate per iteration as the total number of story points approved and closed by the customer in the given iteration. The flow of requirements through the software development life-cycles was the key topic also in Petersen and Wohlin [17]. However, this study was measuring hand-overs within the stories as well as the variance, to better predict the development cycle [17].

7) Release frequency: The only study which used the number of releases was Middleton and Joyce [2], defining it as the number of items released to customers. Time-frame for measuring the frequency of releases was set to one month. Even though this metric does not reveal how much value is being delivered to the customers, it showed an eight-fold increase in releases for a two years period [2]. This is indicating an improvement in configuration management discipline and capability [2], as the more frequent releases are reducing technical and market risks, as the customer can evaluate a real product in smaller increments rather than just seeing temporary results from progress reports.

IV. Conclusion

The goal of this paper was to better understand the "agile-to-lean" transformation process regarding benefits, challenges, and metrics that primary studies reported in transformations within companies. To reach the goal, we conducted a Systematic Literature Review (SLR) [6]. The most represented benefits were reduced lead time, improved flow, continuous improvement and improved defect fix rate. Seeking the challenges faced, the common problems were the adaptation to the lean mindset, teaching and maintaining the "lean mindset", maintaining development flexibility, and the identification of the concept of waste. The most used metric to measure a lean transformation was lead time.

References

[1] X. Wang, "The combination of agile and lean in software development: An experience report analysis," in *2011 Agile Conference*, Conference Proceedings. doi: 10.1109/AGILE.2011.36 pp. 1–9.

[2] P. Middleton and D. Joyce, "Lean software management: Bbc worldwide case study," *IEEE Transactions on Engineering Management*, vol. 59, no. 1, pp. 20–32, 2012. doi: 10.1109/TEM.2010.2081675

[3] M. Kalenda, P. Hyna, and B. Rossi, "Scaling agile in large organizations: Practices, challenges, and success factors," *Journal of Software: Evolution and Process*, p. e1954. doi: 10.1002/smr.1954

[4] M. Poppendieck, T. Poppendieck, and T. Poppendieck, *Lean Software Development: An Agile Toolkit*, ser. The agile software development series. Addison-Wesley, 2003. ISBN 9780321150783

[5] A. Shalloway, G. Beaver, and J. R. Trott, *Lean-agile software development: achieving enterprise agility*. Pearson Education, 2009.

[6] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[7] P. Middleton, "Lean software development: two case studies," *Software Quality Journal*, vol. 9, no. 4, pp. 241–252, 2001. doi: https://doi.org/10.1023/A:1013754402981

[8] K. B. Stone, "Four decades of lean: a systematic literature review," *International Journal of Lean Six Sigma*, vol. 3, no. 2, pp. 112–132, 2012. doi: https://doi.org/10.1108/20401461211243702

[9] E. Kupiainen, M. V. Mäntylä, and J. Itkonen, "Using metrics in agile and lean software development–a systematic literature review of industrial studies," *Information and Software Technology*, vol. 62, pp. 143–163, 2015. doi: https://doi.org/10.1016/j.infsof.2015.02.005

[10] X. F. Wang, K. Conboy, and O. Cawley, ""leagile" software development: An experience report analysis of the application of lean approaches in agile software development," *Journal of Systems and Software*, vol. 85, no. 6, pp. 1287–1299, 2012. doi: 10.1016/j.jss.2012.01.061

[11] T. Hayata, J. Han, and M. Beheshti, "Facilitating agile software development with lean architecture in the dci paradigm," in *2012 Ninth International Conference on Information Technology - New Generations*. doi: 10.1109/ITNG.2012.157 pp. 343–348.

[12] C. R. Jakobsen and T. Poppendieck, "Lean as a scrum troubleshooter," in *2011 Agile Conference*, Conference Proceedings. doi: 10.1109/AGILE.2011.11 pp. 168–174.

[13] R. Kuusela and M. Koivuluoma, "Lean transformation framework for software intensive companies: Responding to challenges created by the cloud," in *2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications*, Conference Proceedings. doi: 10.1109/SEAA.2011.74. ISBN 1089-6503 pp. 378–382.

[14] M. Misaghi and I. Bosnic, "Lean mindset in software engineering: A case study in a software house in brazilian state of santa catarina," vol. 466, pp. 697–707, 2014. doi: https://doi.org/10.1007/978-3-319-11854-3_60

[15] M. Paasivaara, C. Lassenius, V. T. Heikkila, K. Dikert, and C. Engblom, "Integrating global sites into the lean and agile transformation at ericsson," *2013 Ieee 8th Int. Conference on Global Software Engineering (Icgse 2013)*, pp. 134–143, 2013. doi: 10.1109/icgse.2013.25

[16] G. I. U. S. Perera and M. S. D. Fernando, "Enhanced agile software development - hybrid paradigm with lean practice," in *2007 Int. Conference on Industrial and Information Systems*, Conference Proceedings. doi: 10.1109/ICIINFS.2007.4579181. ISBN 2164-7011 pp. 239–244.

[17] K. Petersen and C. Wohlin, "Measuring the flow in lean software development," *Software-Practice and Experience*, vol. 41, no. 9, pp. 975–996, 2011. doi: 10.1002/spe.975

[18] P. Rodríguez, J. Markkula, M. Oivo, and K. Turula, "Survey on agile and lean usage in finnish software industry," in *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '12. New York, NY, USA: ACM, 2012. doi: 10.1145/2372251.2372275. ISBN 978-1-4503-1056-7 pp. 139–148.

[19] P. Rodríguez, J. Partanen, P. Kuvaja, and M. Oivo, "Combining lean thinking and agile methods for software development: A case study of a finnish provider of wireless embedded systems detailed," in *2014 47th Hawaii International Conference on System Sciences*, Conference Proceedings. doi: 10.1109/HICSS.2014.586. ISBN 1530-1605 pp. 4770–4779.

[20] U. Samanta, V. S. Mani, and Ieee, "Successfully transforming to lean by changing the mindset in a global product development team," pp. 135–139, 2015. doi: 10.1109/icgse.2015.17

[21] J. Schnitter and O. Mackert, "Large-scale agile software development at sap ag," vol. 230, pp. 209–220, 2011. doi: https://doi.org/10.1007/978-3-642-23391-3_15

[22] D. I. K. Sjøberg, A. Johnsen, and J. Solberg, "Quantifying the effect of using kanban versus scrum: A case study," *IEEE Software*, vol. 29, no. 5, pp. 47–53, 2012. doi: 10.1109/MS.2012.110

[23] B. Swaminathan and K. Jain, "Implementing the lean concepts of continuous improvement and flow on an agile software development project: An industrial case study," in *2012 Agile India*, Conf. Proceedings. doi: 10.1109/AgileIndia.2012.12. ISBN 2326-6007 pp. 10–19.

[24] J. Trimble and C. Webster, "From traditional, to lean, to agile development: Finding the optimal software engineering cycle," in *2013 46th Hawaii Int. Conference on System Sciences*, Conference Proceedings. doi: 10.1109/HICSS.2013.237. ISBN 1530-1605 pp. 4826–4833.

[25] K. Vilkki, *When Agile Is Not Enough*, ser. Lecture Notes in Business Information Processing, 2010, vol. 65, pp. 44–47. ISBN 978-3-642-16415-6

[26] U. Viswanath, "Lean transformation: How lean helped to achieve quality, cost and schedule: Case study in a multi location product development team," in *2014 IEEE 9th Int. Conference on Global Software Engineering*, Conf. Proceedings. doi: 10.1109/ICGSE.2014.13 pp. 95–99.

[27] M. Walter, R. Tramontini, R. M. Fontana, S. Reinehr, and A. Malucelli, *From Sprints to Lean Flow: Management Strategies for Agile Improvement*, ser. Lecture Notes in Business Information Processing, 2015, vol. 212, pp. 310–318.

[28] M. Pergher and B. Rossi, "Requirements prioritization in software engineering: a systematic mapping study," in *Empirical Requirements Engineering (EmpiRE), 2013 IEEE Third International Workshop on*. IEEE, 2013. doi: 10.1109/EmpiRE.2013.6615215 pp. 40–44.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 975–978
DOI: 10.15439/2018F151
ISSN 2300-5963 ACSIS, Vol. 15

Problems and Solutions of Software Design in Scrum Projects

Jakub Miler
Gdansk University of Technology
Faculty of Electronics, Telecommunications and Informatics
11/12 Narutowicza St., 80-233, Gdansk, Poland
Email: jakub.miler@eti.pg.edu.pl

Kamil Kajdy
IHS Global sp. z o.o.
163 Marynarki Polskiej St.,
80-868, Gdansk, Poland
Email: kamil.kajdy@gmail.com

***Abstract*— The aim of the paper is to identify the problems and solutions of the software design in Scrum project as well as to analyze the effectiveness of the solutions. Through a series of workshops with 4 experts from IT industry and academia we have identified 52 problems and 99 unique solutions. In this paper we present a list of 10 common problems and 5 solutions for each problem selected by the number of sources. The effectiveness of the solutions to the given problems was evaluated in an opinion survey by 39 respondents with experience both in software design and in the Scrum framework. This evaluation provided for our initial recommendations on the choice of solutions to particular problems.**

I. Introduction

Software design is one of the key elements of software engineering [1], [2]. Systematic approach to architecture, code structure, data processing, and other aspects is required for many types of systems based on their size, complexity, distribution, and quality factors e.g. safety, security [3], [4]. Development practices such as pair programming, continuous integration, test driven development [5], [6], design patterns, refactoring [7] or clean code principles [6], [8] provide solutions to many problems, but their application in practice is challenged by the development methodology, technology, team, customer and many more.

Scrum defines only the roles, artifacts and events of the development process on a general level and leaves the room for specific decisions and actions to the Scrum Team [9]. This includes the design, programming and testing of software, where the Scrum Team should be multifunctional to cover all the competencies necessary to deliver the product [10] and include the role of a software architect if necessary [11]. Additionally, Scrum promotes working product increment after each sprint leaving little time for detailed approach to architecture and design [6], [10]. It is recommended to design as little and as late as possible to avoid negating the design by changing requirements [12], [13], [14]. This approach results in increasing technical debt which is related to the low quality of design and code [12]. It is also not possible to apply in case of complex systems and scaled Scrum [15].

Some of the recent research studied the relationship between the architecture-centric design and the agile development, but the authors focused either on the eXtreme Programming framework [16] or the agile projects in general [17]. This shows that the integration of the software design principles and the Scrum framework is not straightforward and calls for a detailed inquiry.

Our research goal was to analyze the problems and solutions of software design in Scrum projects. To achieve this goal, 3 research questions were formulated: (RQ1) What are the problems with software design in the Scrum projects? (RQ2) What are the solutions to the software design problems in the Scrum projects? (RQ3) Which solutions to the problems with software design can be recommended to the Scrum projects?

The contribution of this paper is the identification of the problems and solutions of software design in Scrum projects as well as some initial recommendations of the effective solutions to the most common problems.

The paper is organized as follows. Section II describes the research method, the workshops with experts and the online survey. Section III presents the list of the top problems and their solutions as well as the evaluation of these solutions together with some recommendations. Section IV discusses threats to the validity of this research followed by the conclusions in Section V.

II. Research Method

Our research method comprised two techniques: the workshops with experts to identify the problems and their solutions, and the online survey to evaluate the perceived effectiveness of the solutions to particular problems.

The workshop was designed as a structured multi-phase brainstorming session with the following steps:

1. introduction to the workshop, explanation of the goals and the scope,
2. individual identification of problems,

This work was supported by DS Funds of ETI Faculty, Gdansk University of Technology.

3. discussion of the problems identified in step 2, aggregation of duplicate problems,
4. individual identification of the solutions to the problems resulting from step 3 (a solution may solve more than one problem),
5. discussion of the solutions identified in step 4, aggregation of duplicate solutions.

The workshop involved a domain expert and one of the researchers (K. Kajdy) as a moderator. The scope focused on the specific aspects of the Scrum agile framework: development in short iterations, changing requirements, self-organized teams, and little documentation. Additionally, the aspects of software design were restricted to the following: component integration, architecture, design patterns, NoSQL or relational databases, user interfaces, modularization, and refactoring.

We have carried out workshops with 4 experts with at least 2 years of experience in both software design and the Scrum framework. The experts played the roles of developers and/or Scrum Masters. Each workshop resulted in a distinct list of problems and solutions to these problems. Finally, a compiled list of problems and their solutions was built from the results of all 4 workshops. The merging was based on keyword analysis.

We have selected 10 problems and 5 solutions to each of these problems for the evaluation survey (50 solutions in total) to limit the size of the survey and increase the rate of feedback. The problems and solutions were selected primarily based on the total number of indications in the source workshops.

The effectiveness of each solution in relation to a given problem was assessed in a Likert-type 5 level scale of 1 to 5, where 1 meant "a solution is totally ineffective to the problem" and 5 meant "a solution is very effective to the problem" with an escape answer "I don't know". We have also asked about the respondents' experience in software design and in the Scrum framework. Although Likert-type scale is ordinal, in the data analysis we have treated it as numerical with assigned values of 1 to 5. The evaluation of each solution's effectiveness was calculated as a weighted average, where weights represented the respondents' experience: 0.1 – under 1 year; 0.3 – 1-2 years; 0.6 – 2-3 years; 0.85 – 3-5 years; 1 – above 5 years.

III. Results

The identification workshops were carried out in May and June 2017. On average, the experts identified 25.75 problems, 31 unique solutions and 75.75 total solutions per workshop. In total, they have identified 52 problems, 99 unique solutions and 231 total solutions to all problems. The detailed results of the workshops as well as the full list of merged problems and solutions are available in [18].

The evaluation survey was carried out in August and September 2017 with Google Forms. It was promoted among the IT practitioners via e-mail and social media. In total, 39 respondents took part in the survey. 22 respondents (56%) had at least 2 years of experience with software design and 20 respondents (51%) had at least 2 years of experience with Scrum.

Table I and Table II present the identified problems and their evaluated solutions. The columns are as follows: identifier, problem/solution name, evaluation with a weighted average and weighted standard deviation in parentheses, survey sample size (N), and number of indications in the workshops (n). The problems are ordered by the number of indications (n) and the solutions for each problem are ordered by their evaluation (avg.).

Most of the top evaluated solutions reached a score close to 4 or more than 4. Problem P2 is the exception with a top evaluated solution of 3.49. This indicates the need for further research on its better solutions. Problems P3, P4, P5, P7, P8, and P10 can be assigned a clear leading solution with top score of more than 4. Additionally, more than one solution for problems P7 and P8 have reached the score of 4. The top scoring solutions for problems P1, P6, and P9 have an evaluation of slightly below 4, but the top solutions are still significantly ahead of the rest except for the problem P6, where 4 top solutions are enclosed within the range of 0.1.

As for the lowest scoring solutions, it can be observed that the solutions S1 and S36 are evaluated as least effective for all problems they were assigned to (S1 to problems P1, P7, P9, and P10; S36 to problems P5 and P6) with sample size of more than 30. They were, however the top solutions in the identification phase resulting from 4 and 3 sources respectively. The experts' belief in their effectiveness has been significantly challenged by the survey. It may indicate that these solutions strongly depend on factors specific to business environments (e.g. personnel, culture, type of products), which can be further studied in future research.

It should be noted that some of the solutions can be hard to apply in a strictly agile environment. Formal review of projects (S4) can go beyond the visibility and transparency principles of Scrum and Agile Manifesto leading to an overly monitored and manually managed team. Task estimation and accounting recommendations such as S25 or S31 can also hamper the customer-developer trust Agile is based on. S34 refers to the role of a project manager, which is outside of the Scrum framework and calls for a project management methodology on top of Scrum. This can be considered a non-agile practice.

Some solutions are also technology or architecture dependent e.g. NoSQL databases (S24), API versioning (S52) or microservices (S61), which also limits their application. It may not be beneficial or possible at all to implement such solutions in particular systems.

The proposed list of problems and their evaluated solutions has mostly educational use by Scrum developers, Scrum Masters and coaches. The application of a solution in a particular project shall always be discussed and accepted within the Scrum Team.

TABLE I.
PROBLEMS P1-P5 AND THEIR EVALUATED SOLUTIONS

Id	Name	Avg.	N	n
P1	**Team work assessed mainly with of code increments and new functionalities**			**8**
S3	Promoting quality and designing in the organization and to the client	3.88 (1.35)	34	2
S2	Avoiding creating fast and large increments at the expense of design and quality of code	3.53 (1.34)	36	2
S4	Formal review of projects	3.20 (1.39)	32	2
S5	An organization's policy that only part of the time is devoted to working with the code	3.18 (1.10)	34	2
S1	Professional Scrum Master teaching team communication and promoting issues of architecture and design at the meetings	3.12 (1.38)	33	4
P2	**Problems with expanding and modifying the production database in the client's environment**			**6**
S17	Automation of creating data models from code	3.49 (1.45)	32	1
S24	NoSQL databases	3.05 (1.49)	18	1
S21	Designing the database changes one sprint earlier or at the very beginning of the sprint	2.96 (1.19)	31	1
S20	Small database design at the beginning (the less data collected, the less data to update)	2.78 (1.48)	30	1
S16	Making modifications to the database once every few sprints	2.61 (1.22)	30	3
P3	**Recognizing refactoring as an increment by the client, despite the client's resistance**			**5**
S28	Doing refactoring partially in each sprint, not all in one sprint	4.04 (1.21)	38	1
S29	Using the refactoring automation tools	3.87 (1.02)	29	1
S25	Including the refactoring costs in the price of an expensive task	3.77 (1.08)	33	5
S27	Educating the client and obtaining approval for corrective and maintenance actions	3.50 (1.12)	36	2
S26	Introduction of stabilization sprints for code maintenance	3.36 (1.30)	33	2
P4	**Improperly defined tasks that hamper planning and design**			**5**
S30	Grooming before planning - examining and presenting details of a given User Story	4.09 (0.77)	33	3
S31	Overestimating tasks to leave time for "unpredictable"	3.89 (1.06)	36	1
S32	A business analyst present on the planning and available to the team	3.73 (1.09)	33	1
S34	Project Manager that accurately defines the tasks	3.71 (1.01)	36	1
S35	Behavior Driven Development - Gherkin language	3.07 (1.10)	17	1
P5	**Difficulties with introducing new functionalities due to architectural errors**			**5**
S42	Separation of views, data and business logic	4.50 (0.64)	35	1
S38	Applying the initial conceptual and design phase before the actual implementation	3.97 (0.88)	36	2
S39	A team using design patterns, standards, diagrams	3.90 (1.01)	33	2
S37	Making the client aware of the time needed for the design and that it will pay back	3.82 (1.18)	35	2
S36	Preparation of prototypes and preliminary design in "sprint 0"	3.33 (1.14)	35	3

TABLE II.
PROBLEMS P6-P10 AND THEIR EVALUATED SOLUTIONS

Id	Name	Avg.	N	n
P6	**Problems resulting from the selection of project technology in advance, before the implementation**			**3**
S38	Applying the initial conceptual and design phase before the actual implementation	3.98 (0.95)	35	2
S48	Careful selection of technologies - proven technologies for large projects, experiments with fast Proof of Concepts	3.94 (1.08)	36	1
S44	Checking the technologies available on the market as part of the task of the increment	3.91 (1.02)	34	2
S37	Making the client aware of the time needed for the design and that it will be pay back	3.88 (0.98)	35	2
S36	Preparation of prototypes and preliminary design in "sprint 0"	3.30 (1.07)	36	3
P7	**Problems with developing a uniform communication interfaces between modules**			**3**
S52	API versioning	4.37 (0.76)	32	1
S39	A team using design patterns, standards, diagrams	4.10 (0.95)	31	2
S3	Promoting quality and designing in the organization and to the client	3.90 (0.90)	32	2
S51	The design created 1 sprint earlier or at the very beginning of the sprint	3.37 (1.15)	33	1
S1	Professional Scrum Master teaching team communication and promoting issues of architecture and design at the meetings	2.90 (1.18)	31	4
P8	**Problems with technological debt and poor quality due to rush in implementation**			**3**
S54	Applying SOLID practices and adhering to the rules of clean code	4.37 (0.77)	33	1
S56	Multiphase code reviews	4.15 (0.89)	35	1
S28	Doing refactoring partially in each sprint, not all in one sprint	4.04 (1.18)	37	1
S27	Educating the client and obtaining approval for corrective and maintenance actions	3.85 (1.01)	36	2
S26	Introduction of stabilization sprints for code maintenance	3.75 (1.29)	35	2
P9	**Difficulties with breaking down tasks into smaller tasks**			**3**
S60	Transferring detailed design problems to separate meetings of selected people	3.94 (0.86)	36	1
S25	Including the refactoring costs in the price of an expensive task	3.58 (1.19)	30	5
S61	Application architecture based on microservices	3.38 (1.06)	29	1
S39	A team using design patterns, standards, diagrams	3.34 (0.97)	34	2
S1	Professional Scrum Master teaching team communication and promoting issues of architecture and design at the meetings	2.94 (1.20)	32	4
P10	**Mutual blocking of implementation tasks**			**3**
S30	Grooming before planning - examining and presenting details of a given User Story	4.16 (0.85)	35	3
S50	Informal conversations and arrangements (helping to avoid blocking tasks)	3.83 (1.05)	37	1
S39	A team using design patterns, standards, diagrams	3.71 (0.84)	32	2
S62	Assigning tightly related tasks to one developer	3.61 (0.95)	37	1
S1	Professional Scrum Master teaching team communication and promoting issues of architecture and design at the meetings	2.93 (1.24)	34	4

IV. Validity threats

A. Threats to construct and internal validity

We have controlled the workshop moderator's bias and his impact on experts with the structure of the workshop, which included the steps of individual identification (steps 2 and 4). Only then were the identified problems and solutions discussed and merged. The moderator was open to further expert's explanations.

The incorrect interpretation of the output from experts was controlled by writing down the output on the post-it notes and then discussing it to clearly understand the experts intentions. The moderator preserved all post-it notes after the workshop and built the resulting list of problems and solutions directly after the workshop referring to the notes and his fresh memory. For details on the workshop design see section II of the paper.

The interview experts represent the above average experience of our sample. Only 7 of 39 survey respondents had more experience, which puts the interview experts in the top quartile of the survey sample. What is the most important, the interview experts were able to identify large number of problems and solutions from their experience.

Our weight system is arbitrary at the moment, but it was designed to represent the assumed learning curve of software design and the Scrum framework based on the university syllabus and the authors' work experiences. We plan to study the learning curve of the Scrum framework in the future.

B. Threats to external validity

The number of experts was limited to 4 due to several factors. First, the set of data collected after 4 workshops was very satisfactory and we agreed on finishing this phase of research at this stage. Second, we required our experts to have experience both in software design and in the Scrum framework, which significantly limited the available sample. We have involved experts from various business environments: academia, technological start-up, small company, and large multinational corporation. The number of respondents was also limited due to the specific set of competencies required for the survey.

Our sample is not statistically random, but the experts and respondents were identified and contacted with various channels such as personal contacts, business contacts, social media, and recommendations from identified experts. This provided for a reasonably diverse group of practitioners.

We have based our research on data from experts and respondents working in the Polish market. This forms the natural limitation to our current results.

V. Conclusion

We have carried out 4 workshops with IT practitioners and identified 52 unique problems and 99 unique solutions. We believe that these results form a valuable answer to the research questions RQ1 and RQ2. Due to the limitations of this paper, we have presented only the 10 most commonly indicated problems as well as 5 solutions per problem selected for the survey.

We have acquired some evaluation of the effectiveness of the solutions to the 10 selected problems. We could point out some of the highly evaluated solutions as our initial recommendations as well as indicate the lowest evaluated solutions as risky. It should be considered, however, that the evaluations are based on the opinion poll only. This provides only a preliminary answer to the research question RQ3. Further and more detailed verification of the solutions' effectiveness in practice requires careful observation of a number of projects and can be done in future research.

Acknowledgment

The authors thank all the experts and respondents who took part in the identification workshops and the survey.

References

[1] I. Somerville, *Software Engineering*, 10th edition, Pearson, 2015
[2] R. S. Pressman, *Software Engineering: A Practitioner's Approach*, 8th Edition, McGraw-Hill Education, 2014
[3] J. Valacich, J. George, *Modern Systems Analysis and Design*, 8th edition, Pearson, 2016
[4] L. Maciaszek, *Requirements Analysis and Systems Design*, 3rd edition, Pearson Education Canada, 2007
[5] K. Beck, C. Andres, *Extreme Programming Explained: Embrace Change*, 2nd edition, Addison-Wesley, 2004
[6] M. Lacey, *The Scrum Field Guide: Practical Advice for Your First Year*, Addison-Wesley Professional, 2012
[7] R. C. Martin, *Agile Software Development, Principles, Patterns, and Practices*, Pearson, 2002
[8] R. C. Martin, *Clean Code: A Handbook of Agile Software Craftsmanship*, Prentice Hall, 2008
[9] K. Schwaber, *Agile Project Management with Scrum*, Microsoft Press, 2004
[10] K. Schwaber, J. Sutherland, *The Scrum Guide. Rules of the Game*, Scrum.org, 2017
[11] M. Cohn, *Succeeding with Agile: Software Development Using Scrum*, Addison-Wesley, 2010
[12] K. S. Rubin, *Essential Scrum: A Practical Guide to the Most Popular Agile Process*, Addison-Wesley Professional, 2012
[13] J. Rasmusson, *The Agile Samurai: How Agile Masters Deliver Great Software*, Pragmatic Bookshelf, 2010
[14] J. Sutherland, J. J. Sutherland, *Scrum: The Art of Doing Twice the Work in Half the Time*, Currency, 2014
[15] J. Diaz, J. Garbajosa, J. Perez, A. Yague, *Bridging User Stories and Software Architecture: A Tailored Scrum for Agile Architecting*, Agile Software Architecture: Aligning Agile Processes and Software Architectures, M. Ali Babar, A. W. Brown, I. Mistrik (eds.), Morgan Kaufmann, 2013
[16] R. L. Nord and J. E. Tomayko, "Software architecture-centric methods and agile development", *IEEE Software*, vol. 23, no. 2, pp. 47-53, 2006, DOI: 10.1109/MS.2006.54
[17] C. R. Prause and Z. Durdik, "Architectural design and documentation: Waste in agile development?", *2012 International Conference on Software and System Process (ICSSP)*, Zurich, 2012, pp. 130-134. DOI: 10.1109/ICSSP.2012.6225956
[18] K. Kajdy, *Analysis of software design in Scrum projects*, MSc Thesis, supervisor J. Miler, Gdansk University of Technology, Poland, 2017 (in Polish)

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 979–983
DOI: 10.15439/2018F229
ISSN 2300-5963 ACSIS, Vol. 15

Hard lessons learned: A model that facilitates the selection of methods of IT project management

Krzysztof Redlarski
Gdansk University of Technology
Faculty of Management and Economics, Gdansk, Poland
Department of Applied Informatics in Management
ul. Narutowicza 11/12, 80-233 Gdańsk, Poland;
krzysztof.redlarski@zie.pg.gda.pl

Abstract—*The article presents the results of research conducted in an international enterprise responsible for IT project implementation. The carried out analysis of the case study with the use of surveys and data synthesis allowed the major factors causing problems connected with project management to be identified. The identified factors were aggregated and then, by using four key variables, a rhomboidal model adaptation was proposed to facilitate the choice of the best method of project management. The proposed solution may aid Project Managers choosing the most appropriate method of project management as well as measuring and monitoring risk indicators.*

I. Introduction

The development of methods and tools of project management is a challenge for people responsible for project management [1, 5]. High dynamics of changes taking place in IT projects regarding client expectations and awareness means that it becomes more and more challenging to achieve the final success of the project [4, 8].

It is therefore necessary to continue to explore solutions that, based on the experience of completed projects, facilitate taking the right decision on the choice of methods of IT project management because incorrect decisions made in the initial phase of the life cycle are much more risky for the project. The high costs associated with removing the consequences of wrong decisions transfer into a project and are often not accepted by customers [7].

II. The development of IT project management

Current reference books outline a lot of methods of project execution [2]. They concentrate on three main approaches: classic [5], agile [4] and hybrid [8].

The classic approach, as the name suggests, uses a classic method for project management (i.e. PRINCE2, PMBOK - collection of good practices). This approach places a particular emphasis on precise project planning in such a way that later, in the project execution phase, it works towards maintaining the base plan adopted earlier. Classic methods are also characterized by a high level of formalization and they use a cascade approach to product or service development. The further stages of the project come one after another, without returning to the previous stage. The advantage of this approach is higher control and predictability of the project budget. Less flexibility regarding the necessity of implementing changes and the identification of the customers' needs are disadvantages.

The agile approach uses adaptation methods for project management (i.e. SCRUM, Extreme Programming, Lean). Agile methods have a low level of formalization and they assume an incremental and iterative way of product or service development. The spiral model of development used in this approach allows for returning to the earlier stages of the projects depending on the needs. The advantage is the possibility of having a flexible approach to the project and increased interaction with the product or service users [6]. The frequent problems related to project scope control and, as a consequence, related to budget control are a disadvantage. The higher engagement of the users in relation to classic methods increases the costs of product manufacturing but minimalizes the risk of making mistakes related to wrongly defined customer requirements.

The hybrid approach combines both the above approaches to project management alternately (i.e. classic and agile). Depending on the project structure, it uses both classic methods, mainly in the layer of management processes, and agile methods, mainly in the layer of production processes [7]. Thanks to such an approach, benefiting from the advantages of both methods described above is possible. Difficulty connected with the efficient estimation of the place and scope of using particular methods, as well as the necessity for Project Managers to have higher competences may prove to be disadvantages.

III. Research Methodology

In the presented research the following testing method was adopted: (Fig. 1)

Fig. 1. Research Methodology

This work was not supported by any organization

Step 1: Book references analysis. The objective of this step was to conduct an analysis of the research problem on the basis of the available reference literature. Within the analysis, a review of the main approaches to project management (classic, agile and hybrid), taking into account both advantages and disadvantages, was performed.

Step 2: Variables identification. The objective of this step was to identify and describe the variables causing problems related to project management. It was conducted in three stages:

- Phase 1: Identification of the variables of the analysis of the accomplished IT projects.
- Phase 2: Identification of the variables of the opinions provided by participants of the IT projects.
- Phase 3: Evaluation of the characteristics of real projects for the correctness of choosing the project execution method.

In this step, a survey questionnaire was used. It allowed the acquisition of opinions from participants of the projects. The provided data was subjected to synthesis, which allowed the main factors determining the choice of the particular project management method to be distinguished. Within this research, the characteristics of variables appearing in real projects were identified on the basis of a quality analysis of the current situation present in the examined enterprise.

Step 3: Decision model development. The objective of this step was to create a decision model which, thanks to the adaptation of a rhomboidal model, would ease decision making regarding the choice of the IT project management method.

The criterion of the appropriate selection of the project management method was defined as the general satisfaction of the project stakeholders with the delivered product or service within the executed project. Among the methods of project execution, three approaches were distinguished, i.e. classic, agile and hybrid.

Step 4: Model validation. The objective of this step was to verify the correctness of the assumed diagram model construction, the usage of which would allow the correct selection of the IT project management method on the basis of the identified variables. The model was subject to validation on the basis of the projects conducted in the examined enterprise.

IV. VARIABLES IDENTIFICATION

The examined enterprise accounts for the large business sector, employing approximately two thousand people. Within the overall structure, the IT Department was created. It is responsible for the design, realization and implementation of IT projects for internal clients' needs. The company executes their projects mainly according to the classic approach, i.e. PMBOK or PRINCE2. Every time, project documentation is created before the project launch and completed regularly throughout the project duration. The decision regarding the choice of projects for realization is made by the office in charge of the project portfolio selection, in consultation with the company's management.

Forty-two employees of the company took part in the research. The participants were people involved in the execution of IT projects. The participant structure of the survey is presented in Table 1.

TABLE I.
PARTICIPANT STRUCTURE

Respondent's role in a project	Percentage [%]
Designers	7
Developers	27
Management	18
Implementation Specialists	35
Remaining participants	13

The respondents of the research are people present in all stages of the project life-cycle, which means designing, creating and implementing. A substantial number of people (18%) were represented by management of the project, which means decision-makers, responsible for the possibility of making changes in the scope of the selection of the project realization methods.

The results of the survey research were presented in the form of percentages. In the case of multiple choice questions, the percentage estimates the ratio of people who indicated a particular answer from all of the research respondents.

The first question referred to the evaluation of the effectiveness of the project execution. It required the respondents to list the main problems encountered during the project execution. The obtained results are presented in Fig. 2.

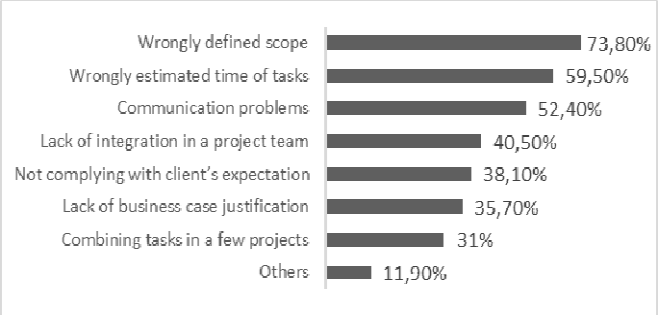


Fig. 2. Problems present in projects

The results of the analysis show that the main problem which occurred in the projects was a wrongly defined scope. This is a particularly essential factor of the project, which, in the event of failure, infringes the classical triple constraint. A significant proportion of the employees also pointed out communication problems. Despite the communication tools available in the company, communication is not satisfactory. This problem may result from an inappropriate information flow, not from a lack of available technical solutions, but from incorrect work organization in a project. This may be caused by the low effectiveness of employees' work or by a bad estimation of the task duration time made by those responsible for project execution.

Another question referred to the identification of the factors influencing the quality of the products or services delivered within the performed projects. The results are shown in Fig. 3.

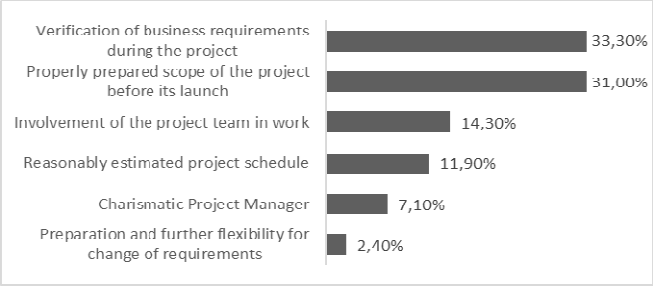


Fig. 3. Factors influencing the quality of products or services delivered

In the respondents' opinion, the biggest influence on the high quality of the product delivered within the project is the possibility of verifying and adjusting the requirements to the business expectations of the company. This is crucial for the changeable environment which IT technologies are. The method of project execution must be flexible enough to be able to adjust to the clients' requirements going beyond the traditional project triple constraint. This factor was probably particularly visible because the leading methodology of conducting projects in the examined company is the classic method concentrating on maintaining the project in a classical triple constraint.

Another question was related to the indication of other factors which according to the respondents' opinion may also influence an increase in the quality of delivered products or services within executed projects. The results are presented in Fig. 4.

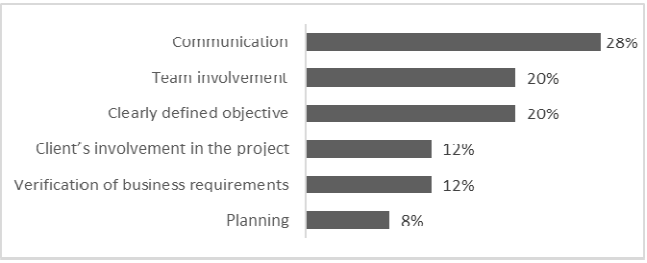


Fig. 4. Factors influencing the quality of delivered products and services – open question

The achieved answers in the majority of cases confirmed problems connected with communication in a project, involvement and willing for cooperation among the participants, a properly defined objective and scope, the verification of requirements and the correct definition of responsibilities in a project. Less frequently, the significance of good planning, work scheduling and solution verification in every stage of project realization were indicated.

The last question required the indication of the preferable methodology of project execution. Thanks to a question formed in such a way, the attitude of the IT project participants towards the possibility of using another (than classic) methodology of project execution could be checked. The results are shown in Fig. 5.

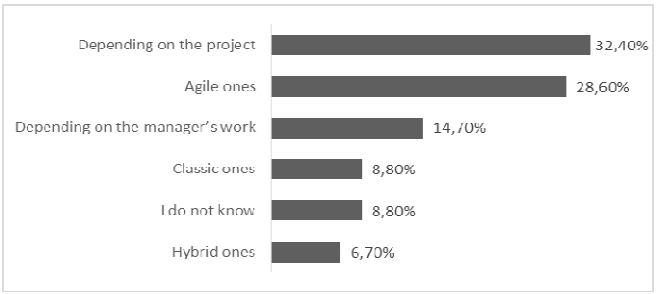


Fig. 5. Preferred methodology of IT project management

The majority of respondents conclude that the best method of project management is a method individually adjusted to the project needs (32,4%). Another group of people support projects run using the SCRUM method, which means performed according to the agile approach (28,6%). Approximately 15% of respondents say that the preferable method of project management depends on the Project Manager's working methods. This means that the method used in the project should undergo changes during its realization under changing project conditions – the hybrid approach. This may imply that the use of the classic approach to project management in a company does not meet the expectations of all of the project participants.

V. Decision Model Development

The conducted surveys allowed key problems to be identified related to the implementation of projects, for which the solution should be the correct choice of IT project management. In connection with the above, in order to solve the identified problems, the usage of a decision model was proposed. This would ease the Project Managers' selection of the best approach to project management (agile, classic or hybrid). The proposed solution should allow for an overview of the project from the perspective of all project participants. It should also allow for its analysis and be an attempt of deviating from the classical triple constraint in the direction of adaptive solutions. Each project is unique and creating one method for all projects may not be efficient enough.

The proposed model is an adaptation of a rhomboidal model [3], which allows for the classification of the project (Fig. 6.)

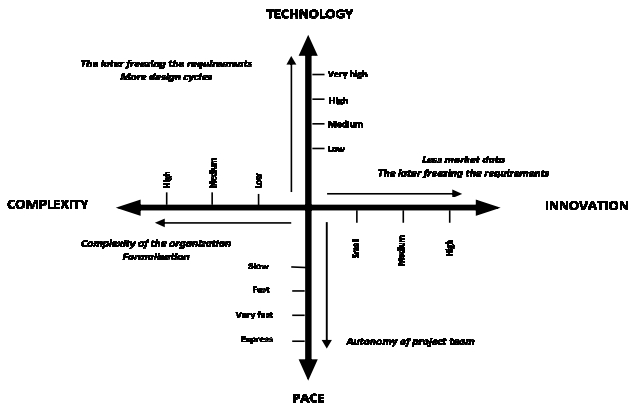


Fig. 6. The rhomboidal model

The presented model covers four basic variables characterizing the project, meaning: [3]

Innovation: the variable estimating the uncertainty of the product or service delivered within the project. It estimates the level in which we are able to evaluate the requirements and needs of further users. The variable is described by three levels of values.

- *Small* – means that the result of the project refers to classic and repetitive products which can possibly expand the products existing on the market,
- *Medium* – means that the result of the project refers to products which are a new series replacing earlier solutions,
- *High* – means that the result of the project refers to totally new projects being innovative solutions.

Technology: the variable estimating technical uncertainty and its influence on the project. It is described by four levels which characterize projects in the following way:

- *Low* – means using well-known technology exclusively,
- *Medium* – means using well-known technology with elements of innovation,
- *High* – means that the majority of used technologies is not well-known yet but is available on the market,
- *Very high* – means that the used technology does not exist yet and the solution will lead to the designated aim using totally new technology.

Complexity: the variable describing the level of difficulty and formalizing the project scope. The variable is described by three levels of values:

- *Low* – means that the product delivered within the project is simple and/or creates a single set of elements,
- *Medium* – means that the product delivered within the project consists of many subsystems influencing each other, and fulfils different functions,
- *High* – means that the product delivered within the project is very complex and is part of a set of distracted systems.

Pace: the variable describing the pressure of time connected with project realization, taking into account the consequences of the non-respect of the project deadline. The variable is described by four levels of values:

- *Slow* – means that the time of making the product or service is not a measurement of the project success,
- *Fast* – means that possibility of faster project accomplishment is an added value and competitive advantage but it does not determine the success of the project,
- *Very fast* – means that the project has a precisely estimated time of accomplishment regarding upcoming developments and its realization in time determines the failure or success of the project,
- *Express* – means that the project is urgent, with no delays accepted and its execution determines the solution of some problem or crisis.

As a result of the conducted analysis of variables in the examined enterprise and the evaluation of the influence of the variables on project quality, a decision model was developed, (Fig. 7). The model indicates certain tendencies for the project variables. Therefore, it is possible to conduct its analysis, which makes the selection of the project management method easier.

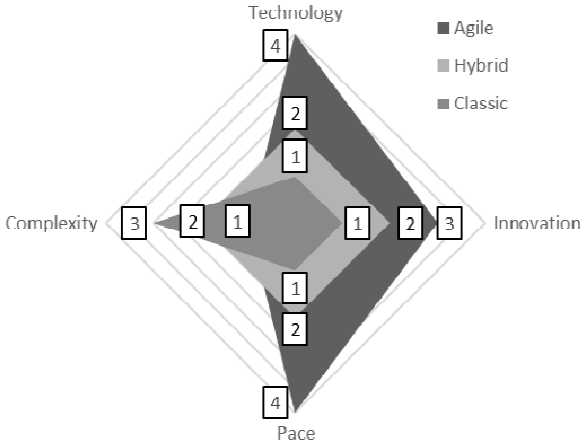


Fig. 7. Rhomboidal model adaptation

The areas where a particular method of management is used were indicated with different colors. In the case of areas not covered with a recommendation, for example, because of high complexity and very advanced technology, the selection of the methodology belongs to the Project Manager.

The more advanced the technology used, the later the freezing of the requirements for the project should take place, because there is more time to change. Therefore, in this case, it is proposed to use the agile approach, which allows for the smooth registration and implementation of changes in particular stages of the project. This is opposed to the situation when the project uses less advanced technology. Often, in such a situation, the requirements are already well-known before launching works. In connection to this, it is suggested to use the classic approach. Similarly, the same is in the case of high innovation. The more innovative the product is and the less the clients' expectations are known, the more agile or hybrid the approach of project management is recommended to be used.

The pace of conducting the project is essential and may be considered in two ways. An additional meeting is a waste of precious time which may be spent on work connected with the direct creation of the product. From the other side, it is time allowing for the earlier verification of the product and the implementation of possible changes. Therefore, in the case of a high pace of project realization, it is recommended to use the agile approach. If time is not so relevant, the classic approach may be used. For a medium pace, the agile or hybrid approach is proposed to be used.

The last issue taken into account in a model is the project complexity, which is a domain of well documented classic methodologies. In the agile approach, the activities are always first and the documents are treated as subsidiary. Therefore, the more complex and documentation demanding

the project is, the more the classic approach is proposed. However, for solutions demanding a medium level of formalizing, the hybrid model is recommended, particularly in the case of huge technical progress and innovation.

VI. Model validation

At the validation stage, the presented model was verified using data coming from the project accomplishment. The data achieved in this way was analyzed using the rhomboidal model. Next, the data was verified taking into account the opinions of Project Managers. This allowed the identification of whether the used model could enable the selection of the correct project execution method.

The analysed project referred to the implementation of a platform for investments and debts services. The project was conducted according to the classic approach (PMBOK) and unfortunately failed.

The innovation of the product was estimated at level 2 (medium). The projected platform was supposed to replace the existing solution by using various channels for information access. It was supposed to be available for the client in a new, friendlier way.

The technology used was evaluated at level 3 (the highest). It means that the product created within the project was very advanced and the majority of technologies were not known in the company. Nevertheless, this technology was already available on the market.

The complexity of the project was evaluated at level 3 (the highest). In the examined organization it was a very complex and demanding project. It required collating a lot of data which often was not presented on a daily basis and had to be taken from clients from different databases.

The pace of the planned work was established at the highest level - 4 (express). The project had to finish fast because the solutions implemented in this company were also used by competitors, which required priority action.

The statement of the above variable values with the use of the rhomboidal model recommended using the hybrid approach for project execution. The project demanded huge autonomy of the project team and was performed at an express pace. It was characterized by late freezing of requirements with a high level of innovation and using advanced technology. Unfortunately, the project was executed according to the classic approach, which meant that it was not realised in a sufficiently flexible way. According to Project Managers, using hybrid methodology would allow for the creation of detailed documentation and agile team management. The project conducted in such a way would involve more frequent verification of requirements in connection with the business. This would imply that the project is accomplished with success.

In the same way, the proposed rhomboid model has been validated on the basis of three completed IT projects. The validation carried out in this way confirmed the effectiveness of its use.

VII. Summary

The presented model is an attempt to support the people responsible for making decisions in the scope of the selection of the project realization method. The proposed solution allows the analysis of key project parameters (complexity, innovation, technology) and eases the selection of the optimal approach for project management (classic, agile, hybrid).

The advantage of the model is an overview from the wider perspective, which means taking into account the factors and indicators surrounding the project. It can be used for supporting the decision-making process regarding the selection of the best project management model.

The disadvantage of the model may be the necessity of conducting an additional analysis, which is time-consuming. Thus, as results from research, the efficient selection of the project management method is the factor which increases the probability of accomplishing the project with success.

The possibility of model adaptation in organizations is wide. It should be remembered that the problem of the examined enterprise was the fact that all the projects finished with apparent success. The reason for such a situation was the fact that the end user was an internal client. It was the internal client who in every case reported the demand for the product or service and defined the requirements. All project changes (scope, duration, etc.) were realized with the acceptance of the internal client and from a formal point of view, the project was always accomplished with success. In a longer perspective, the project was not successful. It was characterized by low effectiveness and a low client satisfaction level. The presented approach in the examined company was not typical. Therefore, the possible limits and doubts should be verified in the process of further research.

References

[1] Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2017). *Agile software development methods: Review and analysis. arXiv preprint arXiv:1709.08439*.

[2] Kerzner, Harold, and Harold R. Kerzner. *Project management: a systems approach to planning, scheduling, and controlling*. John Wiley & Sons, 2017.

[3] Kuchta, D., Skowron D. (2014). Model romboidalny w zarządzaniu projektami badawczo-rozwojowymi. *Zeszyty Naukowe/Wyższa Szkoła Oficerska Wojsk Lądowych im. gen. T. Kościuszki*, (1), 201-220.

[4] Kaczorowska, A. (2015). Traditional and agile project management in public sector and ICT. In *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on* (pp. 1521-1531). IEEE.

[5] Lechler, T. G., Edington, B. H., & Gao, T. (2012). Challenging classic project management: Turning project uncertainties into business opportunities. *Project Management Journal*, *43*(6), 59-69.

[6] Redlarski K., *The impact of end-user participation in IT projects on product usability*, ACM, 2013.

[7] Redlarski, K., & Weichbroth, P. (2016). Hard lessons learned: delivering usability in IT projects. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on* (pp. 1379-1382). IEEE.

[8] Wells, H., Dalcher, D., & Smyth, H. (2015). *The adoption of agile management practices in a traditional project environment*: An IT/IS Case Study. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 4446-4453). IEEE.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 985–994
DOI: 10.15439/2018F100
ISSN 2300-5963 ACSIS, Vol. 15

The Role of a Software Product Manager in Various Business Environments

Olga Springer, Jakub Miler
Gdansk University of Technology
Faculty of Electronics, Telecommunications and Informatics
11/12 Narutowicza St., 80-233, Gdansk, Poland
Email: {olga.springer, jakub.miler}@pg.edu.pl

***Abstract*— The aim of the paper is to identify the role of the software product manager depending on the size of the company and the characteristics of the product they are working on. This has been achieved in cooperation with 15 experts from the IT industry. The companies were divided into 4 levels of size: micro-enterprises, small businesses, medium businesses and large enterprises. The characteristics of the products were divided into business-business (B2B) and business-customer (B2C). This way, 8 personas of software product managers have been developed. Differences in this role were mainly related to the staffing, its scope of responsibility, tools and techniques used as well as the mode of work with the target customers. Many common aspects of this role have also been identified that made it possible to define archetype persona of a software product manager. All personas have been validated by experts who offered their improvements.**

I. Introduction

MARKETING definition of "a product" is: "a product is anything that can be offered to a market for attention, acquisition, or consumption that might satisfy a want or need" [1]. Since 2007, it is known that systematic product management increases the success rate of software projects [2]. Product management responsibilities are crucial in software companies, as they support decision making process and developing products according to the company strategy. By definition, product management is different from project management, however in many software companies the roles of product manager and project manager are mixed [1].

Similarly to project managers, the product managers must have high level management skills, as very often they lead teams, projects or even departments [3]. Communications skills are important in order to efficiently manage stakeholders and cooperate with development teams to execute product roadmap [1]. The approach to product management often follows the principles of lean management as in the Lean Start-up method [4]. Product managers design and verify the business models in the market [5] with various techniques and tools [6].

This work was supported by DS Funds of ETI Faculty, Gdansk University of Technology.

Product management spans the entire product-life cycle: strategy, concept, market entry and development, evolution [7]. Product managers do not only follow the product-life cycle, they own it. They are responsible for the success of the product, which require them to take care of wide technical and business activities: product strategy, product planning, strategic management, orchestration of the organization's functional areas [1]. It is difficult to start career as a product manager without experience in adjacent disciplines, e.g. marketing or development [8].

There is a role defined in Scrum that is responsible for requirements and business value – the Product Owner [9]. Spectrum of activities and responsibilities of the product manager is much wider than the Product Owner role [10]. Depending on the organization's size, the product manager and the Product Owner roles may be separated or a product manager can assume the Product Owner role, however "in most environments, it makes more sense to have the two roles product owner and product manager separated" [1]. Additionally, earlier research has shown that the approach to projects is related to the business environment [11], [12].

The goal of our research was to understand the current role of product managers in software companies. To achieve this goal, 3 research questions were stated: (RQ1) What is the software product manager's role depending on the company size and characteristics of the product? (RQ2) What are the differences in the role of a software product manager depending on the company size and product type? (RQ3) What are the common characteristics of a software product manager independent of the company size and product type?

The contribution of this paper is the identification of the characteristics of the software product manager's role in different business environments as well as the elaboration of the archetype of this role based on the elements common to all environments.

The paper is organized as follows. Section II discusses the work of other researchers to which we relate our study. Section III describes the research method, the experts and techniques. Section IV provides the personas of product managers specific to different business environments as well

as the archetype common to all environments. Section V presents the two phase validation of the proposed personas and archetype. Section VI discusses the threats to the validity of this research followed by the conclusions in Section VII.

II. Related Work

In 2013, research on software product management was conducted in Lappenranta University of Technology, and its main research question was: which common roles do software product manager fulfill in the organizations? The result of the research is a framework that reveals the role of product managers. Four stereotypical roles were identified in the studied organizations: experts, strategists, leaders and problem solvers [8].

An expert is a person who has a deep expertise in some specialization (e.g. marketing or development), but does not have the responsibility for product management activities. It is a kind of informal product manager, who works close to product and knows it very well. He is involved in the implementation to decide on feature priorities. The strategist is a product manager who has an impact on the company product vision and roadmap. As the authority of the strategist grows, he becomes the third profile, the leader, who has access to product resources and can manage it. The fourth profile is the problem solver who manages stakeholders and resolves product-related issues, but it is the management who defines the strategy and roadmaps [8].

The framework was developed empirically based on interviews with companies' representatives and by researching the documentation. Besides the framework mentioned, super categories with properties were also identified based on the grounded theory [21] analysis:

- access to resources – ownership of the product budget, possibility to hire people, information resources,
- influence on the product – orchestration of development, definition of tactical actions, participation in strategy planning, creation of roadmaps,
- authority – product leadership, power of decision making,
- influence on collaboration – ability to resolve problems between departments, level of communication.

In 2014, Christof Ebert and Sjaak Brinkkemper defined product management key success factors, their effect on business and challenges that stand ahead of the product management [13]. They also described 4 best practices for systematic product management. They stated that during the research the majority of product managers that have been found in software companies had: "a strong technical background and rather weak finance, marketing and general management skills" [13].

"Software Product Management" book, published in 2017 by H. Bernand Kittlaus and Samuel A. Fricker [1] is a review of software product management and many terms related to it like: management of software as a business, product strategy, product planning, strategic management, orchestration of the organization's functional areas. It also introduces the Software Product Management Framework, which is an integration and consolidation of three other frameworks that were invented:

- Reference Framework for Software Product Management developed by Inge van de Weerd, Willem Bekkers, Sjaak Brinkkemper, and colleagues at the University of Utrecht, Netherlands in 2006 [14]. This framework describes the core activities of a software product manager in specific areas: portfolio management, product planning, release planning, requirements management.
- Pragmatic Marketing Framework developed by Kittlaus and Clough [15] that describes the aspects of product management, product marketing and defines the role a of product manager and product marketing manager ultimately.
- a paper by Ebert on the impacts of software product management [2], a proposal on how to manage software along the product lifecycle phases.

Software Product Management Framework indicates the main functional areas of a software organization: strategic management, product strategy, product planning, development, marketing, sales and distribution, service and support. Each area has related activities assigned.

Product management is gaining popularity every year, however it has been topic for a scientific research only few times in the recent years. Of 3 researches on software product management described above only one of them is focusing strictly on the product manager's role, however it does not analyze if this role depends on the company size and product type.

III. Research Method

Our research was carried out as a qualitative study using grounded theory [21] as a research method. The identification of the role of the software product manager has been achieved in cooperation with 10 experts with more than 2 years of experience in research and business projects in the IT industry. The experts worked as:

- a product manager,
- an employee with product management responsibilities but named differently,
- an employee working in close cooperation with a product manager.

The companies were divided into 4 levels of size (ranged in size of number of employees): micro-enterprises (1-10), small businesses (10-50), medium businesses (50-250) and large enterprises (250+) [16]. The characteristics of the products were divided according to the product delivery model into business-business (B2B) and business-customer (B2C) [17]. Combining these dimensions, we analyzed the software product manager's role in 8 different environments. The characteristics of the experts are presented in Table I.

TABLE I.
CHARACTERISTICS OF THE INTERVIEWED EXPERTS

Identifier	Position	Company	Size	Delivery model
E1	Vicepresident of Projects	UXpin	medium, large	B2B, B2B, B2C
E2	Vicepresident	Rocket Studio	large, large	B2B, B2C
E3	Product manager	Spartez	medium, large	B2C, B2B
E4	President	RoomAuction .com Ltd.	micro, small	B2B, B2C
E5	Product manager	Better Solutions	small	B2B
E6	Business analyst	Net PC	small	B2B
E7	Product manager	AirHelp	medium	B2C
E8	Project manager	dr Poket	micro	B2C
E9	Product manager	Young Digital Planet	large	B2C
E10	Product manager	AirHelp	micro, small	B2C, B2C

We have used the technique of a persona [18] to describe the identified roles of a software product manager. This technique aims at building a profile of a representative member of a particular group of people. Such profile may include various attributes. Based on our research goals and the literature we have selected a set of attributes of the software product manager's role and taken them as the model of our personas. This model with the description of all the attributes is presented in Table II. Further in the research, this model was used to define 8 personas of product managers from various business environments.

The main instrument of the data collection were the structured interviews with experts conducted in the form of face to face meetings in May 2016. We have asked the experts the following questions based on the persona model:

- What is your experience in product management area?
- How many years you have been working as product manager or as related role?
- Who else was responsible for product management?
- What were the goals and responsibilities of product managers?
- What was the product lifecycle?
- Who did you cooperate with?
- What techniques did you use?
- What tools did you use?
- What are the first steps to become a product managers? How in your opinion young people interested in this path of career can get first experience?

TABLE II.
A MODEL OF A SOFTWARE PRODUCT MANAGER'S PERSONA

Attribute	Description
Introduction to the environment	The specificity of the environment and its approach to product management.
Objectives	The objectives of the Product Manager or the role with his responsibilities, depending on the size of the organization.
Responsibilities	The responsibilities of the Product Manager's role, depending on the size of the organization.
Product Life Cycle	The development phases of the product, on which the approach to project management depends.
Main competences	The required competences of product managers including soft and hard skills.
Cooperation with other teams	Teams and roles that product manager cooperates with.
Roles in the product team	The location of the product team within the organization and its characteristics.
Techniques	List of techniques used by the product manager including techniques for: verification of the product vision and strategy, product delivery management, user research.
Tools	List of tools used by the product manager to: manage tasks and backlog, user research, store documentation, prototype and collaborate with other roles.
First steps	Career opportunities as a product manager and specific steps that can be taken to become a product manager.
Other positions	Other positions or roles responsible for product management.

Each interview lasted from 30 to 60 minutes and was recorded, analyzed, and documented. The recordings allowed to fill up the model properly with the data. The templates for the interviews were filled by the researcher, however later on they were validated by every expert (see section IV). Many interviewed experts provided insights to more than one business environment as a result of their experience. The coverage of the analyzed business environments with the interviews is presented in Table III.

TABLE III.
NUMBER OF INTERVIEWS FOR EACH COMPANY SIZE AND DELIVERY MODEL

Size / Delivery model	B2B	B2C
micro	1	2
small	2	2
medium	1	2
large	3	3

Additionally, we aimed at the identification of the most inherent aspects of the software product manager's role which seem independent from the business environment. To capture this viewpoint we employed the concept of an archetype which is a common recurring motif or pattern assumed to represent fundamental characteristics of a thing [19]. We have built an archetype of a software product

manager which can be viewed as a super-persona, a more general or abstract persona than the personas specific to different business environments (in object-oriented modeling the archetype would be a superclass of the classes of specific personas, hence the name super-persona).

The software product manager's archetype was built using the following rules: (1) include common elements that repeat in each specific persona; (2) find a classification or a generalization of elements that do not repeat directly in the specific personas; (3) include the classes or generalization of elements identified by rule 2. Considering each specific persona as a set of values of the attributes, the resulting archetype can be considered an intersection of these personas focusing on their commonalities.

TABLE IV.
SOFTWARE PRODUCT MANAGER IN A MICRO-ENTERPRISE, B2C

Attribute	Description
Introduction to the environment	Responsibility for the product management and all tasks associated with the role of product manager are handled by the company's founders. Generally there is no strict product manager position.
Objectives	To recognize customer needs and develop a valid business model
Responsibilities	Project management, prototype development (MVP), team-building, tasks related to marketing and sales, gathering users' feedback, fund gathering, resource management.
Product Life Cycle	Idea, business model verification, introducing functional prototype to the market, sales and product development, maturity, decline. Different approach to product management depending on the phase.
Main competences	Understanding of the problem domain, cooperation skills, openness, communication skills, entrepreneurship, visionary, priority management.
Cooperation with other teams	Mutual cooperation of the company's founders (partners), cooperation with the development team (internal or external), cooperation with mentors and investors, cooperation with external entities (i.e. design agency)
Roles in the product team	Multidisciplinary team of several people with different competences.
Techniques	Business Model Canvas, Product Roadmap, Product Backlog, Agile techniques, Personas, Product Map, Statistical analytics, Workshops with end users, Usability labs, Prototyping
Tools	Podio, Trello, Mixpanel (free version), Google Analitycs, Microsoft Office, Physical board, Posts-its
First steps	Learn the company, the product specifics and the users' needs e.g. while working in customer support team or at a technical position.
Other positions	Chief Executive Officer (CEO), Analyst, Project Manager, Developer

IV. RESULTS

This section presents the proposed personas of software product managers depending on the business environment (the organization size and the product delivery model), initial observations on the differences of the product manager's role among business environments as well as the final super-persona of an archetypical software product manager. The section follows the increasing company size. Sections A, B, C, and D are related to RQ1, section E to RQ2 and section F to RQ3.

A. Micro-enterprises

The personas of a software product manager in a micro-enterprise are presented in Table IV (B2C delivery model) and Table V (B2B delivery model). It could be observed that in the companies of this size the product manager's role is not assigned to a distinct post but is rather fulfilled by the company owner or the CEO.

TABLE V.
SOFTWARE PRODUCT MANAGER IN A MICRO-ENTERPRISE, B2B

Attribute	Description
Introduction to the environment	Responsibility for the product management and all tasks associated with the role of product manager are handled by the company's founders. Generally there is no strict product manager position.
Objectives	To recognize customer needs and develop a valid business model
Responsibilities	Project management, creating the product strategy, setting the short-term goals, coordinating work to achieve the goals, market research, tasks related to marketing and sales, acquiring knowledge from experienced "business sharks", establishing relationships and networking with prospecting clients, gathering customer feedback, business analytics, close cooperation with the founders.
Product Life Cycle	Idea, business model verification, introducing functional prototype to the market, sales and product development. Different approach to product management depending on the phase.
Main competences	Consistency and persistence, data analysis and synthesis, understanding customer needs, soft skills, flexibility, inquisitiveness, openness and the ability to establish business contacts.
Cooperation with other teams	Mutual cooperation of the company's founders (partners), cooperation with the development team (internal or external), cooperation with mentors and investors.
Roles in the product team	Multidisciplinary team of several people with different competences.
Techniques	Business Model Canvas, Product Roadmap, Product Backlog, Agile techniques, Personas, Workshops with clients, Usability tests, Web analytics,
Tools	Asana, MindMaster, Trello, Slack, Skype, MS Office, Physical board, Posts-its
First steps	Read a lot, create something of your own or start a career in technical positions.
Other positions	Chief Executive Officer (CEO), founders, president

B. Small enterprises

The personas of a software product manager in a small enterprise are presented in Table VI (B2C delivery model) and Table VII (B2B delivery model). It could be observed that the companies of this size start to hire the product manager as a distinct post.

TABLE VI.
SOFTWARE PRODUCT MANAGER IN A SMALL ENTERPRISE, B2C

Attribute	Description
Introduction to the environment	Responsibility for the product management is on the founders of the company. At some point, the Product Manager is hired to support the founders.
Objectives	Co-creation and execution of the product strategy, support in defining the company's business goals.
Responsibilities	Product development, incremental product delivery, negotiation of priorities with the stakeholders, prototyping, cooperation with external suppliers, solving users' problems.
Product Life Cycle	Idea, introduction to the market, growth, maturity, decline. Very high work dynamics in the phases of introduction and growth. Different approach to product management depending on the phase.
Main competences	Project management, data analysis, ability to work in a dynamic environment, communication, negotiations, empathy.
Cooperation with other teams	Cooperation with the founders, project manager, customer support, and development team.
Roles in the product team	A multidisciplinary team: product manager, project manager, graphic/UX designer.
Techniques	NPS (net promoter score), product roadmap, product backlog, agile techniques, user stories, interviews, prototyping.
Tools	Trello, Google Analitycs, Mixpanel, Skype, Axure, Gdrive
First steps	Begin a career in the development environment, for example as a project manager, tester, or developer. Observe the product development process and get to know the product. Observe and learn from experienced product managers.
Other positions	Chief Executive Officer (CEO), founders, president, Chief Technology Officer (CTO), Chief Marketing Officer (CMO), project manager.

TABLE VII.
SOFTWARE PRODUCT MANAGER IN A SMALL ENTERPRISE, B2B

Attribute	Description
Introduction to the environment	There is or there is not a defined role of the Product Manager depending on the relationship with the client. In the latter case, the roles that share the responsibility for product management are Project Manager and Business Analyst.
Objectives	Co-creation and implementation of the business goals of the company, delivering the product to the customer according to the requirements.
Responsibilities	Defining, clarifying and specifying the requirements, contacting the client, prototyping of interfaces, prioritizing tasks according to the client's needs, proposing solutions, building customer relations, examining customer needs.
Product Life Cycle	If company offer universal product for all b2b customers: Idea, introduction to the market, growth, maturity, decline. If company offers custom solutions for enterprises: Tender specification, cost valuation stage, bidding, product implementation (or its development, maintenance), closing the project. Permanent cooperation with the client is also possible. Different approach to product management depending on the phase.
Main competences	Analytical thinking, accuracy, project management, communication skills, decision-making, negotiations, empathy and understanding of customer needs, forming business relationships.
Cooperation with other teams	Cooperation with project manager, business analyst customer support, and designers
Roles in the product team	A multidisciplinary team: product manager, project manager, business analyst, graphic/UX designer.
Techniques	Product backlog, product roadmap, use cases, wireframes, UML diagrams, interviews, workshops with clients, project plan, requirements specification.
Tools	Jira, Mantis, Axure, Proto.io, UXPin, Balsamiq, Enterprise Architect.
First steps	Read about project management and analysis. Begin a career in the position of: analyst, project manager, programmer, tester.
Other positions	Project manager, business analyst.

C. Medium enterprises

The personas of a software product manager in a medium enterprise are presented in Table VIII (B2C delivery model) and Table IX (B2B delivery model). It could be observed that the companies of this size build product management teams around the product manager.

TABLE VIII.
SOFTWARE PRODUCT MANAGER IN A MEDIUM ENTERPRISE, B2C

Attribute	Description
Introduction to the environment	Software companies aware of the need for product management. Defined role of the Product Manager.
Objectives	Defining the product strategy in line with the company's business goals and its implementation.
Responsibilities	Defining the way to achieve company's business objectives, confirming the expectations of the stakeholders, communication of results to the stakeholders, gathering requirements, defining needs or problems, analyzing numerical data, user research, refining and describing initiatives, projects and tasks, confirming priorities with stakeholders, coordination of project implementation, communication of initiatives and reporting of achieved goals after implementation, cooperation with development team - daily communication in order to specify requirements, participation in Scrum meetings (primarily in planning meetings).
Product Life Cycle	Idea, Introduction, Growth, Maturity, Decline. Different approach to product management depending on the phase.
Main competences	Communicativeness, openness, assertiveness, ability to create visions, analytical mind, understanding of users' needs.
Cooperation with other teams	Cooperation with: designers, developers, testers, stakeholders (leaders of other teams i.e. customer service or sales), management board, other product managers (often goals and initiatives are shared).
Roles in the product team	The product team consists of: product managers, designers, user research specialists, conversion specialists, and translation specialist (multilingual support for the product). Some competencies may be outsourced.

Techniques	Product roadmap, user stories, personas, user research, interviews, wireframes, A/B tests, NPS (Net Promoter Score), workshops with stakeholders, story mapping.
Tools	Jira, Confluence, Slack, Skype, Proto.io, HotJar, Google Analitycs, Mixpanel.
First steps	Participation in workshops, presentations, trainings on product management and project management, individual learning, online courses. Membership in non-profit organizations to develop soft skills recommended. Career can be started at another job position to learn the product and then the change of the role might be possible.
Other positions	Product Owner, brand manager, project manager, business analyst.

TABLE IX.
SOFTWARE PRODUCT MANAGER IN A MEDIUM ENTERPRISE, B2B

Attribute	Description
Introduction to the environment	Software companies aware of the need for product management. Defined role of the Product Manager.
Objectives	Defining the product strategy for business customers, in line with the company's business goals.
Responsibilities	Responsibility for many aspects, both technical and business, defining product strategy, market analysis, researching customer needs, creating a product roadmap, defining product goals, working close with main executives.
Product Life Cycle	Idea, Introduction, Growth, Maturity, Decline. Different approach to product management depending on the phase.
Main competences	Curiosity of the world, data-based analysis, openness, empathy, understanding people, leadership skills.
Cooperation with other teams	Cooperation with all departments in the company: sales, marketing, business development, customer support, project managers.
Roles in the product team	Product managers are gathered in one operational team together with designers and user researchers.
Techniques	Product roadmap, user research, interviews with prospects and clients, user recordings, A/B tests.
Tools	ProdPad, Target Process, Microsoft Office, Keynote, UXPin.
First steps	Start a career in the software company as developer, tester, customer support specialist, project manager. Start your own startup.
Other positions	Designer, tech lead, project manager.

D. Large enterprises

The personas of a software product manager in a large enterprise are presented in Table X (B2C delivery model) and Table XI (B2B delivery model). It could be observed that the companies of this size can employ multiple product managers on the per project or per department basis.

TABLE X.
SOFTWARE PRODUCT MANAGER IN A LARGE ENTERPRISE, B2C

Attribute	Description
Introduction to the environment	In many large enterprises, the role of product manager is taken by persons from departments within which a new project is initiated or a product is developed. The product management is not their only responsibility. They also perform other duties resulting from the specificity of their positions. Only some software companies (mature, product-driven and agile) employ dedicated product managers who define the product strategy and roadmap.
Objectives	Defining a product strategy in accordance with the company's business strategy, achieving goals.
Responsibilities	Defining the scope, contact with stakeholders, collecting requirements and opinions, receiving product increments and accepting the work done by the development team, reporting to the steering committee, cooperation with the development team (sometimes this responsibility is transferred to the project manager), presenting the product inside the company, acquiring the budget for specific actions, data analysis, cooperation with roles with different competences (analyzes, trends, interfaces).
Product Life Cycle	A product is created in a project (initiation, planning, development, closing phases) or along with the product idea a project is launched to verify its business value (concept phase, running an experiment and gathering feedback from the market, implementation of a functional prototype, sales and development). Different approach to product management depending on the phase."
Main competences	Leadership skills, impact on people despite the lack of authority, negotiation skills, assertiveness, ability to prioritize and make decision, business knowledge, creativity, listening to others, taking criticism, teamwork skills, useful technical knowledge, presentation skills, knowledge about techniques and tools.
Cooperation with other teams	Cooperation with: development team, project manager, steering committee, project sponsor, stakeholders, directors of sales, marketing, and customer support departments. Cooperation with subcontractors. Horizontal and vertical communication (reporting to the board).
Roles in the product team	Depends on the organization structure. Mainly multidisciplinary teams working on one product, one product is one department or division. As part of the operational team: product manager, project manager, research specialists, UX experts. Rarely a dedicated product team, possibly in mature technology companies that have grown up on product-driven work (e.g. Atlassian).
Techniques	Product roadmap, product backlog, Business Model Canvas, personas, user research, MindMap, goRoadmap, interviews and surveys, market trends analysis, statistical analysis, predictions, correlations.
Tools	Google Analytics, Mixpanel, Jira, Jira Portfolio, Asana, Redmine, Trello, Sharepoint, Confluence, MS Project, Microsoft Office, Intercom, HotJar, SurveyMonkey.
First steps	Learn about techniques, tools, good interface design practices, learn systematic work e.g. by running a project. You can start your career as: programmer, tester, UX designer, analyst, or starting your own startup. You can also hire as a product manager's assistant and train under his supervision. Develop communication skills and work in a group, e.g. by leading non-profit projects.
Other positions	Project manager, product development manager, product owner, brand manager, marketing director, business analyst (in case the product is an in-house system)

TABLE XI.
SOFTWARE PRODUCT MANAGER IN A LARGE ENTERPRISE, B2B

Attribute	Description
Introduction to the environment	In many large enterprises, the role of product manager is taken by persons from departments within which a new project is initiated or a product is developed. The product management is not their only responsibility. They also perform other duties resulting from the specificity of their positions. Only some software companies (mature, product-driven and agile) employ dedicated product managers who define the product strategy.
Objectives	Collection of requirements from the customer and delivery of the ordered product. The Product Manager the most often has no impact on the company objectives and product strategy, implements elements of the company's overall strategy.
Responsibilities	Contact with the client, collecting requirements and opinions from the clients, cost estimation, managing the budget, ordering additional resources (e.g. usability research), proposing solutions, planning team work, supervising the implementation, presentation of product increments to the client, synthesis of a large number of development threads. Responsibilities depends on how product manager role is set up - sometimes product manager doesn't own budget and doesn't work on any cost estimations.
Product Life Cycle	If company offer universal product for all b2b customers: Idea, Introduction, Growth, Maturity, Decline. If company offers custom solutions for enterprises: Signing a contract for the creation of a new product or contact for the development of an existing system. Initiation, planning, production, closing the project. Different approach to product management depending on the phase.
Main competences	Communication skills, listening to others, data analysis and synthesis, interdisciplinary.
Cooperation with other teams	Cooperation with: project manager, development team, other departments (e.g. to allocate human resources to the project).
Roles in the product team	Depends on the organization structure. Mainly multidisciplinary teams working on one product, one product is one department or division. As part of the operational team: product manager, project manager, research specialists, UX experts. Rarely a dedicated product team, possibly in mature technology companies that have grown up on product-driven work (e.g. Atlassian).
Techniques	Product roadmap, product backlog, workshops with clients, requirements templates, requirements specification.
Tools	Jira, Confluence, Microsoft Office.
First steps	Acquiring broad knowledge from various fields: marketing, sales, UX, project management. Starting a career in positions: project manager, analyst, programmer, tester, architect.
Other positions	Project manager, product development manager, product owner, brand manager, marketing director, business analyst (in case the product is an in-house system)

E. Differences among business environments

Our research suggests several differences in the approach to product management and the role of the software product manager depending on the size of the organization. Different roles are responsible for product management in the early stage, the responsibilities are spread between founders and the core team. As the company grows it is more often when dedicated software product manager appear on board. It can also be seen that large companies use more advanced tools and techniques, and often specialized competences are outsourced. Additionally, the impact of the software product manager on the company's vision and business objectives seems to decrease along with the size of the organization.

Roles in the product team also depend on the company size. In micro-enterprises there is a multidisciplinary team of several people with different competences. Responsibility for the product management and all tasks associated with the role of a product manager are handled by the company's founders. Generally there is no strict software product manager position. In small-businesses there is a team of product manager, project manager, UX designer and sometimes other roles like business analyst (B2B). When company grows and becomes a medium business it invests more in the product team and hires even more specialists such as conversion specialist, translation specialist. In large enterprises the roles in the product team depend on the organization structure.

The product characteristics seem also to influence the role of the software product manager at some point. For example, working in the B2B model requires negotiations with the client, as well as creating additional documentation specifically for business clients. There are some specific techniques that software product managers in the B2B model use i.e. workshops with clients. On the other side, products in the B2C model require understanding of the users, their segmentation, source of origin, and their behavior in order to adapt the product to the target group with the highest business potential.

Regarding tools that software product managers use no differences have been identified when we compare business environments, however there are some techniques and tools that can be used only when the number of customers reaches certain point. For example, our research shows that A/B testing is not used by product managers in micro-enterprises.

The product lifecycle seems to be very similar in different environments, however there are some exceptions e.g. the standard product lifecycle in B2B companies offering custom solutions for enterprises is affected by tender specification, cost of the evaluation stage, bidding, custom product implementations. In large B2C companies the product can be created as a result of an internal project or along with the product idea a project can be launched to verify its business value.

F. Archetype of a software product manager

Table XII presents the archetype of a software product manager based on the product managers' personas specific to different business environments.

TABLE XII.
ARCHETYPE OF A SOFTWARE PRODUCT MANAGER

Attribute	Description
Objectives	Achieving goals by implementing the product strategy and consistent product vision.
Responsibilities	Defining goals, proposing solutions, prioritizing projects or tasks, user research, analysis of requirements, market analysis, stakeholder management, cooperation with the development team.
Main competences	Soft skills: communication, negotiations, teamwork, decision-making, curiosity of the world, open-mindedness, assertiveness, understanding of human behavior, inquisitiveness, networking, leadership predispositions, consistency and perseverance. Hard skills: ability to understand the problem domain, data analysis and synthesis, knowledge of business analysis and project management, interface prototyping, willingness to learn.
Cooperation with other teams	Cooperate with all product stakeholders and development team.
Techniques	Techniques that support: verifying a product vision and strategy, product delivery, user research.
Tools	Tools that support: task and backlog management, data analysis, user research, documentation, prototyping, remote cooperation.

V. VALIDATION

The first phase of the validation began on June 7, 2016. It has started by sending an email to 10 experts (the same who participated in the interviews) with a link to the document containing the initial versions of the personas. Until 5 July 2016, the experts expressed their opinions on the personas and suggested changes in their presentation and content. In total, we have received 33 suggestions of corrections and remarks on the misinterpreted or incomplete content. They were all taken into account to build the second version of the personas.

The second phase of the validation took place in 2018 and began on January 25th and ended on February 21st. The validation method was the same as in the first phase but this time we engaged new experts who didn't participate in the interviews and the first validation phase. The characteristics of the experts of the second validation phase are presented in Table XIII.

New experts provided 29 remarks in the second validation phase. They did not challenge the personas but rather suggested possible extensions, more competencies, techniques, and tools. The final personas presented in this paper include all the expert's remarks.

TABLE XIII.
CHARACTERISTICS OF EXPERTS IN THE SECOND VALIDATION PHASE

Identifier	Position	Company	Size	Delivery model
E11	Technical Product Manager	Dynatrace	large	B2B
E12	Product Manager	Zalando SE	medium, large	B2C, B2C
E13	CEO	Cux.io	micro	B2B
E14	Senior Product Owner	STX Net	micro, medium, large	B2B, B2C, B2B
E15	Product Manager	Smartly.io	micro, medium, large	B2B, B2C

In total, we involved 15 experts who provided input and comments to the proposed personas of software product managers. The coverage of particular business environments with the experts involved in the validation is presented in Table XIV.

TABLE XIV.
NUMBER OF EXPERTS FOR EACH COMPANY SIZE AND DELIVERY MODEL IN THE VALIDATION PHASE

Size / Delivery model	B2B	B2C
micro	4	2
small	2	2
medium	1	5
large	5	5

VI. THREATS TO VALIDITY

A. Threats to construct and internal validity

We have identified the following threats to the construct and internal validity of our research: (a) interviewer's bias and impact on interviewees, (b) misinterpretation of the interviews with experts, (c) invalid model of the software product manager's persona, (d) subjective construction of the software product manager's archetype.

We have controlled the interviewer's bias and their impact on experts with the semi-closed structured interview design. Each expert was asked the same set of predefined open-ended questions with no suggested answer. The interviewer was open to further expert's explanations. For details on the interview design see section III of the paper.

The incorrect interpretation of the output from experts was controlled by recording each interview, listening to it several times while extracting data to the personas as well as further validation of the personas with the same experts. Experts verified if the personas correctly represented their statements. For more details on the validation of personas see section V.

Our model of the software product manager's persona covers the aspects of this role we assumed important and worth studying. The technique of a persona allows to use the

set of attributes which suit the modeling goal. The selected attributes allow for the generalization of the expert's output as well as cover different aspects of the role. More specific aspects can be studied in further research.

We have built the software product manager's archetype using only the data from the personas specific to different business environments. No raw data was added directly to the archetype. We have identified the common elements in all specific personas and put them into the archetype, adding some classification and generalization when appropriate. The archetype was validated by the experts, however it should still be considered a preliminary result of this research. We plan to work on improvements, verification and application of this archetype in the future.

B. Threats to external validity

We have identified the following threats to the external validity of our research: (a) low number of experts, (b) experts selected as a convenience sample, (c) experts sample limited mostly to Polish market.

The first two threats result from the fact that the software product management is a relatively new topic both in research and in practice, therefore access to experienced product managers in IT is naturally limited. We have mitigated these threats by having our experts experienced in several business environments from working in different companies, experiencing the growth of a company (e.g. from small to medium) or working on products with different business models (i.e. both B2B and B2C). As a result, most of the business environments were covered by at least two experts (for details see Tables III and XIV). Furthermore, all experts took part in the first validation phase, where they reviewed all personas, not only those they provided input to. Additional 5 experts from various environment reviewed all personas in the second validation phase, where they mostly proposed some minor extensions (for details see section V).

The experts were identified and contacted with various channels such as personal contacts, business contacts, subscribers and readers of the Product Vision blog [20], and recommendations from identified experts. This method provided for a fairly diverse group of experts with different background and experience (see Tables I and XIII).

We have based our research on data from experts working in the Polish market (apart from one foreign expert involved in the second validation phase). This forms a natural limitation to our current results. The specifics of the product manager's role in different markets, if any, is yet to be investigated. We plan to engage more experts from abroad in our future work.

VII. Conclusion

The aim of this research was to analyze the approach to software product management and the role of the software product manager depending on the size of the organization and the characteristics of the projects.

We have interviewed 10 experts from various IT companies on the role of a software product manager as well as involved a total of 15 experts in the validation of the results. Based on this work we have proposed the personas of a software product manager for 8 different business environments that answer the research question RQ1.

Then we have identified some of the biggest differences between the software product manager's roles in various environments. The key differences are related to the staffing of this role, its scope of responsibility, tools and techniques used as well as the mode of work with the target customers. This partially answers the research question RQ2 and suggests the directions of further research.

Our research also indicated some common elements independent of the area which helped us to create an archetype of the software product manager. The important skills of the product manager are highly developed soft skills, basic hard skills, knowledge of techniques and tools, and empathy towards the customer or user. The product lifecycle seems to be very similar in different environments. The product can be in one of four phases of life and depending on where it is, the product manager must manage its development. In general, the software product manager is responsible for implementing the company's strategy and its business goals. The software product manager's archetype forms the preliminary answer to the research question RQ3.

Additionally, our analysis has revealed the problem of understanding the role of software product manager and common lack of knowledge related to software product management. Interviews with experts have confirmed that people who in fact deal with the software product management may have different job titles (i.e. project manager, function manager, brand manager, product delivery manager), which can result in their lack of awareness of their duties and responsibilities. This may translate later on into making wrong business decisions that have a negative impact on the development of entrepreneurship, companies and products. Also, the relation of the software product manager and the Product Owner and the collaboration of the product manager with agile teams [22], [23] needs further research.

The proposed personas may be used as a guidance and learning targets for software product managers and the top-level management, where the possible practical applications include: (1) education and training of software product managers both in the industry and academia; (2) self-development of the current software product managers, in particular seeking to switch the business environment; (3) support for the managers of the organizations planning to hire software product managers or build a software product management team; (4) increasing the awareness and understanding of the software product manager's role by other members of the team and organization. Users can learn on the various details of the software product manager's role from the personas and possibly find their own missing skills, new tools to master or new techniques to apply.

The research shows that there are many aspects of software product manager's role that require further research: (1) how Software Product Manager interplay with other roles, (2) which roles in the software engineering process have competences to become a software product manager (3) what affects how product teams are structured and how many software product managers are employed.

ACKNOWLEDGMENT

The authors thank all the experts that took part in the collection of data and the validation of the results.

REFERENCES

[1] H.-B. Kittlaus, S. A. Fricker, *Software Product Management: The ISPMA-Compliant Study Guide and Handbook*, Springer, 2017

[2] Ch. Ebert, "The impacts of software product management", *Journal of Systems and Software* 80 (6), pp. 850-861, 2007, DOI: 10.1016/j.jss.2006.09.017

[3] T. Tomaszewski: *Product Manager vs Project Manager*. Available: https://productvision.pl/2013/product-manager-vs-project-manager/

[4] E. Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, Currency, 2011

[5] A. Osterwalder, Y. Pigneur, *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*, John Wiley and Sons, 2010

[6] Aha.io *What tools do product managers use?* Available: https://www.aha.io/roadmapping/guide/product-management/which-tools-do-product-managers-use

[7] G. Geracie, S. Eppinger: *The Guide to the Product Management and Marketing Body of Knowledge: ProdBOK(R) Guide*, 2013

[8] A. Maglyas, U. Nikula, K. Smolander, *What are the roles of software product managers? An empirical investigation*, Journal of Systems and Software 86 (12), 2013, DOI: 10.1016/j.jss.2013.07.045

[9] K. Schwaber, M. Beedle, *Agile Software Development with Scrum*, Pearson, 2001

[10] H.-B. Kittlaus, *Software product management and agile software development: conflicts and solutions*. In: Maedche, A., Botzenhardt, A., Neer, L. (eds.) Springer, Berlin, 2012

[11] A. Kaczorowska, *Traditional and Agile Project Management in Public Sector and ICT*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 5, pages 1521–1531, 2015, DOI: 10.15439/2015F279

[12] H.-B. Kittlaus, *One Size Does Not Fit All: Software Product Management for Speedboats vs. Cruiseships*, International conference on software business (ICSOB 2015), Braga, Portugal, 2015

[13] Ch. Ebert, S. Brinkkemper, "Software product management - An industry evaluation", *Journal of Systems and Software* 95, 2014, DOI: 10.1016/j.jss.2013.12.042

[14] I. van de Weerd, S. Brinkkemper, R. Nieuwenhuis, J. M. Versendaal, A. Bijlsma, *On the Creation of a Reference Framework for Software Product Management: Validation and Tool Support*, International workshop on software product management (IWSPM'06), Minneapolis, MN, USA, 2006

[15] Pragmatic Marketing, Inc., *Pragmatic Marketing Framework*, 2016

[16] European Commission, *The New SME Definition, User guide and model declaration*, 2005

[17] R. Mencarelli, A. de Rivière, "Perceived value in B2B and B2C: A comparative approach and cross-fertilization", *Marketing Theory*, Vol 15, Issue 2, 2015, DOI: 10.1177/1470593114552581

[18] L. Nielsen, *Personas - User Focused Design*, Springer, 2013

[19] Definition of archetype, Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/archetype

[20] Product Crew, Product Vision Blog, www.productvision.pl

[21] A. Strauss, J. Corbin, *Basics of Qualitative Research – Techniques and Procedures for Developing Grounded Theory*, second edition, Sage Publications, London, 1998

[22] A. Przybyłek, M. Zakrzewski, *Adopting Collaborative Games into Agile Requirements Engineering*, 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'18), Funchal, Madeira, Portugal, 2018

[23] A. Przybyłek, D. Kotecka, *Making agile retrospectives more awesome*, Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds), ACSIS, Vol. 11, pp. 1211–1216, 2017, DOI: 10.15439/2017F423

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 995–1004
DOI: 10.15439/2018F300
ISSN 2300-5963 ACSIS, Vol. 15

What Can Go Wrong in a Software Project? Have Fun Solving It

Miguel Ehécatl Morales-Trujillo
University of Canterbury,
Christchurch, New Zealand
miguel.morales@canterbury.ac.nz

Gabriel Alberto García-Mireles
Universidad de Sonora,
Hermosillo, Sonora, México
mireles@mat.uson.mx

Polina Maslova
Universidad Nacional Autónoma de
México, Ciudad de México, México
pmaslova@comunidad.unam.mx

***Abstract* — Providing stimulating and real-life experiences is a key component in teaching software project management in Computer Science or Software Engineering programs. The diversity of topics that need to be addressed and restrictions that should be considered in university courses make a challenging task of it. This paper presents a serious game, called "White Crow PM" whose objective is to make students aware of the risks they might face during software development projects. The paper describes the game design steps and provides results of its validation in Computer Science programs in two Mexican universities. The collected data showed that the participants had fun playing the game and its content is relevant for software project management courses. Although the game needs to be validated in other settings with more participants, we conclude that it fulfills the goal of motivating discussion and increasing awareness of project management concerns among students.**

I. Introduction

DEVELOPING software systems require organizations to implement appropriate software project management practices. Industrial standards, such as ISO/IEC 12207 [1] or ISO/IEC 29110 [2], include specific processes to address activities and tasks for planning, monitoring, and controlling a software project. Indeed, software project management activities represent a core knowledge area in Software Engineering (SE) [3] and teaching them is supported by organizations such as IEEE and ACM through guidelines for curricula development [4].

SE curricula guidelines recommend students carry out activities in which they develop projects for a real client or get involved in software development during professional training [4]. However, in the area of software project management, a future IT professional lacks practical skills given that most project management syllabi are highly theoretical [5]. In addition, to understand technical challenges, software professionals also need to understand nontechnical issues such as management, communication and teamwork [4] [6].

In contrast with current practice of teaching SE topics by means of lectures, addressing these topics in SE courses require learning strategies that promote application and transfer of knowledge to authentic contexts where knowledge supports making decisions in real-life scenarios [7]. Furthermore, other constraints, such as class duration and instructor's effort impact on the extent project management practices can be carried out in either real or simulated scenarios [6] [8].

An attractive approach to address the aforementioned issues, within a risk-free environment, is the use of games with learning objectives as outcomes [5]. Educational games, also known as *serious games*, pursue the purpose of teaching, changing an attitude or behavior, or creating awareness of a certain issue [9].

Games offer enjoyment, motivation, social interaction and gratification; factors that support learning process [10]. Besides, serious games provide a variety of approaches for learning and teaching, becoming a complimentary tool to achieve specific learning outcomes [6]. According to [11], using games motivates students, immerses them in learning materials and supports learning from their own mistakes.

Several serious games have been proposed to achieve learning goals in SE. They address knowledge areas such as software process, software design, and professional practice [6]. However, few serious games proposals have been evaluated in the computer science domain and even less in software project management [5]. Therefore, given the complexity of SE topics, specificities of teaching SE in universities, and the variety of domains related to software project practices, there is a necessity to explore the advantage of using serious games in supporting SE related learning outcomes.

Serious games can be classified as digital or non-digital. While the former might be a computer application game, the latter might include card games and board games, among others [5]. Board games, in particular, are easy to use, allow interactivity between players, and provide a platform to carry out frequent and inexpensive updates [10].

This paper presents a board game whose objective is to make students aware of risks they might face during a software project and practices they can apply to mitigate them. The board game simulates an environment of a small organization carrying out a software project during a month. The game is targeted to undergraduate students in Computer Science and SE programs who already have introductory knowledge of project management.

This paper is organized as follows, Section II presents the concept of serious games and project management board games. Section III presents White Crow, the original board game. The adapted board game and its validation are

described in Sections IV and V respectively. Finally, conclusions and future work are presented in Section VI.

II. Background

A. Serious Games in Software Engineering

A game is defined as "an activity engaged in for diversion or amusement" [12] as well as "an activity…usually involving skill, knowledge, or chance, in which you follow fixed rules and try to win against an opponent or to solve a puzzle" [13]. While preserving the element of entertainment, a serious game focuses on achieving learning outcomes or educational goals, such as learning specific skills and concepts [6].

The increasing use of serious games inject more fun in both learning and training context due to the power of game to motivate players and the capabilities of them to facilitate cognitive gain, awareness and behavioral change [14]. The use of games on different educational levels has been demonstrated to motivate players to achieve goals, to stimulate interaction, and to encourage players to learn by doing, among other benefits [15]. Although digital games provide a lot of opportunities to incorporate multimedia resources to improve the user experience, they can be prototyped in paper [16].

Thus, working with non-digital games, such as board games, provides the opportunity to explore and gather the initial ideas on the game by the developer team and other stakeholders. In addition, learning from one another while at the same time having fun are experiences that a board game can provide [15].

In the context of SE, the use of games is motivated by the fact that traditional lectures barely address in practical way real life experiences where students make decisions and explore scenarios [6] [10]. Limitations on time or students' availability for participating in software projects are challenges that make deployment of software practices in real-life contexts difficult [6]. Indeed, providing real life experience in managing projects becomes almost impossible in SE education [6].

There is a growing presence of studies related to serious games in SE. Souza et al. [6] reported 86 papers that describe the use of serious games in SE education, and whose learning goals are related to the knowledge areas of software process, software design and professional practice. However, the authors do not decompose the 'software process' category in order to identify studies related to software project management. In a literature review on serious games in education, Calderón and Ruiz [5] reported that only 10% of papers correspond to the Computer Science domain and only two papers present games related to software project management.

B. Board Games for Project Management

Board games support understanding and learning of abstract concepts; their immersive nature facilitates attention, concentration and motivation [17]. Board games allow a learning-by-doing approach; in addition, game competitiveness urges players deeply understand the rules behind the game, and promotes reflection and discussion among players [17].

However, board games for project management are barely addressed in the education domain [5] [18]. Telukunta et al. [19] developed a game called StrateJect, which is similar in design to the Monopoly game. It is a computer game in which players experience consequences of executing or neglecting important project management functions in alignment with PMBoK 5 concepts. Taran [10] proposed a risk management board game in which students gather experience in making decisions involving risks. The game does not include teaching specific practices for risk management because it expects students to learn by their own experience.

Another board game is Deliver! [8]; its objective is to reinforce and teach the application of earned value management concepts targeted to students in undergraduate computing programs. In addition to the perceived potential to learn earned value management concepts and procedures, the game is reported to have a positive effect on social interaction, engagement, immersion, attention and relevance to the course objectives [8].

Other proposals address games for either training software practitioners in daily scrum meetings or learning software processes. In the first case, Yilmaz [20] provides a case study in which the focus is on identifying issues in testing scenarios related to the way a daily scrum meeting is conducted. Within a virtual reality environment, learners interact with virtual personas that have distinct personalities traits in order to trainees get clues about the issue presented in a scenario.

With regard to software processes, Aydan et al. [18] proposed a serious game, called "Floors", to introduce a preliminary training about both vocabulary and processes of the ISO/IEC 12207. The game was designed to visualize a virtual office environment and 3D character models that explain definitions and activities of the processes. Based on quests and dialogues, players can follow the processes organized by a software life cycle model.

C. Evaluating a Serious Game

The primary method to assess a serious game is a questionnaire that is typically applied after the game is played [5]. Questionnaires may include both quantitative and qualitative questions, however, the Likert scale is the most common method to gather participants' perceptions of the game [5]. Students' perceptions are measured by such variables (constructs) as belief, motivation state, expectations and emotions [8]. Actually, "evaluation from student's perceptions represents a simple, quick and less intrusive alternative to obtain feedback" [8].

We used a questionnaire from Wangenheim et al. [8], which focuses on evaluating students' feeling after playing a serious game; we adapted this standardized questionnaire and added concepts from the Kirkpatrick's evaluation model [21]. Altogether the questionnaire assesses motivation, game user experience, learning aspects, and students' perceptions.

Observation is another method used to evaluate serious games [5]. Observations are carried out by facilitators who observe game sessions either to get a general impression or to monitor any particular aspect of interest.

Several aspects can be evaluated in a serious game. Calderón and Ruiz [5] report that learning outcomes, usability and user experience are the most commonly evaluated. The first aspect refers to what learners should know or be able to do as a result of playing a serious game. Usability evaluates both ease of use and learnability while user experience assesses behavior, attitude and emotions.

Other commonly assessed aspects are user's satisfaction (user's attitude towards the serious game), motivation (how the serious game influences users' attention and behavior towards learning outcomes), and enjoyment (whether the serious game is able to provide a fun experience), among others [5].

There are few frameworks to design and evaluate a serious games [16] [22] [23] where mechanics-dynamics-aesthetics (MDA) model is a relevant one. Hunicke et al. [24] proposed this model to understanding games considering both designer and player perspectives in the context of three levels of abstraction: mechanics, dynamics, and aesthetics. The game creators design the mechanics of the game in order to provide a player experience, while player address the game from the aesthetics of the game, i.e., the user experience of playing the game. The model can support the specification of design goals, provides a means to discover game enhancement opportunities and allow to determine the measures to assess the progress in an improvement effort.

The MDA abstractions levels are described as follows [24]. Mechanics defines the game components, including actions, behaviors and control mechanisms. Some game mechanics are levels, tokens, questions and answers, game turns, resource management, and movement [23]. On the other hand, dynamics focused on the behavior of the game considering inputs and outputs. This abstraction level works at systemic level in order to create aesthetic experiences. The aesthetics level is related to the emotional responses of the players when interact with the game. This level can include a fun vocabulary composed of concepts such as: sensation, fantasy, narrative, challenge, fellowship, discovery, and expression.

D. Agile and Lean Approaches for Teaching Project Management by Means of Serious Games

Several games have been proposed to learn principles and practices of lean and agile methods in managing a software project. For lean approaches, Przybylek and Olszewski [25] proposed a game-based extension to Open Kanban. They suggested 12 games to address the four principles: visualization of the workflow, learn and improve, limit work in progress and lead using a team approach. As a result of validating the game-extension of Open Kanban, authors reported improvements in participants' communication, commitment, motivation, and the teams understood the main values and practices of Open Kanban [25]. Another proposal presented a collaborative Kanban board game for a software project management course [26]. The game's learning goals addressed both general description of the Kanban process and detailed aspects of the relationship between work in progress limits, lead time, and bottlenecks. The empirical results showed that learning goals were partially achieved. As for the attitudes towards the game, the participants' feedback reported a positive and highly motivating experience [26].

Incorporation of Scrum in the industrial sector has also been addressed. Several researchers have developed serious games to introduce and reinforce scrum practice. For instance, Przybylek et al. [27] [28] proposed to equip Scrum teams with a set of serious, collaborative games to address social aspects of software development. In turn, De Souza et al. [29] designed SCRUMI, an electronic serious board game for teaching Scrum concepts. In order to move forward through the game, participants have to answer questions on Scrum practices, which are grouped into five phases: preparation, analysis, execution, monitoring and control, and closing. Having evaluated the game, the authors reported that students felt motivated, satisfied and had fun playing the game [29].

SCRUMIA, on the other hand, is a manual paper and pencil game to reinforce the application of Scrum in undergraduate computing programs [30]. The game's main objective is to create artifacts while executing the Scrum process. In the validation results, the authors reported that SCRUMIA was effective, efficient and engaging for teaching Scrum practices [30]. Authors in [30] [31] mention several other games centered on different learning goals related to Scrum; however, despite the relevance of agile and lean methods, there is still a shortage of games addressing these topics [26].

III. WHITE CROW BOARD GAME

The board game presented in this paper is based on a Russian board game called Belaya Vorona (Белая ворона), which is White Crow in English. The label "white crow" in Russian describes a person who is different and stands out in the crowd; it may have a negative connotation similar to "black sheep". As the inventors of the game wrote, "try on the feathers of a white crow: break free and get individuality". The game was created in the post-soviet period, pursuing the idea of changing and embracing the world of business.

The game simulates a month of economic life with the objective of maintaining a healthy economic status by doing business. In the course of the game, the players are constantly affected by unexpected situations that impact their finances for the better or for the worse and, if they do not act accordingly, may result bankrupt.

The original board game consists of:

- A dice
- Five tokens
- A game board (see Fig. 1)
- Play money
- 64 mail cards
- 16 business cards
- A notepad for tracking loans and investments.

During the game, the players get familiar with such concepts of business and banking as investments, deposits, loans, rates, bank shares, clients, profits and expenses. It also shows the importance of risks, luck, enterprise and caution.

Fig. 1. Game board. Image taken from the original game Belaya Vorona produced by Design Studio Art Lestnitsa, Centre of Prospective Projects.

Each player decides their banking policy in advance: whether they will invest or ask for loans. Besides, at the beginning of each round, players receive an amount of money, with which they can buy businesses (square Бизнес "business") and cover unexpected expenses. Any business they own pays back in case the player gets on the square Клиент "client". The unexpected expenses are either laid out on the game board or come by mail (square Почта "mail"). Each month corresponds to one round, and the winner is the player who saves the biggest amount of money by the end of the game.

A business can be bought by any player who lands on the square Business. The player must take a business card and decide whether to use the opportunity or to discard it. For example, Fig. 2 shows a business card Subway "Falcon". If a player decides to invest in this business opportunity, she/he must pay 1,500 units of play money[1], which is indicated in the first upper line as плата "charge", to the bank.

The investment pays off when the player lands on the square Client. In the case of the card in Fig. 2, the lucky player receives 2,000 rubles as indicated in the second line by the word цена "price", thus obtaining a 500 rubles profit.

The third, bottom line комиссионные "commission" indicates the amount of money that is given to another player. To do so, all the players except the business card owner roll the dice, and the player with the highest points receives the "commission" money, 150 rubles in the case of Fig. 2 card. In the adapted version of the game, this transaction was omitted for reasons of simplification.

Fig. 2. Business card. Image taken from the original game Belaya Vorona produced by Design Studio Art Lestnitsa, Centre of Prospective Projects.

If a player lands on the square Mail, he/she must take a mail card and to act according to its instructions. In Fig. 3 mail card the player must pay 150 rubles, indicated in the left top corner as оплатить "to pay", to the bank for an emergency surgery the player had to undergo, which is explained in the right top corner of the card while the bottom of the card contains the name of the charging establishment.

We considered the White Crow board game to be suitable for adaptation due to its similarity with running a software project. The adapted game, White Crow PM, is based on simulating software project related activities and is meant to convey concepts of project management, risks and decision making. Similar to the original game, in the adapted version, players decide how to manage their money working on a software project during a month.

Fig. 3. Mail card. Image taken from the original game Belaya Vorona produced by Design Studio Art Lestnitsa, Centre of Prospective Projects.

One of the main goals of the adapted game is to make players aware of unexpected risks and associated expenses in software project management alongside with helping software engineering students realize what can go wrong or right and how well they can be prepared for it.

[1] In the case of the original game, rubles are used.

IV. WHITE CROW PM FOR SOFTWARE PROJECTS

A. Ad-hoc methodology to adapt the game

In order to narrow the general business-related scope of the original game and to shift its focus onto software projects context, we adopted an ad-hoc methodology, which was composed of 7 major steps.

The first step was to play the game to ensure it had an acceptable level of entertainment and challenge. Then it was translated from Russian into Spanish by two of the authors.

The second step was to analyze the original translated game, looking for the elements to be modified in order to simulate a software project without affecting the game's logic and flow. The 64 mail cards and the 16 business cards were reviewed, concluding that the prices were to be preserved but the number of cards might be increased. Moreover, the commission related rules were eliminated in order to simplify the game.

The third step consisted in rewriting the mail and business opportunity cards to suit software project related situations. The situations were designed by Computer Science students and professors, which was part of the final assignment of a Software Engineering advanced course and were based on their practical experience. In addition, selected papers and books were provided by the professors as a source of ideas for the students.

The fourth step was to develop a proof of concept for the new version of the game. The game was played by Computer Science students; the following aspects were examined:

- A1. Readability and clarity of the game rules.
- A2. Coherency and accuracy of the situations.
- A3. Students' perception of gaining knowledge.
- A4. Level of entertainment of the game.
- A5. Duration of the game.

Once the aspects to be improved were identified, a new version of the game was created.

During the fifth step we carried out a beta test of the game with students enrolled in an introductory Project Management course. 10 Applied Mathematics and Computation students, divided in 2 groups of 3 and 1 group of 4, played the game.

Afterwards, they answered an online questionnaire that measured aspects A1 to A4 using a Likert scale (1 to 5 scale values) while aspect A5 was measured by the two authors of the paper who observed the activity.

The results of the survey were as follows. For A1. Readability and clarity of the game rules, a median of 4.0 (42 out of 50 points) was obtained. For A2. Coherency and accuracy of the situations described in the cards and board, a median of 5.0 (46 points) was obtained. A3. Students' perception of gaining knowledge produced a median of 4.5 (45 points). Lastly, A4. Level of entertainment reached a median of 5.0 (50 points).

Regarding A5. Duration of the game, it was noted that a round to the board took 19 minutes in average while a recommended duration was two rounds. Overall, a 60 minutes class provides enough time to present the game, explain its rules, play two rounds and discuss the outcomes.

Through the same questionnaire we collected improvement suggestions to the board game, most of which concerned the game board and cards aesthetics. However, an important adjustment to the rules was suggested. The first round of the games was perceived as "slow" by the participants because of the "small" amount of money they received at the beginning of the game, hampering the possibility to invest and take advantage of the business opportunities early in the game. With that in mind, the players will receive a one-time initial bonus at the beginning of the first round.

The sixth step consisted in validating the new version of the game. This step is described in detail in Section V.

The last step was the socialization of the game, making it available to other professors and students.

B. White Crow PM [2]

In this subsection, the latest version of the board game is presented.

At the beginning of the game a bank manager is appointed, and each player decides their banking policy: they can either make savings or ask for loans. The purpose of this restriction is to offer the players a moment of decision making, facing up to its consequences later. The rates and interest can be agreed upon, those recommended by the game are 10%. Each player receives a one-time bonus of 1,000 units of play money and decides how much money they want to keep or deposit.

The game is best suitable for 3 to 5 players and it includes: a game board (see Fig. 4), 72 email cards, 20 Opportunity cards, tokens, a dice and play money. These components correspond with the Mechanic abstraction level of the game, according to the MDA model.

The Dynamic abstraction level of the game is described as follows: When a player gets to:

Email square: they have to take an email card. Email cards describe various aspects of a software project, such as Process related (Planning, Monitoring, Closure, Requirements, Design, Construction, Integration, Testing); Product related (Functionality, Reliability, Usability, Efficiency, Maintainability, Portability, Compatibility, Security) or Team related (Organization, Collaboration, Communication, Environment), which are indicated on the left side of the card. At the top, there is a short title of the situation, followed by an indication to pay or to receive, and a more extended description of the situation. The majority of the email cards requires the player to pay for some unexpected expenses while the rest recover money for the player.

[2] The White Crow PM latest version (in Spanish) can be downloaded from https://goo.gl/NjqMKo

Fig. 4. White Crow PM board

Fig. 5 shows email cards related to the Team communication and the Integration phase in a project. All the transactions must be made at the moment of getting the card. In case an email card indicates the amount to be paid per month (see Fig. 5, right), the player must multiply this amount by the number of months already played. This type of cards contains situations that have a wider impact on software projects, for example, an appearance of bugs in modules from first iterations. The rationale behind this modification was the importance of timing in discovering and solving issues.

Improvement opportunity square: the player has to take an Opportunity card and decide whether they want to use it or not. An Opportunity card offers an improvement related to software projects and mitigates unexpected expenses from email cards. The player must pay an indicated price to use it, however, they may choose to discard it without using.

The Improvement opportunity cards replaced the Business opportunity cards from the original game. An opportunity card does not provide any financial benefit for the player; it, however, mitigates risks or expenses from Email cards.

Fig. 6 shows an Improvement opportunity card, in which an investment in a software architect mitigates possible extra costs related to software construction and maintainability, while Fig. 7 shows an agile oriented opportunity.

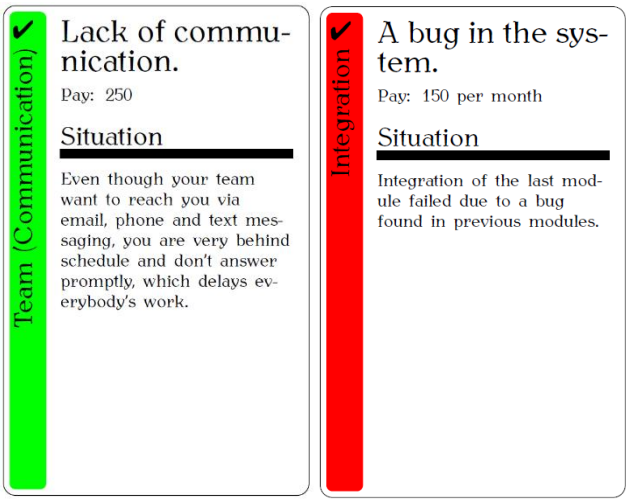


Fig. 5. Email cards, from White Crow PM

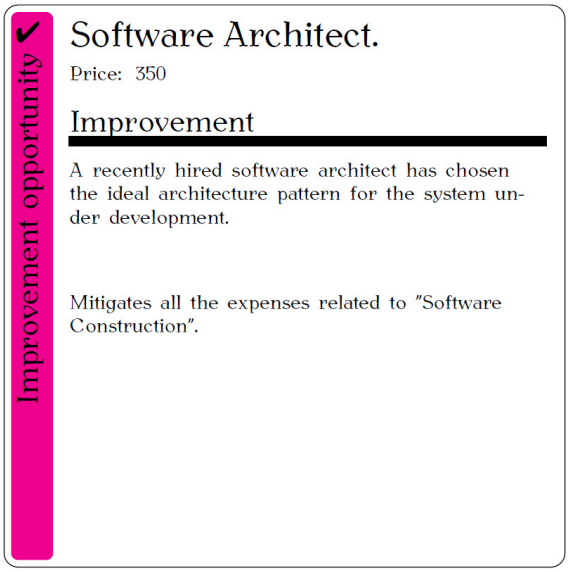


Fig. 6. Opportunity card, from White Crow PM

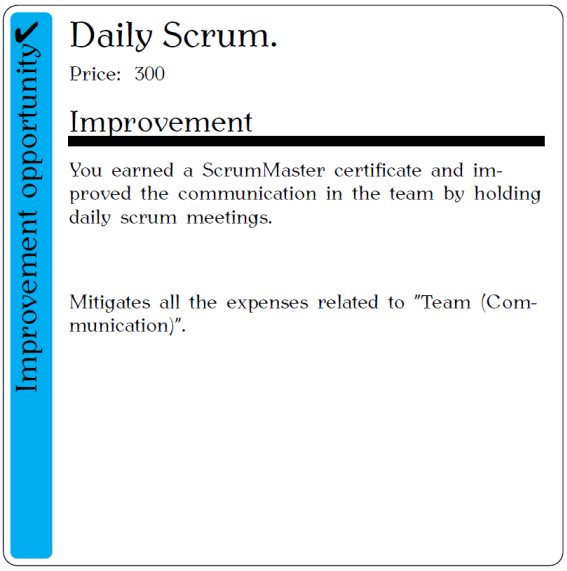


Fig. 7. Opportunity card, from White Crow PM

Bonus: players follow the instructions on each respective Bonus space, which always have a positive effect.

Expense: players follow the instructions on each respective Expense space, which always have a negative effect.

Gambling: each player bets 100 units of play money and throws the dice. The player to get the highest number wins all the money. This feature is preserved from the original game; it highlights occasional good or bad luck and gaming experience.

Last minute change: every player must go back one space and follow its respective instructions. Students become aware of unexpected changes in the project development.

A new project manager on the team: each player must hand over 50 units of play money. The money is kept in the space at the bottom of the board till one of the players gets a six on a dice roll and collects all the money.

Special offer: the player's expenses for the next two weeks will be reduced by 50%.

Break: players take a "day off".

Day of the White Crow: each player must pay any rates generated by bank loans, and receive 325 units of play money and interests, if any.

The game finishes by the end of the agreed number of rounds, and the player with the biggest amount of money is announced the winner.

V. Validation and Results

The hypothesis that guided this work is defined as follows: educational games contribute positively to achievement of learning objectives, motivate students, and promote a pleasant user experience.

With the aim of testing this hypothesis, the board game was validated in four groups of students from two Mexican universities. In total, 15 participants played the game; Table I presents a summary of the groups composition. The 4 groups were considered to have a homogenous academic background, cultural context and interest in the topic.

Table I.
Groups composition

Summary of groups			
University	**ID**	**Size**	**Degree**
A	Group 1	4	3 Engineering 1 Mathematics
	Group 2	3	3 Computer Science
B	Group 3	4	4 Computer Science
	Group 4	4	4 Computer Science

To obtain feedback and the student's perception of the game, a 4-section questionnaire was applied to the participants:

- **Basic information (5 questions):** questions in this section targeted students' academic background and experience, if any, in software projects.
- **Gaming Experience (15 questions):** questions in this section were based on the questionnaire presented in [8], and evaluated motivation, user experience and learning components. We used a Likert scale with response alternatives ranged from "strongly disagree" to "strongly agree" on a five points scale from 1 to 5.
- **Emotions (18 options):** this section consisted of a list of emotions, and each participant was invited to choose those experienced during the game.
- **General opinion:** the last section invited the participants to suggest improvements and to write their perceptions and opinion. We also included a question to evaluate how similar email and opportunities situations were to real projects, which was graded using Likert scale.

Table II shows a summary of the results of the Gaming Experience section, column M represents the median.

The analysis of the results points out that the participants had fun playing the game (5.0) and found it easy to understand (5.0). Moreover, they believe that the game content is relevant to their interests (4.0), is connected to other knowledge they already acquired (4.0) and suits their way of learning (4.0). As a whole, the dimensions of relevance, fun, satisfaction and learning were graded highly.

TABLE II.
GAMING EXPERIENCE SECTION RESULTS

Motivation	
Attractiveness	**M**
1. The game design is attractive.	**4.0**
2. The form, content and activities helped me to stay focused on the game.	**4.0**
Relevance to learning interests	
3. The content of the game is relevant to my interests.	**4.0**
4. The way the game works suits my way of learning.	**4.0**
5. The game content is connected to other knowledge I already have.	**4.0**
Confidence	
6. It was easy to understand the game and start playing it as learning material.	**5.0**
7. While playing the game, I felt confident that I was learning.	**4.0**
Satisfaction	
8. I feel positive because I know I will have opportunities to use what I learned playing this game in practice.	**4.0**
User experience	
Social interaction	
9. I had fun playing with other people.	**5.0**
10. The game promotes moments of cooperation and/or competition among players.	**5.0**
Fun	
11. This game has an adequate level of challenge for me; the tasks are neither too easy nor too difficult.	**4.0**
12. The game progresses at an adequate pace and does not become monotonous; it offers new obstacles, situations or task variations.	**4.0**
Competence	
13. I had fun playing the game.	**5.0**
14. When interrupted at the end of the class, I was disappointed that the game was over.	**4.0**
Learning	
Short-term learning	
15. I achieved the goals of the game applying my knowledge.	**5.0**

The participants' comments highlight learning aspects of Risks, Planning and Resource Management, in their own words: "Risks can affect and hamper my projects"; "It is better to be prepared to face them"; and "Good investment can help me to avoid risks". The comments also point to an increased PM-related awareness among students; measuring PM-related learning is out of the scope of this research stage.

An important factor to take into consideration was students' feelings and emotions while playing. Their evaluation allows us to gain insight into the user's experience and attitude towards the game.

This information was collected in the Emotions section, in which the participants were offered a list of emotions to choose from. These data is a first approach to describe the Aesthetics abstraction level of the game.

The results are shown in Table III. Positive emotions are preceded by a plus sign (+) while negative emotions are preceded by a minus sign (-). The number in the second column indicates how many participants chose the respective emotion.

TABLE III.
EMOTIONS QUESTIONNAIRE RESULTS

Emotions	
+ Interested	13
+ Eager	7
+ Optimistic	7
+ Focused	6
+ Confident	6
+ Happy	5
+ Capable	5
+ Relaxed	4
+ Immersed	4
+ Challenged	4
+ Satisfied	3
+ Encouraged	2
+ Useful	2
- Angry	2
- Bored	2
- Sad	1
- Lost	1
- Stressed	0

It can be noted that the negative emotions where chosen by far less frequently than the positive ones.

Finally, the last section of the questionnaire provided improvement suggestions and opinions of the game. On the one hand, the general opinion expressed was positive; it was fun, interesting, and a good idea that could be used more frequently in classrooms. The similarity of the presented situations with real projects obtained a median of 4.0, which we consider a positive score.

The participants mentioned: "It was a game but those things (situations presented) might happen in real life"; "My

interest for the subject increased"; "Very entertaining and easy to play, moreover, it gave me a good perspective of a software project"; and "I had a lot of fun, I like it, the situations are real".

On the other hand, creation of more opportunity cards was suggested; reading them before the game starts in order to have a wider vision of the game possibilities was also mentioned.

Another repeatedly commented aspect was the possibility to increase opportunities for collaboration between players. Lastly, an inclusion of a tip or advice in the email cards is seen as a possible path to provide more leaning opportunities. These improvements will be tackled in a future version of the game.

A. Threats to validity

In order to mitigate threats to validity, various factors were considered during the board game design and validation. Different causal relations were examined:

- **The number of participants.** Although the number of participants is not large, we consider the sample to be representative of students enrolled in Computer Science courses. However, validating the game with a small set of participants reduces the possibility of generalizing the results.
- **The questionnaire's trustworthiness.** The applied questionnaire is an adaptation of a recognized and proven instrument specialized in the evaluation of training and learning resources [8]. We are aware, however, that such aspects as game appropriateness and engagement are difficult to measure and were captured through subjective measures.

In spite of the limited number of students that participated in the validation, each student can be categorized as a typical Computer Science student, which is the target audience of White Crow PM. The selection of the participants was not intentional; the students who participated in the validation expressed their interest in participating.

Another factor to highlight is that the game applications were guided by several facilitators, 3 out of 4 groups were guided by professors not familiar with the game before. This factor supports the easy to use and apply claim.

The case studies were carried out by three researchers and the results were constantly triangulated to third parties, such as colleagues and members of the research group.

In order to improve the validity of this study the following approaches cited in [32] were taken into account:

- **Triangulation** was possible due to an active participation of two professors in the data collection process. Thus we were able to analyze different data sources: questionnaires and direct observations.
- **Peer debriefing** took place in all game applications. In addition, findings and results were periodically discussed with other members of the research group.

The game board applications demonstrated that the objective for which it was created was achieved, as it made students aware of risks they might face during a software project and of practices they can apply to mitigate them. In addition, the board game motivated students and promoted a pleasant experience.

Finally, the limitations of the case studies can be summarized in two points:

- The sample size is small, which therefore limits the power of generalization. It is necessary to apply the game in bigger populations.
- Bias in the game application could have occurred due to the participants' feeling of being observed and evaluated. This may have led to an alteration in their actual behavior.

VI. Conclusions and Future Work

Simulating a software project and making students face consequences of their decisions pose a challenge due to diverse internal and external factors. Therefore, games and simulations play an important role in creating abstractions and simplifications of real life software development [6].

Serious games represent a powerful learning tool in the field of education. Their teaching potential and entertaining aspect provide an alternative to traditional learning process in the classroom. Moreover, serious board games are cheaper and easier to apply, and allow educator to identify on-site advantages.

White Crow board game, adapted to the software projects context, resulted in a successful and useful resource to introduce students into software project management, and make them aware of its complexity, unexpected variables and the importance of decision making.

The students' opinions obtained through the questionnaire and the observations allowed us to conclude that the game supports enjoyment, motivation, social interaction and gratification; factors that support learning process.

As future work, improvements on the cooperative factor of the board game are considered, for example, being able to make agreements between players and the bank as well as between players. Besides, since cultural aspects also count, and this version of the game is based on the Mexican way of running a project, the situations in the cards could be tailored to other cultures.

Another interesting future research line might be creating an agile-oriented version, based on a 30-day sprint with players using a product backlog and specific PM events explicitly included in the board. More validation forms, such as in-depth interviews with experts and participants, are considered in order to enrich the qualitative analysis of the results.

Acknowledgment

The authors would like to thank Computer Science students Mauricio Esquivel Reyes, Anahí Quiróz Jiménez, Nancy

Matias Hernández and Karla Andrea Contreras Maya, who collaborated in the adaptation and validation of the game.

Also, special thanks to the professors Guadalupe Ibargüengoitia González, Gabriela Martínez Quezada and Maribel Santiago Luna, who kindly agreed to apply the game in their groups.

REFERENCES

[1] ISO/IEC 12207:2008 Systems and software engineering — Software life cycle processes (2008)

[2] ISO/IEC TR 29110-5-1-2:2011 Software engineering—lifecycle profiles for Very Small Entities (VSEs)—part 5-1-2: Management and engineering guide: Generic profile group: Basic profile (2011)

[3] Bourque, P, Fairley, R.E. eds., *Guide to the Software Engineering Body of Knowledge*, Version 3.0, IEEE Computer Society (2014)

[4] ACM and IEEE, Software Engineering Curriculum Guidelines; http://securriculum.org (2014)

[5] Calderón, A. and Ruiz, M.: *A systematic literature review on serious games evaluation: An application to software project management*. Computers & Education, Vol. 87, pp. 396-422, DOI: 10.1016/j.compedu.2015.07.011 (2015)

[6] Souza, M. R. A., Veado, L., Moreira, R. T., Figueiredo, E. and Costa, H.: *A Systematic Mapping Study on Game-related Methods for Software Engineering Education*. Information and Software Technology, Vol. 95, pp. 201-218, DOI: 10.1016/j.infsof.2017.09.014 (2017)

[7] Choi, J-I. and Hannafin, M.: *Situated cognition and learning environments: Roles, structures, and implications for design*. Educational Technology Research and Development, Vol. 43, No. 2, pp. 53-69, DOI: 10.1007/BF02300472 (1995)

[8] von Wangenheim, C. G., Savi, R. and Borgatto, A. F.: *DELIVER!–An educational game for teaching Earned Value Management in computing courses*. Information and software Technology, Vol. 54, No. 3, pp. 286-298, DOI: 10.1016/j.infsof.2011.10.005 (2012)

[9] Alonso-Fernández, C., Calvo, A., Freire, M., Martínez-Ortiz, I. and Fernandez-Manjon, B.: *Systematizing game learning analytics for serious games*. In Proc. Of the IEEE Global Engineering Education Conference, pp. 1111-1118, DOI: 10.1109/EDUCON.2017.7942988 (2017)

[10] Taran, G.: *Using games in software engineering education to teach risk management*. In Proc. Of the 20th IEEE Conference on Software Engineering Education & Training, pp. 211-220, DOI: 10.1109/CSEET.2007.54 (2007)

[11] Teed, R.: *Game-Based Learning*. SERC, Carleton College. https://serc.carleton.edu/introgeo/games/index.html (2018)

[12] Game, in Merriam-Webster.com. Retrieved May 15, 2018, from https://www.merriam-webster.com/dictionary/game

[13] Game, in Collins Dictionary. Retrieved May 15, 2018, from https://www.collinsdictionary.com/dictionary/english/game

[14] Arnab, S.,\ and Clarke, S.: *Towards a trans-disciplinary methodology for a game-based intervention development process*. British Journal of Educational Technology, Vol. 48, No. 2, pp. 279-312, DOI: 10.1111/bjet.12377 (2017)

[15] Retalis, S.: *Creating adaptive e-learning board games for school settings using the ELG environment*. J. UCS, Vol. 14, No. 17, pp. 2897-2908 (2008)

[16] Kosa, M. and Yilmaz, M.: *The Design Process of a Board Game for Exploring the Territories of the United States*. Press Start, Vol. 4, No. 1, pp. 36-52 (2017)

[17] Chiarello, F. and Castellano, M. G.: *Board games and board game design as learning tools for complex scientific concepts: some experiences*. International Journal of Game-Based Learning, Vol. 6, No. 2, pp. 1-14, DOI: 10.4018/IJGBL.2016040101 (2016)

[18] Aydan, U., Yilmaz, M., Clarke, P. M. and O'Connor, R. V.: T*eaching ISO/IEC 12207 software lifecycle processes: a serious game approach*. Computer Standards & Interfaces, Vol. 54, pp. 129-138, DOI: 10.1016/j.csi.2016.11.014 (2017)

[19] Telukunta, S., Kota, M. S. K., Potti, M. S., Shashank, M. H. and Triloknath, M.: *StrateJect: An Interactive Game for Project Management Experiential Learning.* PMP Conference, PMI Bangalore chapter (2014)

[20] Yilmaz, M.: *Virtual Reality-Based Daily Scrum Meetings*. In: Encyclopedia of Computer Graphics and Games, Publisher: Springer, Editors: Newton Lee, pp. 1-6. DOI: 10.1007/978-3-319-08234-9_160-1 (2017)

[21] Kirkpatrick, D.L. and Kirkpatrick, J.D.: *Evaluating Training Programs: The Four Levels*. Berrett-Koehler Publishers (2006)

[22] Duarte, L. C. S. and Battaiola, A. L.: *Distinctive features and game design. Entertainment computing*. Vol. 21, pp. 83-93 (2017)

[23] Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., De Freitas, S., Louchart, S., and De Gloria, A.: *Mapping learning and game mechanics for serious games analysis*. British Journal of Educational Technology, Vol. 46, No. 2, pp. 391-411, DOI: 10.1109/TETC.2015.2504241 (2015)

[24] Hunicke, R., LeBlanc, M. and Zubek, R.: *MDA: A formal approach to game design and game research*. In Proc. of the AAAI Workshop on Challenges in Game AI, Vol. 4, No. 1, p. 1722, DOI: 10.1.1.79.4561 (2004)

[25] Przybylek, A. and Olszewski, M. K.: *Adopting collaborative games into Open Kanban*. In Proc. of the Federated Conference on Computer Science and Information Systems, IEEE, pp. 1539-1543 (2016)

[26] Heikkilä, V. T., Paasivaara, M. and Lassenius, C.: *Teaching university students Kanban with a collaborative board game*. In Proc. of the 38th International Conference on Software Engineering Companion, ACM, pp. 471-480, DOI: 10.1145/2889160.2889201 (2016)

[27] Przybylek, A. and Kotecka, D.: *Making agile retrospectives more awesome.* In Proc. of the Federated Conference on Computer Science and Information Systems, IEEE, DOI: 10.15439/2017F423 (2017)

[28] Przybylek, A. and Zakrzewski, M.: *Adopting Collaborative Games into Agile Requirements Engineering.* In: 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'18), Funchal, Madeira, Portugal (2018)

[29] De Souza, A. D., Seabra, R. D., Ribeiro, J. M. and Da S. Rodrigues, L. E.: *SCRUMI: a board serious virtual game for teaching the SCRUM framework.* In Proc. of the 39th International Conference on Software Engineering Companion, IEEE Press, pp. 319-321, DOI: 10.1109/ICSE-C.2017.124 (2017)

[30] von Wangenheim, C. G., Savi, R. and Borgatto, A. F.: *SCRUMIA—An educational game for teaching SCRUM in computing courses.* Journal of Systems and Software, Vol. 86, No. 10, pp. 2675-2687, DOI: 10.1016/j.jss.2013.05.030 (2013)

[31] Mahnič, V.: *Scrum in software engineering courses: an outline of the literature*. Global Journal of Engineering Education, Vol. 17, No.2, pp. 77-83 (2015)

[32] Runeson, P., Host, M., Rainer, A. and Regnell, B.: *Case Study Research in Software Engineering: Guidelines and Examples*. Wiley, DOI: 10.1002/9781118181034 (2012)

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1005–1008
DOI: 10.15439/2018F137
ISSN 2300-5963 ACSIS, Vol. 15

Usability attributes revisited: a time-framed knowledge map

Paweł Weichbroth
Faculty of Management, WSB University in Gdansk,
ul. Grunwaldzka 238A, 80-266 Gdansk, Poland
Email: pawel.weichbroth@hotmail.com

***Abstract*—Software usability plays a major role in the quality perceived by its users. However, a variety of definitions and associated attributes shows that there is still no consensus in this area. The overall purpose of this paper is to present the results of a critical and rigorous literature review, the aim of which is to demonstrate all the relevant usability definitions and related attributes introduced till now. This comprehensive view, depicted by a time-framed knowledge map, provides an in-depth understanding of the observed evolution on the one hand, and also serves as a guide for usability engineers to address some non-functional requirements, on the other.**

I. Introduction

DISCUSSING usability in general is one thing, but explaining its nature in detail in an applicative manner is another. An ordinary user of any application or device would point to ease of use when asked about the primary attribute of its usefulness. Indeed, in an article for Time Magazine, Tim Bajarin, while discussing "*6 Reasons Apple Is So Successful*", placed this answer in second place [O1]. Beyond this, market success is the hard work of not only the sales department, but mostly, visionary usability engineers.

To design and produce usable software, the concept of usability must be understood. However, till now, a plethora of definitions and related attributes, on the one hand, do not make it easy, while on the other, a flood of buzz words make it even worse. To fill this gap, we investigate existing models and standards to find answers to the research questions: (RQ_1) what is software usability? and (RQ_2) which attributes most frequently contribute to the software usability?

The reasons for performing this study are twofold. Usability is a major concern in every stage of the design process, and now design thinking is regarded as a system that fosters innovation. Secondly, in face of the observable shift from desktop to mobile application, nowadays, usability again captivates the interest of software vendors in adapting their products to the new settings.

The rest of the paper is organized as follows. In Section II, the research method is depicted, followed by the recognition and definition of usability attributes identified in multiple standards and models (Section III). Finally, we conclude the paper and point out future research directions (Section IV). The paper's contributions are: outlining past research, highlighting its drawbacks and providing one synopsis in the form of a time-framed knowledge map (Table I).

II. Research Method

In order to provide the relevant answers to the research questions, the literature survey was based on the approach of Webster and Watson [O2]. A rigorous search process was applied to the existing body of knowledge to identify norms, scientific papers and books, as well as technical reports that contribute to the subject field. The term "rigorous" stands for the reliability and validity of the search process. The former refers to the extent to which its outcome is consistent over time and an accurate representation of the total population under study, and if the results of the study can be reproduced under a similar procedure. The latter is the degree to which the literature search accurately uncovers the relevant sources. The research was composed of three stages, namely, literature search, literature analysis and selection of attributes.

The literature search embraced databases of widely recognized publishers whose scopes correspond to computer science, information systems or similar (e.g. ergonomics). In particular, ACM, Elsevier, IEEE and Springer were searched for the keywords *usability* and *software usability*, followed by a search of aggregated databases that store records of numerous publishers, namely EBSCOhost, Scopus and Web of Science. An electronic search was performed against the metadata of all the publications. Next, only articles published in relevant journals or books were taken into account. However, the abundance of search results pushed us to redefine our query by limiting the search fields to the title, keywords and abstract. Although this search was not exhaustive, it submitted a comprehensive input.

The literature analysis aimed to identify adequate content for the research questions (see Section I). To achieve this goal, firstly, the publications were reviewed to eliminate those biased by a context (user-specific attributes such as: age, occupation, sex or system-specific support features like visually impaired, disability), unrelated to computer systems or tailored to specific project constraints. Secondly, the remaining publications were read completely with a focus on their parts devoted to the subject, in order to catalog valid data.

The selection of attributes stage was based on the following, mutually nonexclusive criteria, applied to all the classified items: (*a*) published in English, (*b*) related to usability studies, (*c*) published by an international standard-

setting body or governmental institution, (*d*) referenced in research papers. The 'charting' technique was adopted to synthesize and interpret gathered qualitative data by sifting, charting and sorting information chunks according to key issues and themes. The final outcome is twofold: a narrative review of the recent research, and a time-framed knowledge map (Table 1) as the research summary.

III. Usability Attributes

Among efforts to explain what the term means, Shackel claims that the definition of usability was probably first attempted by Miller in 1971 in terms of measures for "ease of use" [1].

In 1977, McCall defines usability as "the effort required to learn, operate, prepare the input and interpret the output of a program" [O3]. However, the fact that it was already included as a quality factor did not really imply attention from software engineers, but it was regarded as an extension of general data processing system design. This perspective was typical of its time.

From the late 1970s, through the 1980s, it was in fashion to discuss software and other hardware-based artefacts which were easy to use as being "user-friendly". The question of "what is user-friendly" had little credibility; however, the problem of the efficiency of computer programs was often deliberated. These two issues were recognized together by several researchers, including Bennet (1984) and Shackel (1986), who introduced comparable definitions [O4, O5].

Shackel framed usability in terms of the system: *effectiveness*, the *easiness to learn*, its *flexibility* and user *attitude* [O5]. A formal definition, established in 1991, says that it is "the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfil the specified range of tasks, within the specified range of environmental scenarios" [1]. The notion is next split into four attributes: *effectiveness*: "in terms of performance (e.g. time, errors, number of sequence activities) in learning, relearning and carrying out a representative range of operations", *learnability*: "within a specified time from the installation and start of user training" and "the amount of training and user support", *flexibility*: which allows "adaptation to some specified percentage variation in tasks and/or environments beyond those first specified" and *attitude*: "within acceptable levels of human cost in terms of tiredness, discomfort, frustration and personal effort". However, despite the interest that has been aroused around this approach, it appears to have limitations. Flexibility is particularly difficult to specify, communicate and test in a real system development environment. Among others, Preece et al. [O6] also draw heavily on its rationale. Today, flexibility is seldom considered explicitly.

In 1987, the FURPS model was first introduced by Grady and Caswell [2]. In 1992, the original model "was extended to empathize various specific attributes" and re-designated to FURPS+ in which usability is described by four attributes: *aesthetics, consistency*, *documentation* and *human factors* [3].

In response to the need of the software industry to standardize the evaluation of software products using quality models, in 1991, the ISO Organization issued a standard, namely ISO 9126, that specifies six areas of importance for software evaluation, including usability, defined as "a set of attributes of software which bear on the effort needed for use, and on an individual assessment of such use, by a stated or implied set of users" [4]. This standard puts forward three attributes (named as sub-characteristics): *understandability*, *learnability*, *operability* (see also: [O7] and [O8]).

In 1993, Kirakowski and Corbett presented the SUMI (Software Usability Measurement Inventory) method to measure users' perception of software usability, providing three different layers of output. The second layer has five sub-scales: *affect*, *efficiency*, *learnability*, *helpfulness* and *control* [5].

Nielsen (1993) associates usability with five attributes: *learnability*, *efficiency*, *memorability*, *errors* and *satisfaction* [6]. Every definition starts with words "the system should be or have" which, in particular, applies to its capability of being easy to learn, efficient to use, easy to remember, free of errors and pleasant to use. Nielsen emphasizes learnability as "the most fundamental attribute of usability, since most products need to be easy to learn, and since the first experience that most people have with a new product is that of learning to use it", and also relates it to novices' ability to reach a reasonable level of performance rapidly, which indicates the direct relation between learnability and efficiency, i.e. the user interface should be easy to learn so that the user is able to complete a given task successfully in a certain time.

ISO 9241-11 (1993, 1998) [7, 10], along with Bevan (1995) [9], consider *effectiveness*, *efficiency* and *satisfaction* as usability measures. These standards relate to usability as a high level quality objective, which is reflected by its definition: "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Today, this is the most recognizable usability definition.

IBM's CUPRIMDSO (1994) quality assurance system includes customer *satisfaction* only, with a five-point scale for evaluating the value of the product that results from its effective usage [8].

Constantine and Lockwood (1999) propose five facets, as different aspects of a system and its user interface that contribute to usability: *learnability*, *rememberability*, *efficiency in use*, *reliability in use* and *user satisfaction* [11]. The authors declare that software which "leads its users to make fewer mistakes will be more reliable in use", that is, "in how it functions in combination with its users and in how it promotes reliable human performance". Reliability in use

is more closely tied with the user interface design than with coding and debugging.

Ten years after the introduction of ISO/IEC 9126:1991, the standard was refined by a group of software engineer experts to ISO/IEC 9126:2001. Even though it does not cover all aspects of software quality from the product perspective, it is still the most comprehensive model developed to date. Here, usability is specified as "the capability of the software product to be understood, learned and liked by the user, when used under specified conditions", construed as five attributes: *understandability*, *learnability*, *operability*, *attractiveness* and *usability compliance* [12].

Abran et al. (2003) developed an Enhanced Usability Model, by integrating process-related (ISO 9241) and product-related (ISO 9126) standards, which includes five attributes: *effectiveness*, *efficiency*, *satisfaction*, *learnability* and *security* [13]. To justify the last one, they provide a list of five arguments, particularly referring to ITSEC (Information Technology Security Evaluation Criteria) and to three normative standards (IEC 300, ISO 13407: 1999, ISO/IEC 9126), and eventually closing by indicating that "security is a characteristic of CHI, which is particularly important in an industrial context".

Seffah et al. (2006) elaborate the QUIM model by consolidating ten attributes (originally named factors): *efficiency*, *effectiveness*, *productivity*, *satisfaction*, *learnability*, *safety*, *trustfulness*, *accessibility*, *universality* and *usefulness*, each of which corresponds to a specific facet of usability, identified in an existing standard or model [14]. These ten factors are decomposed into a total of 26 sub-factors or measurable criteria that are further broken down into 127 specific metrics. It seems that the novelty of this work brings into view a complete and legitimate instrument to evaluate usability.

In ISO 9241-210:2010 [15], the usability definition is adapted from ISO 9241-11:1998 [10], along with its set of attributes as well.

The ISO 25010:2011 standard on quality models updates

TABLE I.
USABILITY ATTRIBUTES OF VARIOUS STANDARDS AND MODELS

Attribute / [Ref. No] & Year	**[1] 1991**	**[2, 3] 1987**	**[4] 1991**	**[5] 1993**	**[6] 1993**	**[7] 1993**	**[8] 1994**	**[9] 1995**	**[10] 1998**	**[11] 1999**	**[12] 2001**	**[13] 2003**	**[14] 2006**	**[15] 2011**	**[16] 2016**	**[17] 2018**
Effectiveness	●					●		●	●			●	●	●		●
Learnability	●		●	●	●					●	●	●	●		●	
Flexibility	●															
Attitude	●[2]															
Aesthetics		●														
Consistency		●														
Documentation		●														
Human factors		●														
Understandability			●								●				●[1]	
Operability			●								●				●	
Affect				●												
Efficiency				●	●	●		●	●	●[1]		●	●	●		●
Helpfulness				●												
Control				●												
Memorability					●					●[1]						
Errors					●										●[1]	
Satisfaction					●	●	●	●	●	●[1]		●	●	●		●
Reliability in Use										●						
Attractiveness											●				●[1]	
Usability Compliance											●					
Security												●				
Productivity													●			
Safety													●			
Trustfulness													●			
Accessibility													●		●	
Universality													●			
Usefulness													●			

●: Included. [1]: Defined under a similar name. [2]: Attitude has a moderate connotation contrary to satisfaction, thus they are not combined.

and brings together previous standards, defining three views of quality: internal quality, external quality and quality in use [16]. Of particular interest in ISO 25010 is the standard's new breakdown of quality in use and usability. The former is defined as "the degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk, and satisfaction in specific contexts of use", while the latter is defined the same as in [10, 15]. Usability can either be specified or measured as a product quality characteristic in terms of its subcharacteristics, or specified directly by measures that are a subset of quality in use [O8]. The standard delineates its six attributes (subcharacteristics) [16]:

- *appropriateness recognizability*: degree to which users can recognize whether a product or system is appropriate for their needs;
- *learnability*: degree to which a product or system can be used by specified users to achieve specified goals of learning to use the product or system with effectiveness, efficiency, freedom from risk, and satisfaction in a specified context of use;
- *operability*: degree to which a product or system has attributes that make it easy to operate and control;
- *user error protection*: degree to which a system protects users against making errors;
- *user interface aesthetics*: degree to which a user interface enables pleasing and satisfying interaction for the user;
- *accessibility*: degree to which a product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.

Compared with the 2001 edition (ISO/IEC 9126-1) [12], two new subcharacteristics were introduced: *user error protection* and *accessibility*, while usability compliance was withdrawn. It is worth noting that understandability was renamed to appropriateness recognizability, along with attractiveness to user interface aesthetics [16].

Similarly (see: ISO 9241-210:2010 and ISO 25010:2011), in ISO 9241-11:2018 [17] the usability definition is adapted from [15] together with its three attributes.

IV. Conclusions

Table 1 shows the final list of all usability attributes, together with the representative references, including also standards from the International Organization for Standardization (ISO), covering usability in human-computer interaction. From the list, the most frequent are *efficiency* and *satisfaction* (each supported by 10 references), *learnability* (9) and *effectiveness* (8). The least frequent are *understandability* and *operability* (3), *memorability*, *errors*, *attractiveness* and *accessibility* (2), while the rest attributes occur only once. The next research step is to determine usability facets relevant for mobile applications.

References

[1] B. Shackel, Usability – Context, framework, definition, design and evaluation. in B. Shackel and S. Richardson (Eds.), *Human Factors for Informatics Usability*, Cambridge 1991, pp. 21–38.

[2] R. B. Grady and D. Caswell, *Software metrics: Establishing a company-wide program*. Englewood Cliffs: Prentice-Hall 1987.

[3] R. B. Grady, *Practical software metrics for project management and process improvement*. Englewood Cliffs, NJ: Prentice-Hall 1992.

[4] International Organisation for Standardisation, ISO/IEC 9126. *Information Technology – Software Product Evaluation – Quality Characteristics and Guidelines for Their Use*. Genève 1991.

[5] J. Kirakowski and M. Corbett, SUMI: *the software usability measurement inventory*. British Journal of Educational Technology, Vol. 24(3), 1993, pp. 210-212.

[6] J. Nielsen, *Usability Engineering*, Academic Press, London 1993.

[7] International Standards Organization, ISO DIS 9241-11. *Ergonomic requirements for office work with visual display terminals. Part 11: Guidance on usability*. London 1993.

[8] D. L. Bencher, *Programming quality improvement in IBM*. Technical forum. IBM Systems Journal, 33(1), 1994, pp. 215–219.

[9] N. Bevan, *Measuring usability as quality of use*. Software Quality Journal, 4, 1995, pp. 115–130.

[10] International Organization for Standardization, ISO: 9241-11:1998. *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability*. Geneva 1998.

[11] L. L. Constantine and L. A. D. Lockwood, *Software for Use: A Practical Guide to the Models and Methods of Usage-Centered Design*. Addison-Wesley, New York 1999.

[12] International Organization for Standardization/International Electrotechnical Commission, ISO/IEC 9126 *Standard, Software Engineering, Product Quality*. Part I & Part IV. Geneva 2001.

[13] A. Abran, A. Khelifi, W. Suryn, and A. Seffah, *Usability meanings and interpretations in ISO standards*. Software Quality Journal, 11(4), 2003, pp. 325–338.

[14] A. Seffah, M. Donyaee, R. B. Kline and H. K. Padda, *Usability measurement and metrics: A consolidated model*. Software Quality Journal, 14(2), 2006, pp. 159–178.

[15] International Organization for Standardization, ISO 9241-210:2010. *Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems*. Geneva 2010.

[16] International Organization for Standardization, ISO 25010:2011. *Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models*. Geneva 2011.

[17] International Organization for Standardization, ISO 9241-11:2018. *Ergonomics of human-system interaction -- Part 11: Usability: Definitions and concepts*. Geneva 2018.

Other References

[O1] T. Bajarin, *6 Reasons Apple Is So Successful*. Time. May 07, 2012.

[O2] J. Webster, and R. T. Watson, R. T., Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, 2002, xiii-xxiii.

[O3] J. A. McCall, P. K. Richards, G. F. Walters, *Factors in Software Quality*. US Rome Air Development Center Reports, Vol I, II, III. US Department of Commerce, USA, 1977.

[O4] J. L. Bennett, Managing to Meet Usability Requirements: Establishing and Meeting Software Development Goals. in J. L. Bennett, B. Case, J. Sandelin and M. Smith (Eds.), *Visual Display Terminals: Usability Issues & Health Concerns*. Prentice-Hall 1984, pp. 161–184.

[O5] B. Shackel, Ergonomics in Design for Usability. in M. D. Harrison and A. Monk (Eds), *People and Computers: Designing for Usability*. Proc. of Sec. HCI Conference of the BCS, York 1986, pp. 45–64.

[O6] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland and T. Carey, *Human-Computer Interaction*. Pearson, 1994.

[O7] A. Przybyłek, *An empirical study on the impact of AspectJ on software evolvability*. Emp. Soft. Eng., 23(4), pp. 2018–2050, 2018.

[O8] P. Weichbroth, *Delivering Usability in IT Products: Empirical Lessons from the Field*. International Journal of Software Engineering and Knowledge Engineering, 28(07), 1027-1045, 2018.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1009–1012
DOI: 10.15439/2018F311
ISSN 2300-5963 ACSIS, Vol. 15

MaliciousIDE – software development environment that evokes emotions

Michal R. Wrobel
Faculty Of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
Email: wrobel@eti.pg.edu.pl

Adam W. Zielke
Faculty Of Electronics, Telecommunications And Informatics, Gdansk University of Technology, Poland
Email: adam.wojciech.zielke@gmail.com

***Abstract*—Emotions affect every aspect of human live, including work. Numerous studies in software engineering have shown that negative emotions can lower the productivity of programmers. Unlike traditional approaches to managing software development, modern methods, such as Agile and Lean, take into account human aspects of programming. To thoroughly investigate the impact of negative emotions on the work of programmers, a malicious integrated development environment (IDE) was developed. This tool allow a observer to trigger malicious behavior of the IDE. Conducted study have proved its usefulness. Participants reported that it mostly invoked frustration and angry.**

I. Introduction

RESEARCH on the role of emotions in the work of software developers has been deeply conducted in recent years. Finding the relationship between emotional state and productivity, or the quality of the developed code, may lead to the new methods of IT project management, and consequently to improve software development processes. This type of research corresponds to the current trend of research on social aspects in software development projects [1].

The field of affective computing in recent years has provided many algorithms and tools for recognizing the emotions of computer users. However, despite continuous development, they are still imperfect, especially in applications in unusual situations. Therefore, there is a huge demand for input channel data, such as a video clips or biometric data that are labelled by emotions. They can be used to teach algorithms or check new methods of recognizing emotions. Such data sets, especially those consisting of multiple channels, are not widely available to the public.

In the case of the analysis of emotions in the work of programmers, generic solutions are mainly used. With popular algorithms, such as recognizing emotions based on facial expressions, this approach is acceptable. However, the specificity of the work of programmers allows the development of other, dedicated algorithms. One of such approaches is recognizing emotions based on the analysis of patterns of typing and mouse movements [2]. To develop and learn such algorithms, it is necessary to gather data labelled with emotions. However, such data are extremely difficult to obtain from the real working environment. Therefore, laboratory studies are used for this purpose.

Quigley et al. identified thirteen laboratory emotions induction techniques [3]. However, most of them are not suitable for evoking the emotions of programmers during work. They usually rely on inducing emotions through exposure of the participant to visual or audible stimulus. The exception are three approaches: *peripheral physiological manipulations*, *motivated performance* and *physically real stimuli* [3]. In the case of software engineering studies, the first one is not suitable for general use because it requires direct stimuli in physiological systems, e.g. by injecting drugs. Others can be adapted to induce emotions of programmers.

For the purposes of laboratory studies on the emotions of programmers, a malicious integrated development environment, *MaliciousIDE*, has been developed at the Gdańsk Technical University since 2015. The goal of the project is to provide a solution that will induce negative emotions of programmers while working. It is adaptation of the *physically real stimuli* induction technique. In addition, the *motivated performance* was studied during a validation, which involved putting participants under the time pressure.

The rest of the paper is organized as follows: Section II describes the methods used so far in recognizing the emotions of software developers, Section III describes the concept of the MaliciousIDE plug-in, a survey that verified assumptions and architecture of the solution. In Section IV a study in which MaliciousIDE was used as a method to evoke emotions is described. Finally, Section V concludes.

II. Background

So far, a number of research has been conducted in the field of software engineering, which involved recognizing the emotions of IT teams members. Numerous attempts were made to identify emotions using various available channels. However, to the best of our knowledge, there have not yet been proposed methods of inducing emotions developed for the specifics of the work of programmers.

In some studies, general approaches to evoking emotions were used. Khan et al. in their studies tried to induce emotions of software developers by showing video clips, before performing the tasks [4][5].

However, in most studies there were no attempts to evoke emotions at all. In most cases, the researchers only collected

the data useful for recognizing emotions and, at most, obtained information on emotional states based on self-assessment of participants.

The most comprehensive research on utilizing physiological sensors during software developers work was conducted by Müller and Fritz[6][7][8]. During their study [6] on 17 software developers, they collected the following data: electroencephalography (EEG) using a Neurosky MindBand sensor, temperature, electrodermal activity (EDA) and blood volume pulse (BVP) using an Empatica E3 wrist band, and eye-tracking using Eye Tribe. The results of the experiment showed that EDA tonic signal, the temperature, brainwave frequency bands, and the pupil size were most predictive to classify progress of software developers, and brainwave frequency bands, the pupil size, and the heart rate to classify their emotions. Nevertheless, they noted strong individual differences with respect to the correlation and classification of physiological data [6]. Similar differences have also been found in our other studies on the use of sensors to monitor the physiology of computer game players [9].

A relatively new approach, which can be well suited to recognize the emotions of programmers is keystroke dynamics and mouse movements analysis. It is completely non-intrusive and does not require any additional hardware [10]. There have already been attempts to use this method to monitor software developers [2].

However, one of the most popular methods of recognizing emotions, widely used also in the field of software engineering, is the analysis of facial expressions [11][12][13]. It has gained popularity mainly as a universal and non invasive approach. Algorithms analyze video frames to identify face muscle movements and based on the Facial Action Coding System (FACS) [14] assess user's emotional state.

III. Solution

In studies that used real world stimuli methods to evoke emotions, researchers used, among others, widely regarded as horrible, or disgusting animals (spiders, snakes), extreme sports (sky diving, mountaineering), or various types of food [3]. In order to exert an equally significant impact on the emotional state of software developers, we decided to prepare a programmer's nightmare – an unstable programming environment.

An integrated development environment (IDE) is an software which is used to write code, build project and run it. It is a fundamental tool used by the vast majority of programmers. To verify the importance of IDE for the work of programmers and check if its behaviour can affect emotional states, a survey was carried out on a group of 44 software developers, most of whom (61.4%) have professional experience longer than 4 years. Respondents answered three questions regarding the impact of the programming environment on emotions:

- How often do you encounter errors in the IDE?
- Would you change the IDE if it hinders your work?
- Are you getting angry when you can not complete the task because of problems with IDE?

Provided answers, shown in Fig. 1, confirmed that the IDE work and behaviour has an impact on the emotions of programmers. Among all, 68.2% of respondents confirmed that problems with IDE make them angry, and 77.3% would consider changing the IDE in such a situation.

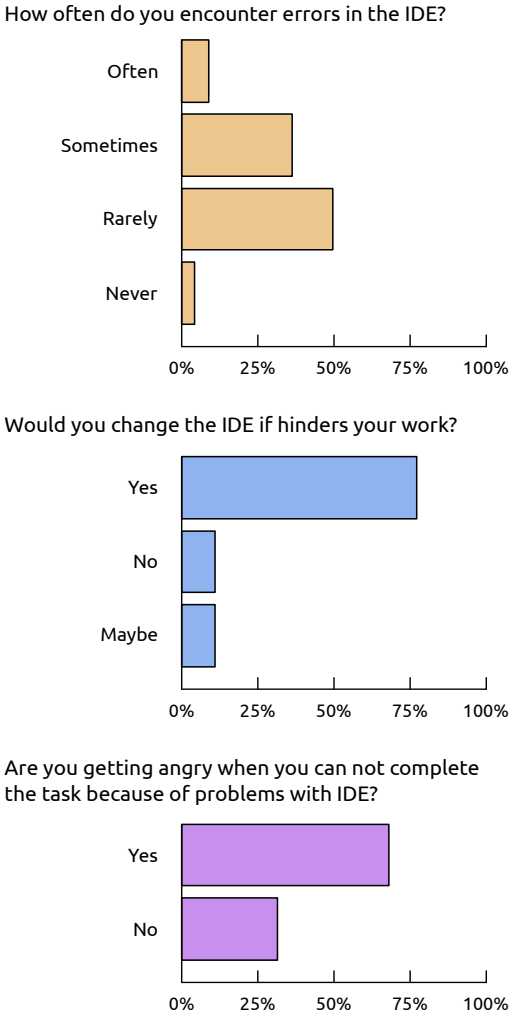


Fig. 1. Survey on the impact of IDE on the emotions of programmers.

In addition, information on the errors encountered by respondents in development environments were collected. It allowed to define seven malices that has been implemented in MaliciousIDE:

1) Clean the clipboard.
2) Delete all semicolons.
3) Freeze whole environment.
4) Hide IDE window.
5) Duplicate keys pressed by participant
6) Minimalize IDE window.
7) Move mouse pointer.

Instead of developing MaliciousIDE from scratch, it was decided to use the plug-in mechanisms, which allows third-party developers to extend the functionality of the program.

The NetBeans was chosen as IDE, on the basis of which malicious behaviour was implemented in the form of a plug-in. It is Open Source, well recognized environment written in Java language.

Preliminary tests of the tool have shown that automatically triggered malices may cause insufficient number of events. For example, users may not notice that the content of the clipboard have been cleaned if it was not used in a particular task. Therefore, it was decided that malices will be triggered manually by the observer, using a remote dashboard (Fig. 2). Thus, an appropriate number of events was ensured. Too few events might not induce emotions, but too much could lead to the disclosure of the malicious activity of the observer.

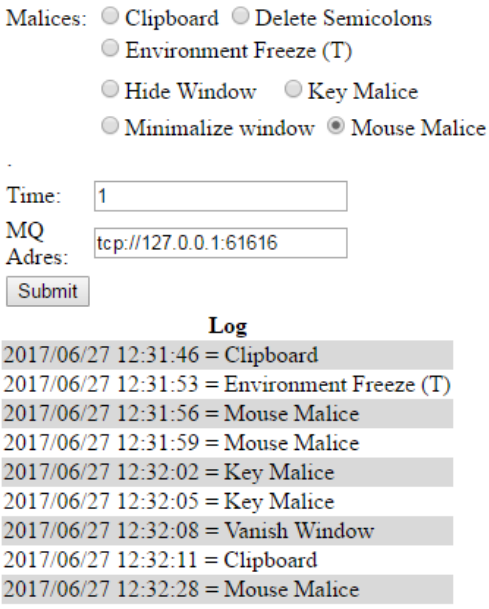


Fig. 2. MaliciousIDE dashboard

Both the plug-in and the dashboard were developed in Java, and in the latter, the Struts 2 framework was additionally used. In order to ensure reliable communication between the two parts of the tool, the Apache ActiveMQ message broker was utilized. MaliciousIDE is released as an Open Source project[1] and can be used freely to conduct research on the emotions of software developers.

IV. Validation

The developed tool, MaliciousIDE, was used during the study conducted in April and May 2017 at a biometric stand [15] in the "Laboratory of Innovative IT Applications" at Gdansk University of Technology (GUT). The aim of the study was to check which input channels can be used to recognize the emotions of programmers during work [16]. Together 35 undergraduate computer science students, including 6 women

[1]http://www.popi.pl/maliciouside

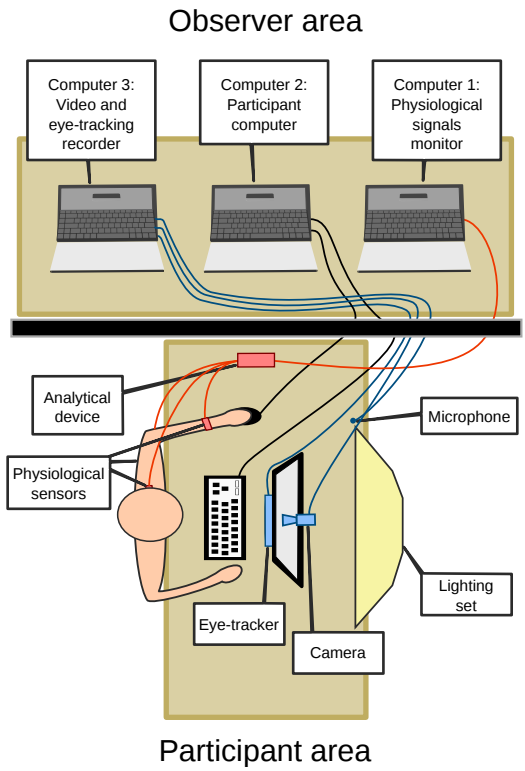


Fig. 3. Study stand

and 29 men, participated in the study. A single session consisted of 5 programming tasks and lasted between 30 and 45 minutes.

During the entire session, the observer was present and controlled MaliciousIDE to evoke the participants' emotions. The most commonly triggered malices were: adding additional characters while entering text, changing the position of the mouse pointer, freezing the environment for 7 seconds, clearing the contents of the clipboard, and temporarily hiding the IDE screen. These actions were carried out to disrupt the work, but in a way that would seem to be a natural behaviour of the application. The frequency of events was manually adjusted so that users remain unaware of the observer's intended actions.

In addition, during the last task of the session, the observer was verbally announcing that the time required to complete the task must be shortened. The purpose of this action was to induce time pressure.

The room was divided into two parts, one for the participant and one for the observer (Fig. 3). On the participant's desk there was a monitor, a keyboard and a mouse connected to the Computer 2, on which the tasks were performed. At the top of the monitor a video camera was located, followed by a lighting set, supplied with Noldus FaceReader software, which was used to recognize emotions based on facial expressions [17]. Underneath the monitor eye-tracker device was located. A number of sensors were attached to the participant body, and were linked through the Coder FlexComp Infiniti by

Thought Technology analytical device with the Computer 1, which was located in the observer area. The observer also has a monitor, a mouse, and a keyboard connected to the participant's Computer 2. A data acquisition program for keystrokes' analysis was also available on the same computer [2]. MaliciousIDE Dashboard was running on Computer 3 and was used to control NetBeans plug-in on participant computer. It also collected all other data for the emotions recognition.

After completing all tasks, the participant was asked to complete a survey implemented using the Google Forms service. Among other questions, participants were asked to name the emotions caused by the unstable IDE. In the answers the most frequently appeared irritation (42.86%), anger (28.57%), followed by nervousness (25.71%) and frustration (11.43%). Four of the respondents (11.43%) indicated amusement (Fig. 4). A post-study informal interview revealed that it was related to the fact that these participants figured out that this unstable work was due to the deliberate action of the observer. The other emotions were pointed out by only one or two people.

On the other hand, the attempt to simulate the approaching deadline by shortening the time of the last task almost completely failed. Nearly half of the respondents indicated that this had no effect on their emotions at all. However, 7 participants indicated that shortening the time was a mobilizing factor, thus having a positive impact on the work attitude.

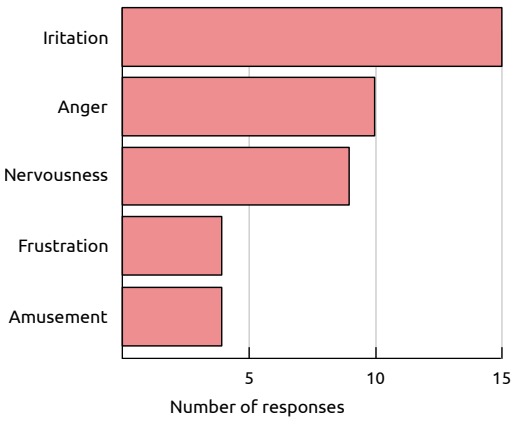


Fig. 4. Emotions induced by malicious behavior of the IDE

V. Summary

MaliciousIDE has been developed to allow researchers to induce the emotions of programmers in a laboratory environment. The presented approach and developed tool can be included to the group of real world stimuli methods of inducing emotions, as it provides a nightmare of programmers – an unstable development environment, which is additionally controlled by the observer.

During the study, emotion inducing method based on the MaliciousIDE has been evaluated. The developed plug-in has proven to be successful mechanism to trigger negative emotions of programmers. While the second approach, shortening the duration of one task, did not turn out to evoke emotions on a significant number of participants.

References

[1] A. Przybyłek and W. Kowalski, "Utilizing online collaborative games to facilitate agile software development," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS'18)*. IEEE, 2018.

[2] A. Kolakowska, "Towards detecting programmers' stress on the basis of keystroke dynamics," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016. doi: 10.15439/2016f263 pp. 1621–1626.

[3] K. S. Quigley, K. A. Lindquist, and L. F. Barrett, "Inducing and measuring emotion and affect," *Handbook of Research Methods in Social and Personality Psychology*, p. 220–252, 2014. doi: 10.1017/CBO9780511996481.014

[4] I. A. Khan, "Mood Independent Programming," in *ECCE '07. Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, no. August, 2007. doi: 10.1145/1362550.1362606 pp. 269–272.

[5] I. A. Khan, W.-P. Brinkman, and R. M. Hierons, "Do moods affect programmers' debug performance?" *Cognition, Technology & Work*, vol. 13, no. 4, pp. 245–258, 2011. doi: 10.1007/s10111-010-0164-1

[6] S. C. Müller and T. Fritz, "Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress," in *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, vol. 1. IEEE, 2015. doi: 10.1109/icse.2015.334 pp. 688–699.

[7] T. Fritz and S. C. Müller, "Leveraging biometric data to boost software developer productivity," in *Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on*, vol. 5. IEEE, 2016. doi: 10.1109/saner.2016.107 pp. 66–77.

[8] S. C. Müller and T. Fritz, "Using (bio) metrics to predict code quality online," in *Proceedings of the 38th International Conference on Software Engineering*. ACM, 2016. doi: 10.1145/2884781.2884803 pp. 452–463.

[9] A. Landowska and M. R. Wróbel, "Affective reactions to playing digital games," in *Human System Interactions (HSI), 2015 8th International Conference on*. IEEE, 2015. doi: 10.1109/hsi.2015.7170678 pp. 264–270.

[10] A. Kołakowska, "A review of emotion recognition methods based on keystroke dynamics and mouse movements," in *Human System Interaction (HSI), 2013 The 6th International Conference on*. IEEE, 2013. doi: 10.1109/hsi.2013.6577879 pp. 548–555.

[11] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005. doi: 10.1016/j.neunet.2005.03.006

[12] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 4. IEEE, 2005. doi: 10.1109/icsmc.2005.1571679 pp. 3437–3443.

[13] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognition*, vol. 47, no. 3, pp. 1282–1293, 2014. doi: 10.1016/j.patcog.2013.10.010

[14] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *Journal of Nonverbal Behavior*, vol. 25, no. 3, pp. 167–185, 2001. doi: 10.1023/A:1010671109788

[15] A. Landowska, "Emotion monitor-concept, construction and lessons learned," in *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*. IEEE, 2015. doi: 10.15439/2015f264 pp. 75–80.

[16] M. R. Wrobel, "Applicability of emotion recognition and induction methods to study the behavior of programmers," *Applied Sciences*, vol. 8, no. 3, p. 323, 2018. doi: 10.3390/app8030323

[17] G. Brodny, A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions," in *Human System Interactions (HSI), 2016 9th International Conference on*. IEEE, 2016. doi: 10.1109/hsi.2016.7529664 pp. 397–404.

Joint 38th IEEE Software Engineering Workshop (SEW-38) and 5th International Workshop on Cyber-Physical Systems (IWCPS-5)

THE IEEE Software Engineering Workshop (SEW) is the oldest Software Engineering event in the world, dating back to 1969. The workshop was originally run as the NASA Software Engineering Workshop and focused on software engineering issues relevant to NASA and the space industry. After the 25th edition, it became the NASA/IEEE Software Engineering Workshop and expanded its remit to address many more areas of software engineering with emphasis on practical issues, industrial experience and case studies in addition to traditional technical papers. Since its 31st edition, it has been sponsored by IEEE and has continued to broaden its areas of interest.

One such extremely hot new area are Cyber-physical Systems (CPS), which encompass the investigation of approaches related to the development and use of modern software systems interfacing with real world and controlling their surroundings. CPS are physical and engineering systems closely integrated with their typically networked environment. Modern airplanes, automobiles, or medical devices are practically networks of computers. Sensors, robots, and intelligent devices are abundant. Human life depends on them. CPS systems transform how people interact with the physical world just like the Internet transformed how people interact with one another.

The joint workshop aims to bring together all those researchers with an interest in software engineering, both with CPS and broader focus. Traditionally, these workshops attract industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practices. This joint edition will also provide a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

TOPICS

The workshop aims to bring together all those with an interest in software engineering. Traditionally, the workshop attracts industrial and government practitioners and academics pursuing the advancement of software engineering principles, techniques and practice. The workshop provides a forum for reporting on past experiences, for describing new and emerging results and approaches, and for exchanging ideas on best practice and future directions.

Topics of interest include, but are not limited to:

- Experiments and experience reports
- Software quality assurance and metrics
- Formal methods and formal approaches to software development
- Software engineering processes and process improvement
- Agile and lean methods
- Requirements engineering
- Software architectures
- Design methodologies
- Validation and verification
- Software maintenance, reuse, and legacy systems
- Agent-based software systems
- Self-managing systems
- New approaches to software engineering (e.g., search based software engineering)
- Software engineering issues in cyber-physical systems
- Real-time software engineering
- Safety assurance & certification
- Software security
- Embedded control systems and networks
- Software aspects of the Internet of Things
- Software engineering education, laboratories and pedagogy
- Software engineering for social media

EVENT CHAIRS

- **Bowen, Jonathan,** Museophile Ltd., United Kingdom
- **Hinchey, Mike**(Lead Chair), Lero-the Irish Software Engineering Research Centre, Ireland
- **Szmuc, Tomasz,** AGH University of Science and Technology, Poland
- **Zalewski, Janusz,** Florida Gulf Coast University, United States

PROGRAM COMMITTEE

- **Ait Ameur, Yamine,** IRIT/INPT-ENSEEIHT, France
- **Banach, Richard,** University of Manchester, United Kingdom
- **Bensalem, Saddek,** VERIMAG, France
- **Broy, Manfred,** Technische Universitaet Muenchen, Germany
- **Čaplinskas, Albertas,** Vilnius University, Lithuania
- **Carter, John,** University of Guelph, Canada
- **Cicirelli, Franco,** Universita della Calabria, Italy
- **Denney, Ewen,** SGT/NASA Ames, United States

- **Derrick, John,** Unversity of Sheffield
- **Ehrenberger, Wolfgang,** Hochschule Fulda, Germany
- **Eleftherakis, George,** The University of Sheffield International Faculty, CITY College, Greece
- **Fantechi, Alessandro,** DSI - Universita' di Firenze, Italy
- **Fidge, Colin,** Queensland University of Technology, Australia
- **Forbrig, Peter,** University of Rostock
- **Fortiers, Stephen,** George Washington University
- **Friesel, Anna,** Technical University of Dennmark, Denmark
- **Fujita, Masahiro,** University of Tokyo, Japan
- **Golatowski, Frank,** University of Rostock, Germany
- **Gomes, Luis,** Universidade Nova de Lisboa, Portugal
- **Gracanin, Denis,** Virginia Tech, United States
- **Grega, Wojciech,** AGH University of Science and Technology, Poland
- **Gumzej, Roman,** Faculty of Logistics, University of Maribor, Slovenia
- **Havelund, Klaus,** Jet Propulsion Laboratory, California Institute of Technology, United States
- **Hsiao, Michael,** Virginia Tech, United States
- **Kornecki, Andrew J.,** Embry Riddle Aeronautical University, United States
- **Laplante, Phillip A.,** PennState University, United States
- **Letia, Tiberiu,** Technical University of Cluj-Napoca, Romania
- **Li, Jianwen,** Iowa State University, United States
- **Liu, Zhiming,** Southwest University, China
- **Lopezo, Oscar Pastor,** Valencia
- **Malloy, Brian,** Clemson University, United States
- **Marwedel, Peter,** Technische Universität Dortmund, Germany
- **Minchev, Zlatogor,** Bulgarian Academy of Sciences, Bulgaria
- **Monostori, László,** Hungarian Academy of Sciences, Hungary
- **Nesi, Paolo,** DSI-DISIT, University of Florence, Italy
- **Obermaisser, Roman,** Universität Siegen, Germany
- **Palanque, Philippe,** ICS-IRIT, University Toulouse 3, France
- **Pu, Geguang,** East China Normal University
- **Pullum, Laura,** Oak Ridge National Laboratory, United States
- **Qin, Shengchao,** Teesside University, United Kingdom
- **Reeves, Steve,** University of Waikato, New Zealand
- **Roman, Dumitru,** SINTEF / University of Oslo, Norway
- **Rouff, Christopher,** Lockheed Martin, United States
- **Rozier, Kristin Yvonne,** NASA Ames Research Center
- **Ryan, Kevin,** Lero-the Irish Software Research Centre, Ireland
- **Rysavy, Ondrej,** Brno University of Technology, Czech Republic
- **Sachenko, Anatoly,** Ternopil National Economic University, Ukraine
- **Sanden, Bo,** Colorado Technical University, United States
- **Seceleanu, Cristina,** Mälardalen University, Västerås, Sweden
- **Sekerinski, Emil,** McMaster University, Canada
- **Selic, Bran,** Simula Research Lab, Norway
- **Sojka, Michal,** Czech Technical University, Czech Republic
- **Sun, Jing,** The University of Auckland, New Zealand
- **Taguchi, Kenji,** AIST, Japan
- **Trybus, Leszek,** Rzeszow University of Technology, Poland
- **van Katwijk, Jan,** Delft University of Technology, The Netherlands
- **Vardanega, Tullio,** University of Padova, Italy
- **Velev, Miroslav,** Aries Design Automation, United States
- **Vilkomir, Sergiy,** East Carolina University, United States
- **Waeselynck, Hélène,** LAAS-CNRS Toulouse, France
- **Zhu, Huibiao,** Software Engineering Institute - East China Normal University
- **Zoebel, Dieter,** University Koblenz-Landau, Germany

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1015–1024
DOI: 10.15439/2018F95
ISSN 2300-5963 ACSIS, Vol. 15

Improved Analogy-based Effort Estimation with Incomplete Mixed Data

Ibtissam Abnane and Ali Idri
Software Project Management Research Team,
ENSIAS, University Mohammed V, Rabat, Morocco
{Ibtissam_abnane, ali.idri }@um5.ac.ma

***Abstract*—Estimation by analogy (EBA) is one of the most attractive software effort development estimation techniques. However, one of the critical issues when using EBA is the occurrence of missing data (MD) in the historical data sets. The absence of values of several relevant software attributes is a frequent phenomenon that may cause inaccurate EBA estimations. The MD can be numerical and/or categorical. This paper evaluates four MD techniques (toleration, deletion, k-nearest neighbors (KNN) imputation and support vector regression (SVR) imputation) over four mixed data sets. A total of 432 experiments were conducted involving four MD techniques, nine MD percentages (from 10% to 90%), three missingness mechanisms (MCAR: Missing Completely at Random, MAR: Missing at Random and NIM: Non-Ignorable Missing) and four data sets. The evaluation process consists of four steps and uses several accuracy measures such as standardized accuracy (SA) and prediction level (Pred).**

The results suggest that EBA with imputation techniques achieved significantly better SA values over EBA with toleration or deletion regardless of the mechanism of missingness. Moreover, no particular MD imputation technique outperformed the other techniques overall. However, according to Pred and other accuracy criteria, EBA with SVR was the best, followed by KNN imputation; we also found that toleration instead of deletion improves the accuracy of EBA.

***Index Terms*—Estimation by analogy, missing data, imputation.**

I. Introduction

SOFTWARE development effort estimation (SDEE) is the process of predicting the effort required to develop a software system. It is a challenging and substantial activity when managing a software project. The challenge arises due to the complex relationship between effort and various software attributes related to the personal, product, and/or platforms used in the project [1], [2].

Machine learning (ML) based estimation techniques are gaining increasing attention in SDEE research, as they can model the complex relationship between effort and software attributes (cost drivers), especially when this relationship is not linear and does not seem to have any predetermined form [2]. Estimation by Analogy (EBA) is one of the most attractive ML techniques in the SDEE field, and is essentially a form of Case-Based Reasoning (CBR)[3]. The idea of analogy based estimation is to determine the effort of the new project as a function of the known efforts from similar historical projects. Wen et al. [2] carried out a systematic literature review of ML SDEE techniques published between 1991 and 2010 and found that EBA is the most investigated ML technique in SDEE (37% of selected studies).

The intensive and increasing use of EBA is due to its several advantages including simplicity, mimicking human reasoning, ease to understand and no assumption is made about the form of the relationship [1], [4]–[10]. Moreover, EBA can handle both quantitative and qualitative data [5]–[7], [11], [12]. Nonetheless, several studies pointed out that EBA still has some limitations such as dealing with missing data (MD) which is a widespread problem that can affect the ability to use data to construct effective EBA techniques [3], [8], [13], [14]. However, little attention has been given to handling missing data in EBA [3]. In fact, the mapping study of Idri et al.]21] found that until 2015, only one paper was published to deal with MD in EBA.

In a prior work [8], we evaluated two EBA techniques in terms of SA criterion on seven data sets when used in conjunction with three MD techniques (toleration, deletion and KNN imputation method), different missingness mechanisms (MCAR, MAR and NIM) and nine percentages of MD (from 10% to 90%) [8]. This was with the aim of determining whether the KNN imputation method, instead of deletion and toleration techniques, could improve the performance of EBA when predicting software development effort with incomplete data. The findings suggest that EBA using KNN imputation outperformed EBA using deletion or toleration regardless of the missingness mechanism and the MD percentage.

However, the study [8] has three limitations: (1) it only dealt with numerical data , (2) it used one imputation technique (KNN), and (3) it used one accuracy criterion (SA), which is insufficient to conclude about EBA accuracy [15], [16].

Thus, this paper improves our previous work with: (1) the use of both numerical and categorical data, 2) the use of a new imputation technique: Support Vector Regression (SVR), in addition to KNN, and (3) the use of a set reliable accuracy criteria (e.g.. Pred(0.25), Mean Absolute Error (MAE), Mean Balanced Relative Error (MBRE), Mean Inverted Balanced Relative Error (MIBRE) and Logarithmic Standard Deviation (LSD)), in addition to SA, in order to investigate if they would confirm the findings of [8].

Therefore, this study carry out an empirical evaluation of EBA using four MD techniques: toleration, deletion, KNN imputation, and SVR imputation with different percentages (from 10% to 90%) and three missingness mechanisms (MCAR, MAR and NIM) over four mixed datasets including

both numerical and categorical attributes (ISBSG R8, COCOMO81, USP05_FT and USP05_RQ).
Toward this aim, four research questions were discussed (RQs):

- (RQ1) Is there evidence that the use of KNN and SVR imputations rather than toleration/deletion improves the accuracy of EBA in terms of SA when using mixed datasets?
- (RQ2) Is there evidence that SVR imputation instead of KNN imputation would improve EBA accuracy measured in terms of SA?
- (RQ3) Is there evidence that the missingness mechanism and the MD percentage affect the accuracy of EBA over mixed datasets?
- (RQ4) Does the performance of EBA in terms of Pred(0.25), MAE, MBRE, MIBRE and LSD confirm the findings of SA?

The structure of the paper is the following: Section II presents the concepts of MD and EBA. Section III describes the data sets used. Section IV presents the empirical design which includes the process of generate MD and the empirical evaluation process. The results are presented and discussed in Section V. Section VI concludes by discussing the findings as well as some directions for future work.

II. BACKGROUND

This section presents the concepts of MD and an overview of the software effort estimation by analogy process we used in this study.

A. Concepts of MD

This section gives an overview of the different missingness mechanisms (i.e., different ways in which data can be missing) and the different MD techniques.

1) Missigness mechanism

Understanding the missing data mechanism is a key stage in comprehending the impact of the missing data on a specific analysis, or missing data methods [17], [18]. Rubin's classification of Missing Data Mechanisms has been regarded as being "fundamental to the modeling of incomplete data" [19] and is in common use in the field of missing data. Little and Rubin classified missing data mechanisms as[17]:

- **Missing completely at random (MCAR)** is when the probability that an observation is missing does not depend on either the observed or the missing values.
- **Missing at Random (MAR)** means that the probability that an observation is missing depends only on the values of the observed data.
- **Non-Ignorable Missing (NIM):** means that the missing data mechanism is related to the missing values.

2) MD techniques

There are three approaches to this problem: MD deletion technique, MD toleration techniques, and MD imputation techniques [13], [20].

a) Toleration

MD toleration technique is an embedded strategy in which analysis is performed directly on the data set with MD[8], [18]. Despite its simplicity, toleration is not a reliable approach, sometimes even providing estimates that are less efficient than estimation from deletion technique [8], [17].

b) Deletion technique

Deletion is the most commonly used technique for dealing with missing data among researchers [8], [21]. It omits all cases with missing values from the analysis and only includes those cases for which all measurements are observed. This method has many advantages since it is easy to use. Also, it produces unbiased estimates for the parameters if the assumption that the data are MCAR holds. Nevertheless, deletion is not generally recommended since omitting cases with MD would result in a significant loss in power and precision due to the reduced sample size. Moreover, if the MCAR assumption does not hold, this method can result in biased parameter estimates as it is ignoring potential systematic differences between the complete and incomplete cases. Consequently, deletion can only be justified if the missing data mechanism in operation is MCAR and the MD percentage is small [22].

c) Imputation technique

MD imputation replaces missing values by suitable estimates and then applies standard complete-data methods to the filled in data [17].This method is attractive to practitioners because the resulting completed data can be handled using standard software for rectangular data sets. Imputation uses available data to impute the missing data and hence, an important characteristic of a good imputation method is that it makes good use of information in the incomplete cases. Moreover, it is important to take into account the missingness mechanism while using imputation technique [8], [18].

B. Effort estimation by analogy: An Overview

EBA is based on the use of historical information from completed projects with known effort [10]. It is based on the following affirmation: *similar software projects have similar costs*. It has been deployed as follows. EBA has been proposed since a long time as a valid alternative to effort estimation by parametric effort estimation and/or expert judgment [23]. In 1997, it has been presented in the form of a detailed estimation methodology and has been applied on a set of historical software projects data sets [10]. EBA consists of three steps:

1) Identification of cases: each project is described by a set of attributes that are believed to be significant in determining similarity and can influence effort.
2) Retrieval of similar cases: several distance metrics can be used to calculate how much the new target project differs from the other projects based on their attribute values. In this study, we used the Euclidean and the overlap distances for numerical and categorical software attributes respectively [10], [24].
3) Case adaptation: involves two phases in order to generate an estimate of the new project. First, we decide on the number of similar projects and second we define the

adaptation strategy. The number of analogues (*k)* refers to the number of most similar projects used to generate the estimation. Several studies in SDEE analyzed the impact of the number of analogues [9], [25], [26]. This study varied the number of analogues between 2 and 5. The second phase consists on selecting the adaptation strategy to provide an effort estimate. We used the arithmetic mean [10], arithmetic median [27] and inverse ranked weighted mean [28].

In order to select the best variant of ABE, we varied the adaptation strategy and the number of analogues as described above and chose the best configuration of ABE, i.e. having the lowest Mean of Absolute Error (MAE).

III. DATA DESCRIPTION

This study uses four available data sets: ISBSG repository (release 8), COCOMO81[23], USP05_FT[9] and USP05_RQ[9]. Table I provides an overview of these datasets, including number of attributes (numerical and categorical), observations, and previous use. The minimum, mean and maximum of effort and size are given.

Since the aim of this study is to deal with missing numerical and categorical data, the solution adopted was to use (1) all the attributes for the COCOMO81 data set, (2) 11 attributes for USP05_FT and USP05_RQ data sets, and (3) 20 attributes for the ISBSG data set. The attributes chosen for USP05_FT, USP05_RQ and ISBSG data sets are the results of our previous studies related to software effort estimation [4][7][8] . Table III shows the attributes chosen for ISBSG, USP05 and COCOMO81 data sets, where (N) and (C) indicate numerical and categorical attributes respectively.

IV. EXPIREMENT DESIGN

This section describes the experimental process used in this study. It consists of four main steps: data removal, complete data set generation, EBA evaluation using SA, and EBA evaluation using Borda count method based on Pred(0.25), MAE, MBRE, MIBRE and LSD. A similar process was followed in [8]. The study was designed to apply EBA with nine percentages of incomplete mixed data (from 10% to 90%), three different MD mechanisms (MCAR, MAR and NIM), and four MD techniques (toleration, deletion, KNN imputation, and SVR imputation) over four data sets. Hence, the experimental design consists in evaluating 9 percentages × 4 MD techniques × 3 MD mechanisms × 4 datasets = 432 different effort estimation experiments.

A. Step 1: Data removal

The first step in the experimental process requires a complete data set to work with. For this purpose, we first preprocessed the four datasets by deleting cases with MD to obtain the corresponding seven complete data sets. We then used the complete datasets to artificially generate MD by mimicking the different mechanisms. A similar approach was followed in [8], [20]. By combining the four datasets, three missingness mechanisms and nine percentages, we obtained 4×3×9 =108 incomplete data sets at this stage.

B. Step 2: Complete data set generation

This step uses four MD techniques: toleration, deletion, KNN imputation and SVR imputation to generate complete datasets from those of Step 1. After applying the four MD techniques on the 108 incomplete datasets of step 1, we obtained 432 complete datasets.

a) Deletion

Under the deletion technique, projects with missing values at any attribute are omitted in the experiments.

b) Toleration

The toleration technique uses a special value *NULL* to replace a missing value in a data set. A similar approach was used in [8], [13], [14]. Hence, the following operations on *NULL* are defined for distance metrics:

$(P_1)\ \delta(\text{b}, \text{NULL}) = \delta(\text{NULL}, \text{b}) = (\text{NULL}, \text{NULL}) = \text{NULL}$

(P_2) W+ NULL = NULL + w = w

where δ is the distance used in Classical Analogy (e.g. Euclidean distance).

It can be seen from the above discussion that the effect of the *NULL* is to ignore the participating attributes that have MD in searching similar objects. Therefore, the more *NULLs* in the data set, the fewer attributes will be participating in searching analogues through similarity measures.

c) KNN imputation

Figure 1 shows the KNN imputation (KNNI) process. KNN imputation belongs to the analogy based algorithms; it is computationally simple and has proven to be effective approach to estimate missing values of attributes in different software engineering datasets [21]. Using KNN for imputation requires adapting the three analogy steps of Section II.B: (1) Identification of cases, (2) Retrieval of similar projects, and (3) Case adaptation. In the following, we present how we adapted each step to serve the imputation process.

The identification of cases step mainly aims to calculate the distance between each incomplete project and the complete projects. The most similar complete projects were used as source analogues in the imputation process. To determine the distance between the incomplete project and the complete projects, we use a combination of two distance measures. Hence, the distance between an incomplete case P_i and a complete case P_j is calculated using Equation (1).

$$\mathbf{d}(\mathbf{P_i}, \mathbf{P_j}) = \mathbf{d_n}(\mathbf{P_i}, \mathbf{P_j}) + \mathbf{d_c}(\mathbf{P_i}, \mathbf{P_j}) \qquad (\mathbf{1})$$

where:

- $d_n(P_i, P_j)$is the Euclidean distance used to calculate the similarity between P_i and P_j taking into consideration only the numerical attributes. It is calculated using Equation (2):

$$d_n(P_i, P_j) = \sqrt{\sum_{l=1}^{n}(P_{il} - P_{jl})^2} \qquad (2)$$

- $d_c(P_i, P_j)$ is the hamming distance used to evaluate the similarity between P_i and P_j taking into consideration only the categorical attributes. The formula of $d_c(P_i, P_j)$ is given by Equation (3):

$$d_c(P_i, P_j) = \sum_{l=1}^{n} \delta(P_{il}, P_{jl}) \qquad (3)$$

TABLE I Selected Software Attributes from ISBSG, USP05, and COCOMO81 Data Sets. (N: numerical and C: categorical)

ISBSG		USP05	COCOMO81	
Value adjustment factor (N)	Reference table approach (C)	Data file (N)	KDSI (N)	VEXP (C)
Maximum team size (N)	Recording method (C)	Data entry (N)	RELAY (C)	LEXP (C)
User-based business units (N)	Development platform (C)	Data output (N)	DATA (C)	MODP (C)
User-based locations (N)	Programming language (C)	Unadjusted function point (N)	CPLX (C)	TOOL (C)
User-based concurrent users (N)	Used methodology (C)	Internal complexity (C)	TIME (C)	SCED (C)
Input count (N)	Development technique (C)	Language (C)	STOR (C)	MODE (C)
Output count (N)	Organization type (C)	Tools (C)	VIRT (C)	
Enquiry count (N)	Business area type (C)	Applications experience (C)	TURN (C)	
Interface count (N)	Application type (C)	Database systems (C)	ACAP (C)	
Measurement technique (C)		Methodology (C)	AEXP (C)	
		Application type (C)	PCAP (C)	

TABLE II Description Statistics of the Selected Data Sets

Data set	#of Projects	#of attributes	Effort					
			Min	Max	Median	Mean	Skewness	Kurtosis
USP05RQ	102	11	0.5	50	3	8.05	2.01	3.60
USP05FT	58	11	0.5	24	1	3.21	3.03	8.84
ISBSG	89	20	24	36286	2101	3779.52	3.49	14.74
COCOMO81	63	17	6	11400	98	683.44	4.47	21.87

where $\delta(P_{i1}, P_{j1})$ is the hamming distance between P_{i1} and P_{j1}. The Hamming distance between two sets of binary digits is the number of corresponding binary digit positions that differ divided by the number of comparisons made [31]

The case adaptation step matches the imputation phase. First, we decide on the number of analogous projects, *k*. we varied k from 1 to 5. Thereafter, to impute the missing values, we had to decide also on the adaptation strategy. For numerical attributes, we choose the weighted mean since it allows the higher similar projects to have more influence than the lower ones. For categorical attributes, we imputed a missing value with the attribute value of the most similar project to the incomplete project.

d) SVR imputation

Support vector machine has been developed by [32] and it is a supervised learning approach based on statistical theory. It has been gaining popularity due to its attractive features and promising empirical performance. Based on the structural risk minimization (SRM) principle, SVM is able to control the complexity of the model and its generalization ability, which can be used for solving two-class or multi-class classification and regression problems in various fields [33]–[35]. SVM possesses many advantages including fast-learning, global optimization, and excellent generalization abilities due to minimizing the tradeoff between the complexity of the model and its generalization ability compared with other approaches such as artificial neural networks [36], linear regression and radial Basis functions neural networks (RBFNs) [37].

With the introduction of Vapnik's ε-insensitivity loss function, the regression model of SVMs, called support vector regression (SVR), has received increasing attention to solve nonlinear regression problems. The investigation of SVR for software development effort estimation was originally carried out by Oliveira [37]. They found that SVR outperforms both linear regression and RBFNs for software effort estimation over NASA data set.

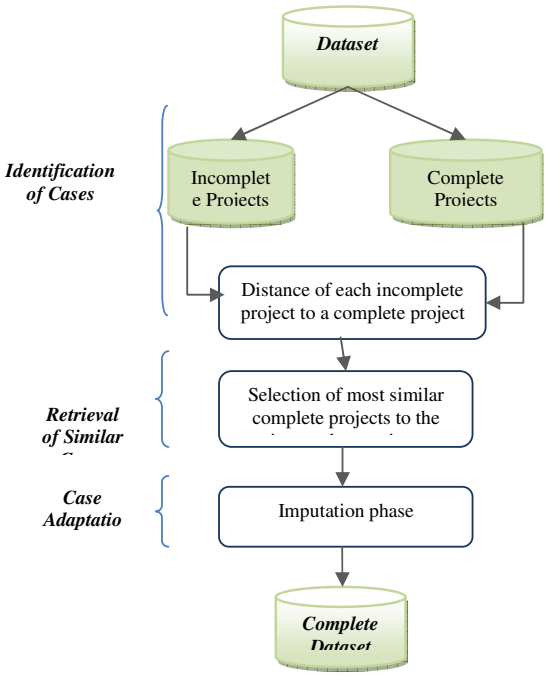


Figure 1 KNN imputation process

The main challenge when dealing with SVR is how to solve the dual problem [38]. The traditional quadratic programming (QP) algorithms solvers are slow, particularly for large problems[34], [35]. In addition, those algorithms can be complex, subtle, and difficult for an engineer to implement [34].

Specific algorithms were developed in order to make easier the use of SVR, such as Vapnik's chunking [39] and Osuna's decomposition [40]. Those algorithms make the training of SVR possible by breaking the large QP problem into a series of smaller QP ones and optimizing only a subset of training data patterns at each step. The subset of training data patterns optimized at each step is called the working set. Thus, these approaches are categorized as the working set methods. Based on the idea of the working set methods, [34] proposed the Sequential Minimal Optimization (SMO) algorithm that selects the size of the working set as two and uses a simple analytical approach to solve the reduced smaller QP problems. Thereafter, [38] ascertained inefficiency associated with Platt's SMO and suggested a modified version of SMO that can solve the SVR QP problem without any extra matrix storage and without using numerical QP optimization steps at all. Hence, this work uses the SMO-SVR algorithm of [38]. Moreover, An important factor that influences the performance of SVR is how to adequately select model parameters (C, ε, γ), which play an important role for a good generalization performance [41]. This paper uses a selection methodology based on Particle Swarm Optimization (PSO) to search global solutions of the optimal parameters (ε, C, γ) [42], [43].

Before proceeding to imputation, SVR transforms categorical variable into numerical ones. In fact, SVR maps each possible value for a categorical attribute into a number.

Unlike KNN imputation technique, SVR imputation (SVRI) requires building a model for each missing value in the dataset. Fig.2 shows the imputation method based on SMO-SVR. Let X be a N×D matrix of N projects described by d *attributes*. For each attribute i, we construct:

1. A complete dataset, Complete_X_i containing all projects P_j for which the values $x_{j,i}$ were not missing.
2. An incomplete dataset, Incomplete_X_i, containing all projects P_j for which the values $x_{j,i}$ were missing

Next, the i^{th} attribute X_i is set as the dependent attribute of Complete_X_i and Incomplete_X_i datasets. Then, the SMO-SVR model is trained using Complete_X_i. Firstly, the PSO algorithm is used to determine the optimal values of ε, C and γ. Next, those optimal values were used to train the SMO-SVR model. Finally, the missing values were imputed by the SMO-SVR model by using the Incomplete_X_i as the test set.

C. Step 3: EBA accuracy evaluation using SA

The accuracy of EBA was assessed using the Jackknife method in which the target project is excluded from the historical dataset and its effort estimation is calculated using the actual effort values of the remaining projects[44].

The accuracy of EBA was assessed in four steps:

1) Evaluation using SA

The first evaluation step aims to compare the accuracy of EBA with random guessing using the Standardized Accuracy (SA) suggested by Shepperd and MacDonell [45]. SA evaluates how good a SDEE technique is in comparison to random guessing. It is based on the Mean of Absolute Error (MAE) and is defined by Equation (4):

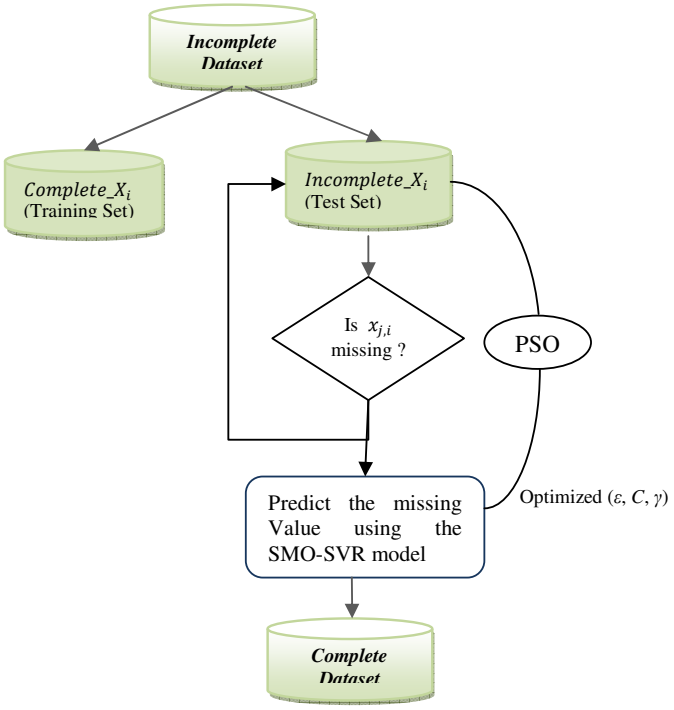


Figure 2 SVR imputation process

$$SA_{P_i} = 1 - \frac{MAE_{P_i}}{\overline{MAE}_{p0}} \times 100 \quad (4)$$

where MAE_{P_i} is defined to be the MAE of the estimation method P_i and MAE_{P_0} is the mean of a large number of (in our case 1000) random guessing. In the random guessing procedure, a training instance is randomly chosen with equal probability from the training set and its effort value is used as the estimate of the test instance. SA gives us an idea of how good an estimation method is in comparison to random guessing. Since the term MAE_{P_i} is in the nominator, the higher the SA values, the better an estimation method is.

The interpretation of SA is that the ratio represents how much better it is as a predictive model (P_i) than the mean or random guessing (P_0). A value close to zero is discouraging and a negative value would be worrisome. The positive sign of SA means the predictive models are better than mean or random guessing. Meanwhile the negative sign is shows how bad the predictive models are against the mean as an estimator. Unlike MRE-based error measures which have been criticized for being biased and favoring underestimation, SA is an unbiased and standardized accuracy measure.

2) Hypothesis testing

The second step aimed to statistically investigate the significance of the results found in step 1. Awareness about statistical validation of the published results had increased among machine learning researchers, in particular is software effort estimation [46].

Hypothesis testing is the process of inferring from a sample whether or not a given statement about the population appears to be true [47]. The first step in hypothesis testing is

establishing a null hypothesis. The null hypothesis is typically a statement contrary to what the researcher is attempting to confirm; we assume the null hypothesis to be true, and use data to try and refute it [47]. The statement that the researcher would like to prove is called the alternative hypothesis. We often establish a significance level (i.e. α-levels). It is is a limit on how unusual a result we will accept. An α-level of 0.05 means that if our observations from our collected data would occur less than 5% of the time given that the null hypothesis is true, then we will reject the null hypothesis [47].

However, null hypothesis testing is not sufficient to analyze and interpret data. A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty of the size of an effect [45],[48]. The American Psychological Association (2001) has suggested that researchers report the confidence interval for research data. A confidence interval is an inference to a population in terms of an estimation of sampling error. More specifically, it provides a range of values that fall within the population with a level of confidence of 100*(1 - α) % (i.e. the level of confidence is 95% for α =0.05). Confidence intervals (CIs) offer much more information and allow us to move beyond dichotomous thinking and adopt an "*estimation thinking"*. Estimation thinking focuses on how big an effect is; this is usually more valuable than knowing whether or not the effect is zero, which is the focus of dichotomous thinking. Confidence intervals convey information about magnitude and precision of effect simultaneously, keeping these two aspects of measurement closely linked [49]–[51].

This study used the Wilcoxon statistical test which is a non-parametric procedure used to test whether there is sufficient evidence that the median of two probability distributions differ in location [52]. All statistical tests were two-sided and performed at α=0.05 significance level. Confidence intervals were calculated using Hodges-Lehmann estimates of shift [53], [54]. To adjust for multiple testing, we used the Holm-Bonferroni method [55].

3) Effect size results

To verify how meaningful is the improvement and how important are the results, the effect size criterion defined by Equation (5) was used:

$$\mathbf{\Delta} = \frac{MAR - \overline{\mathrm{MAR}_{\mathrm{P}_0}}}{SP_0} \quad (5)$$

where SP_0 is the sample standard deviation of the random guessing. The Δ values can be interpreted in terms of the categories proposed by Cohen [56] of small (around 0.2), medium (around 0.5) and large (around 0.8). A medium or large value of Δ indicates an acceptable degree of confidence on the model predictions over random guessing.

D. Step 4: EBA accuracy evaluation using Borda Count

The use of SA enabled us to explore the influence of the missingness mechanisms and MD percentages and techniques on the prediction ability of the EBA technique. In fact, SA determines if EBA is reasonable (i.e., is actually predicting, and how much better is it than random guessing), but does not evaluate its accuracy [16]. Thus, SA alone is not sufficient to conclude about EBA accuracy and should be used with other metrics [15], [16].

Hence, we evaluated The EBA technique using a set of reliable accuracy measures that are believed to be less sensitive to bias and asymmetry. These measures are: Pred(0.25), Mean Absolute Error (MAE), Mean Balanced Relative error (MBRE), Mean Inverted Balanced Relative Error (MIBRE) and logarithmic standard deviation (LSD) as shown in Equations (8)–(12), respectively. Using a set of accuracy measures would ensure that different aspects are captured and would give more confidence in the results obtained compared with using only one accuracy measure. Similar approach was used in [57]–[59].

We evaluated the EBA variants (i.e. with four MD techniques and three missingness mechanisms) according to those performance measures and used Borda counting method to rank them over the four datasets in order to identify which variants were the most accurate. The Borda count method was used for the first time in SDEE by Azzeh et al. [59] and then by Idri et al. [58] and it allows to take into consideration different aspects of prediction performance since it is based on five performance measures.

$$AE_i = |e_i - \hat{e}_i| \quad (6)$$

$$MRE = \frac{AE_i}{e_i} \quad (7)$$

$$Pred(0.25) = \frac{100}{N}\sum_{i=1}^{N} \begin{cases} 1 \; if \; MRE_i \leq 0.25 \\ 0 \; otherwise \end{cases} \quad (8)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} AE_i \quad (9)$$

$$MBRE = \frac{1}{N}\sum_{i=1}^{N} \frac{AR_i}{\min(e_i, \hat{e}_i)} \quad (10)$$

$$MIBRE = \frac{1}{N}\sum_{i=1}^{N} \frac{AR_i}{\max(e_i, \hat{e}_i)} \quad (11)$$

$$LSD = \sqrt{\frac{\sum_{i=1}^{n}(\lambda_i + \frac{s^2}{2})^2}{n-1}} \quad (12)$$

where:

- e_i and $\hat{e}_i$ are the actual and predicted effort for the *i*th project.
- $\lambda_i = \ln(e_i) - \ln(\hat{e}_i)$
- s^2 is an estimator of the variance of the residual λ_i.

V. RESULTS

This section presents and discusses the experimental results when evaluating the accuracy of EBA using MD.

A. EBA Evaluation using SA (RQs 1-3)

The first objective is to investigate the effect of the MD techniques on the accuracy of EBA in terms of SA. Figures 3a-c show the median SA values for EBA applied to the four data sets with three mechanisms of missingness, different MD percentages and four MD techniques. In general, we notice that the SA values decrease as the MD percentage increases regardless the mechanism of missingness.

For the MCAR mechanism, we observe that the imputation techniques outperformed toleration and deletion (SA values at 10% of MD: 41% for SVRI, 41% for KNNI, 37% for toleration and 34% for deletion). Moreover, Fig.3-a shows that toleration outperformed deletion (SA values at 10% of MD: 37% for toleration and 34% for deletion). Note that SVR and KNN imputations performed almost the same.

For MAR mechanism, Fig.3-b shows that the imputation techniques have the same performance (SA values at 10% of MD: 35% for SVRI, 34% for KNNI). Moreover, they both outperformed toleration and deletion (SA values at 10% of MD: 31% for Toleration, 31% for Deletion). Moreover, we notice that toleration and deletion gave the same performance.

As for NIM mechanism, we notice that the imputation techniques gave similar performances (SA values at 10% of MD: 32% for SVRI and 32% for KNNI). Moreover, toleration and deletion performed the same (SA values at 10% of MD: 26% for toleration and 27% for deletion).

1) Hypothesis testing

In the previous section, we found that imputation techniques (KNN or SVR) instead of toleration and deletion improved the performance of EBA. This section investigates whether this improvement is statistically significant. Moreover, we investigate whether the improvement varies with the mechanisms of missingness. To do that, we compared the median of SA values across data sets for each MD percentage using the Wilcoxon t-test. We drew the following hypothesis:

NH1: The prediction performance of EBA is not affected by the MD technique.

NH2: The prediction performance of EBA when using MD techniques is not affected by the mechanism of missingness.

Each null hypothesis was evaluated separately for the MD techniques and the three missingness mechanisms. Tables II and III sum up the results of the Wilcoxon t-test conducted to evaluate NH1 and NH2 respectively, where p(α) denotes the p-value of the Wilcoxon test, α' denotes the significance level corrected by Holm-Bonferroni correction and CI denotes the confidence interval. Table II shows the results of Wilcoxon test on NH1. We notice that the difference between SVR and KNN imputations is not significant regardless of the mechanism of missingness. The confidence intervals also reflect this finding. We notice that for MCAR and MAR mechanisms, the confidence intervals have negative values. This means that no imputation technique is always superior to the other. For MCAR mechanism, we observe that the difference between KNN and SVR is between -1.679 and 0.416. This means that KNN outperformed SVR by 1.679 and SVR outperformed KNN by 0.416. The case of MAR is the similar to MAR. However, under NIM, we observe that SVR outperformed KNN and the magnitude of the difference is between 0.484 and 0.7

Table II also shows that imputation techniques often significantly outperformed toleration and deletion. We notice that the magnitude of the difference is larger with deletion compared to toleration. The improvement given by the imputation techniques over toleration/deletion is larger when using MCAR or NIM compared to MAR. This is due to the fact that, under MCAR mechanism, imputation techniques outperformed largely toleration and deletion. Moreover, under NIM mechanism, toleration and deletion gave the worst results (negative values of SA). Under MAR mechanism, the performance of imputation techniques decreased and the toleration/deletion gave acceptable performance which explains the small CI.

To evaluate the impact of the missingness mechanisms on the accuracy of EBA, Table III reports the results of the statistical tests of NH2; it can be noticed that:

- The difference between MCAR and MAR is significant when using SVR, KNN and toleration. However, the difference is not significance when using deletion.
- The difference between MCAR and NIM is in general significant.
- The difference between MAR and NIM is significant for SVR, toleration and deletion. However, it is not significant in the case of KNN.

2) Effect size results

To ensure that the results are not generated by chance and to assess if there is effect improvement over random guessing, we evaluate the effect size measured by means of Equation (5). Table VII reports the median values of the effect size Δ of EBA using the four MD techniques under three mechanisms of missingness and nine MD percentages across the four datasets where the baseline method is random guessing.

From Table VII, we notice that the Δ values are higher than 0.5, which means that the results obtained by EBA in terms of SA are more likely not to be due to chance under:

- MCAR mechanism when using: 1) imputation, 2) toleration/deletion with MD percentage less than 80%.
- MAR mechanism when using: 1) KNN imputation with MD percentage less than 90%, 2) SVR imputation, toleration or deletion with MD percentage less than 80%.
- NIM mechanism when using: 1) SVR/KNN imputation with MD percentage less than 90%, 2) toleration with MD percentage less than 60%, 3) deletion with MD percentage less than 70%.

B. EBA Evaluation using Borda Count (RQ4)

Although SA results would confirm if EBA outperform random guessing, they are not sufficient to conclude about the accuracy of EBA [15], [16]. This section evaluates EBA with four MD techniques and three missingness mechanisms using a set of reliable performance measures as explained in Section V.C. Thereafter, we rank the four variants of EBA (i.e. with the four MD techniques) using the Borda count method. Table VI shows the ranking over the four datasets.

We notice that EBA with SVR imputation generally outperformed the other EBA variants except for USP05FT under MAR and USP05RQ under NIM.

Moreover, we notice that toleration generally improves the EBA accuracy compared to deletion except for ISBSG under MAR/NIM and USP05FT under NIM.

Furthermore, we compared the four EBA variants across the four datasets. Table VII shows the results of this ranking. We notice that EBA with SVR imputation was the best, followed by KNN imputation, toleration and lastly deletion regardless the missingness mechanisms.

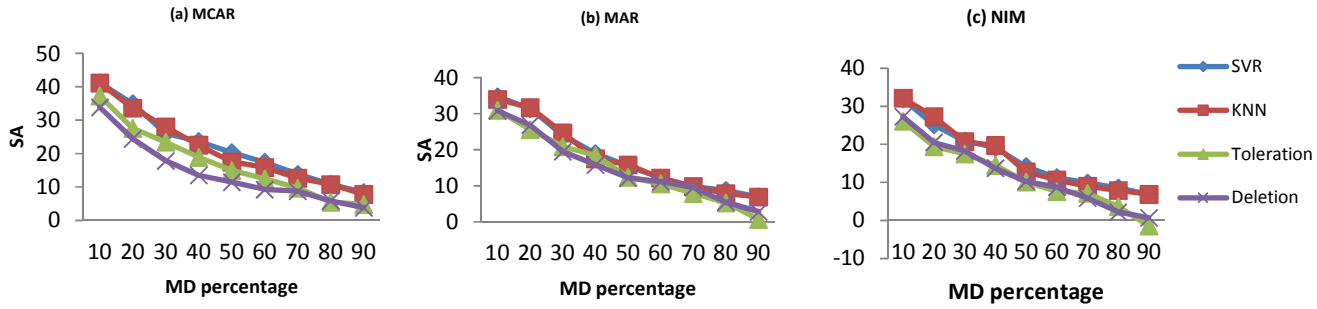


Figure 3(a-c) SA values of Classical Analogy applied to four data sets with three mechanisms of missingness, different MD percentages and four MD techniques

TABLE III Results of Wilcoxon test of NH2

		KNN				Toleration				Deletion			
		p(α)/ α'	Z	CI		p(α)/ α'	Z	CI		p(α)/ α'	z	CI	
				Min	Max			Min	Max			Min	Max
MCAR	SVR	0.123/0.05	1.540	-1.679	0.416	0.008/0.0167	2.666	3.688	5.783	0.008/0.01	2.666	5.097	9.408
	KNN					0.008/0.0125	2.666	2.996	4.911	0.008/0.0083	2.666	4.879	9.060
	Toleration									0.011/0.025	2.547	1.012	4.491
MAR	SVR	0.499/0/025	0.676	-0.766	0.362	0.008/0.0125	2.666	1.842	4.759	0.008/0.01	2.666	1.780	4.217
	KNN					0.011/0.0167	2 .547	1.264	4.814	0.008/0.0083	2.666	1.387	4.438
	Toleration									0.767/0.05	0.296	-1.299	1.230
NIM	SVR	0.790/0.05	0.757	0.484	0.700	0.008/0.0167	2.666	3.530	6.160	0.008/0.01	2.666	3.262	5.508
	KNN					0.008/0.0125	2.666	2.836	6.782	0.008/0.083	2.666	2.509	6.069
	Toleration									0.515/0.025	0.652	-0.732	1.206

TABLE IV Results of Wilcoxon test of NH1

		MAR				NIM			
		p(α)/ α'	Z	CI		p(α)/ α'	Z	CI	
				Min	Max			Min	Max
SVR	MCAR	0.008/0.0167	2.666	2.126	5.292	0.008/0.025	2.666	3.088	7.783
	MAR					0.038/0.05	2.073	0.004	3.222
KNN	MCAR	0.008/0.0125	2.666	1.909	5.071	0.008/0.025	2.666	2.855	6.884
	MAR					0.086/0.05	1.718	-0.200	3.108
Toleration	MCAR	0.008/0.05	2.666	1.121	4.135	0.008/0.025	2.666	3.350	8.043
	MAR					0.008/0.0167	2.666	1.828	4.558
Deletion	MCAR	0.314/0.05	1.007	-1.050	2.001	0.021/0.025	2.310	0.381	3.919
	MAR					0.008/0.0167	2.666	1.900	4.194

TABLE V Median values of effect size of EBA across the four datasets based on comparison with random guessing baseline method

	MCAR				MAR				NIM			
	SVRI	KNNI	Toleration	Deletion	SVRI	KNNI	Toleration	Deletion	SVRI	KNNI	Toleration	Deletion
10%	-2.34	-2.43	-2.1	-2.06	-2.28	-2.49	-1.98	-1.93	-2.09	-2.4	-1.76	-1.76
20%	-2.15	-1.89	-1.82	-1.53	-1.95	-2.07	-1.49	-1.68	-1.84	-1.72	-1.49	-1.49
30%	-2	-1.58	-1.53	-1.27	-1.74	-1.6	-1.27	-1.32	-1.48	-1.33	-1.19	-1.28
40%	-1.62	-1.27	-1.27	-0.91	-1.45	-1.51	-1.08	-1.18	-1.19	-1.28	-0.82	-1.01
50%	-1.29	-0.99	-1.04	-0.82	-1.04	-1.16	-0.81	-0.94	-1.06	-0.96	-0.63	-0.76
60%	-1.08	-0.89	-0.9	-0.67	-0.85	-0.94	-0.71	-0.83	-0.82	-0.75	-0.48	-0.63
70%	-0.83	-0.8	-0.65	-0.63	-0.69	-0.67	-0.55	-0.7	-0.71	-0.64	-0.46	-0.43
80%	-0.64	-0.71	-0.4	-0.43	-0.45	-0.63	-0.4	-0.4	-0.59	-0.56	-0.2	-0.16
90%	-0.56	-0.55	-0.28	-0.26	-0.35	-0.47	-0.06	-0.18	-0.4	-0.47	-0.09	-0.05

TABLE VI Borda Count ranking of the four MD techniques under the three missingness mechanisms.

	EBA											
	MCAR				MAR				NIM			
	SVRI	KNNI	Toleration	Deletion	SVRI	KNNI	Toleration	Deletion	SVRI	KNNI	Toleration	Deletion
ISBSG	1	3	2	4	1	3	4	2	1	2	4	3
COCOMO81	1	2	3	4	1	3	2	4	1	2	3	4
USP05FT	1	2	3	4	2	1	3	4	1	2	4	3
USP05RQ	1	2	3	4	1	3	2	4	2	1	3	4

TABLE VII Borda Count ranking of the four MD techniques under the three missingness mechanisms across data sets.

	SVRI	KNNI	Toleration	Deletion
MCAR	1	2	3	4
MAR	1	2	3	4
NIM	1	2	3	4

VI. Conclusion and Future Work

This study evaluated EBA using four MD techniques: toleration, deletion, KNN imputation, and SVR imputation with different percentages (from 10% to 90%) and three missingness mechanisms (MCAR, MAR and NIM) on four datasets (ISBSG R8, COCOMO81, USP05_FT and USP05_RQ) with mixed data (numerical and categorical).

four research questions RQs 1-4 have been discussed. The findings when answering RQs 1-4 are as follows:

(RQ1): Is there evidence that the use of KNN and SVR imputations rather than toleration/deletion improves the accuracy of EBA in terms of SA when using mixed datasets?

We found that EBA with imputation techniques achieved significantly better SA values over EBA with toleration or deletion regardless of the mechanism of missingness.

EBA with toleration provided significantly better SA over EBA with deletion when the missingness mechanism is MCAR. However, under MAR or NIM mechanisms, the improvement provided by toleration over deletion is not significant. Moreover, EBA outperformed random guessing when using imputation techniques. However, when using toleration/deletion at high percentages of MD, EBA underperformed random guessing.

(RQ2): Is there evidence that SVR imputation instead of KNN imputation would improve EBA accuracy measured in terms of SA?

In terms of SA, we found that the performance difference between the imputation techniques KNN and SVR was not significant.

(RQ3): Is there evidence that the missingness mechanism and the MD percentage affect the accuracy of EBA measured in terms of SA over mixed datasets?

We found that EBA with MCAR instead of MAR achieved significantly better SA values when using imputation or toleration. However, when using deletion, the improvement of EBA with MCAR instead MAR is not significant. EBA with MCAR instead of NIM achieved significantly better SA values regardless of the MD technique used. EBA with MAR instead of NIM achieved significantly better SA values when using toleration/deletion. However, when using imputation, the difference is not significant.

(RQ4): Does the performance of EBA in terms of Pred(0.25), MAE, MBRE, MIBRE and LSD confirm the findings of SA?

When using the Borda count based on five accuracy criteria, we found that EBA with SVR was the best, followed by KNN imputation. We also notice that toleration instead of deletion improves the accuracy of EBA. These findings confirm those of [14], [15] stating that different measures capture different aspects of EBA performance: SA determines if a technique is reasonable (i.e., is actually predicting, and how much better is it than random guessing) while other accuracy metrics measure how close a prediction is to its correct value.

Future work aims to confirm our finding with different variants of EBA techniques (e.g. Fuzzy Analogy) and other software effort estimation techniques. Moreover, other imputation techniques may give different results.

References

[1] S. K. Sehra, Y. S. Brar, N. Kaur, and S. S. Sehra, "Research patterns and trends in software effort estimation," *Inf. Softw. Technol.*, vol. 91, p. , 2017.

[2] J. Wen, S. Li, Z. Lin, Y. Hu, and C. Huang, "Systematic literature review of machine learning based software development effort estimation models," *Inf. Softw. Technol.*, vol. 54, no. 1, pp. 41–59, 2012.

[3] A. Idri, F. A. Amazal, and A. Abran, "Analogy-based software development effort estimation: A systematic mapping and review," *Inf. Softw. Technol.*, vol. 58, pp. 206–230, 2014.

[4] F. A. Amazal, A. Idri, and A. Abran, "Improving Fuzzy Analogy based Software Development Effort Estimation," in *21st Asia-Pacific Software Engineering Conference (APSEC)*, 2014, pp. 1–4.

[5] F. A. Amazal, A. Idri, and A. Abran, "An analogy-based approach to estimation of software development effort using categorical data," in *Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process* and Product Measurement, 2014, pp. 252–262.

[6] A. Idri and A. Abran, "Evaluating software project similarity by using linguistic quantifier guided aggregations," *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference,* Volume: 1, 2001, pp. 470 - 475.

[7] A. Idri, F. A. Amazal, and A. Abran, "Accuracy Comparison of Analogy-Based Software Development Effort Estimation Techniques," *Int. J. Intell. Syst.*, vol. 31 (2), pp. 128–152, 2016.

[8] A. Idri, I. Abnane, and A. Abran, "Missing data techniques in analogy-based software development effort estimation," *J. Syst. Softw.*, vol. 117, pp. 595–611, 2016.

[9] J. Li, G. Ruhe, A. Al-Emran, and M. M. Richter, "A flexible method for software effort estimation by analogy," *Empir. Softw. Eng.*, vol. 12, no. 1, pp. 65–106, 2007.

[10] M. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Trans. Softw. Eng.*, vol. 23, no. 12, pp. 736–743, 1997.

[11] A. Idri and I. Abnane, "Fuzzy Analogy Based Effort Estimation: An Empirical Comparative Study," in *IEEE International Conference on Computer and Information Technology (CIT)*, 2017, pp. 114–121.

[12] F. A. Amazal, A. Idri, and A. Abran, "Software Development Effort Estimation Using Classical and Fuzzy Analogy: a Cross-Validation Comparative Study," *Int. J. Comput. Intell. Appl.*, vol. 13, no. 3, p.

1450013, 2014.
[13] J. Li, A. Al-Emran, and G. Ruhe, “Impact Analysis of Missing Values on the Prediction Accuracy of Analogy-based Software Effort Estimation Method AQUA,” First Int. Symp. Empir. Softw. Eng. Meas. (ESEM 2007), pp. 126–135, 2007.
[14] I. Abnane and A. Idri, “Evaluating Fuzzy Analogy on Incomplete Software Projects data,” in IEEE Symposium Series on Computational Intelligence (SSCI), 2016.
[15] M. Azzeh and A. B. Nassif, “A hybrid model for estimating software project effort from Use Case Points,” Appl. Soft Comput. J., pp. 1–9, 2016.
[16] A. Idri, I. Abnane, and A. Abran, “Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation,” J. Softw. Evol. Process, no. September, 2017.
[17] R. J. A. Little and D. . Rubin, “Statistical Analysis with Missing Data,” Wiley, New York., 1987.
[18] D. . Little, R.J.A., Rubin, “Analysis of social science data with missing values,” Sociol. Methods Res., pp. 292–326, 1989.
[19] G. Molenberghs and M. G. Kenward, Missing Data in Clinical Studies, vol. 61. John Wiley & Sons, 2007.
[20] Q. Song, M. Shepperd, X. Chen, and J. Liu, “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation,” J. Syst. Softw., vol. 81, no. 12, pp. 2361–2370, 2008.
[21] A. Idri, I. Abnane, and A. Abran, “Systematic Mapping Study of Missing Values Techniques in Software Engineering Data,” in International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS, 2015, pp. 1–8.
[22] J. Schafer, Analysis of Incomplete Multivariate Data. 1997.
[23] B. W. Boehm, “Software Engineering Economics,” IEEE Trans. Softw. Eng., vol. SE-10, no. 1, 1984.
[24] L. C. Briand, K. El Emam, D. Surmann, I. Wieczorek, and K. D. Maxwell, “An assessment and comparison of common software cost estimation modeling techniques,” Proc. 21st Int. Conf. Softw. Eng. - ICSE ’99, pp. 313–322, 1999.
[25] E. Mendes, “A Comparative Study of Cost Estimation Models for Web Hypermedia Applications,” Empir. Softw. Eng., vol. 8, no. 2, pp. 163–196, 2003.
[26] M. Azzeh and Y. Elsheikg, “Learning Best K analogies from Data Distribution for Case-Based Software Effort Estimation,” in The Seventh International Conference on Software Engineering Advances, 2012, no. 2, pp. 341–347.
[27] L. Angelis and I. Stamelos, “A Simulation Tool for Efficient Analogy Based Cost Estimation,” Empir. Softw. Eng., vol. 5, no. 1, pp. 35–68, 2000.
[28] G. K. Michelle, M. Cartwright, and L. Chen, “Experiences Using Case-Based Reasoning to Predict Software Project Effort,” no. Ml, pp. 1–22, 2000.
[29] A. Idri, A. Abran, and T. M. Khoshgoftaar, “Investigating soft computing in case-based reasoning for software cost estimation,” International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications, 10 (3), 2002. p. 147-157.
[30] A. Idri, A. Zahi, and E. Mendes, A. Zakrani, “Software Cost Estimation Models Using Radial Basis Function Neural Networks”, Mensura-IWSM: Software Process and Product Measurement , 2007, pp 21-31.
[31] S. Yenduri, “an Empirical Study of Imputation Techniques for Software Data Sets,” Louisiana State, 2005.
[32] V. N. Vapnik, The Nature ofStatistical Learning Theory. New York, 1995.
[33] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to the SMO algorithm for SVM regression,” IEEE Trans. Neural Networks, vol. 11, no. 5, pp. 1188–1193, 2000.
[34] J. C. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” Adv. Kernel MethodsSupport Vector Learn., vol. 208, pp. 1–21, 1998.
[35] A. Smola and B. Scholkopf, “A tutorial on support vector regression,” Stat. Comput., vol. 14, no. 3, pp. 199–222, 2004.
[36] X. Chen, Q. Zhou, and H. Xiao, “Combination of Support Vector Regression with Particle Swarm Optimization for Hot-spot temperature prediction of oil-immersed power transformer,” Prz. Elektrotechniczny, no. 8, pp. 172–176, 2012.
[37] A. L. I. Oliveira, “Estimation of software project effort with support vector regression,” Neurocomputing, vol. 69, no. 13–15, pp. 1749–1753, 2006.
[38] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to platt’s SMO algorithm for SVM classifier design,” 1999.
[39] V. N. Vapnik, “An overview of statistical learning theory.,” IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988–99, 1999.
[40] E. Osuna, R. Freund, and F. Girosi, “Training support vector machines: an application to face detection,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.
[41] H. Hsieh, T. Lee, and T.-S. Lee, “A Hybrid Particle Swarm Optimization and Support Vector Regression Model for Financial Time Series Forecasting,” Int. J. Bus. Adm., vol. 2, no. 2, pp. 48–56, 2011.
[42] C. W. Hsu, C. . Chang, and C. J. A. Lin, “A practical guide to support vector classification.,” 2003.
[43] Q. Zong, W. Liu, and L. Dou, “Parameters selection for SVR based on PSO,” in 6th World Congress on Intelligent Control and Automation, 2006, no. 1, pp. 2811–2814.
[44] E. Kocaguneli and T. Menzies, “Software effort models should be assessed via leave-one-out validation,” J. Syst. Softw., vol. 86, no. 7, pp. 1879–1890, 2013.
[45] M. Shepperd and S. MacDonell, “Evaluating prediction systems in software project estimation,” Inf. Softw. Technol., vol. 54, no. 8, pp. 820–827, 2012.
[46] J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.
[47] G. M. Foody, “Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority,” Remote Sens. Environ., vol. 113, no 8, pp. 1658–1663, 2009.
[48] S. Greenland et al., “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations,” Eur. J. Epidemiol., vol. 31, no. 4, pp. 337–350, 2016.
[49] C. J. Geyer, “Nonparametric Tests and Confidence Intervals,” In Pract., pp. 1–14, 2003.
[50] G. Cumming and S. Finch, “Inference by eye: Confidence intervals and how to read pictures of data.,” Am. Psychol., vol. 60, no. 2, pp. 170–180, 2005.
[51] M. J. Gardiner and D. G. Altman, Statistics with confidence: confidence intervals and statistical guidelines. 1989.
[52] D. Sheskin, Handbook of Parametric and Non-parametric Procedures. CRC Press, 1997.
[53] E. Lehmann, “Nonparametrics: Statistical methods based on ranks,” Prentice Hall New Jersey, 1998.
[54] J. L. Hodges and E. L. Lehmann, “Estimates of Location Based on Rank Tests,” Ann. Math. Stat., 1963.
[55] H. Abdi, “1 Overview 2 Preliminary : The different meanings of alpha,” Encycl. Res. Des., pp. 1–8, 2010.
[56] J. Cohen, “Quantitative Methods in Psychology,” Psychol. Bull., vol. 112, no. 1, pp. 155–159, 1992.
[57] M. Hosni, A. Idri, A. Abran, and A. B. Nassif, “On the value of parameter tuning in heterogeneous ensembles effort estimation,” Soft Computing, Springer Berlin Heidelberg, pp. 1–34, 2017.
[58] A. Idri, M. Hosni, and A. Abran, “Improved Estimation of Software Development Effort Using Classical and Fuzzy Analogy Ensembles,” Appl. Soft Comput., vol. 49, pp. 990–1019, 2016.
[59] M. Azzeh, A. B. Nassif, and L. L. Minku, “An empirical evaluation of ensemble adjustment methods for analogy-based effort estimation,” J. Syst. Softw., vol. 103, pp. 36–52, 2015.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1025–1034
DOI: 10.15439/2018F237
ISSN 2300-5963 ACSIS, Vol. 15

A Cost Model for Hybrid Storage Systems in a Cloud Federations

Amina Chikhaoui, Kamel Boukhalfa
University of Science and Technology Houari Boumediene, Algiers, Algeria
{achikhaoui, k.boukhalfa}@usthb.dz

Jalil Boukhobza
Univ. Bretagne Occidentale
UMR 6285, Lab-STICC F-29200, Brest, France
boukhobza@univ-brest.fr

***Abstract*—A cloud federation gives to cloud service providers (CSP) the opportunity to collaborate in order to offer a better QoS to customers at a lower cost. To do so, CSPs make some spare resources available to others at a reduced cost. One of the most critical resources is the storage system as it represents the main system bottleneck. From this point of view, how to efficiently place data in a federation of Clouds with heterogeneous storage systems is a real challenge. To address this issue, one needs to accurately estimate the data placement cost. In this paper, we propose a cost model for hybrid storage systems in a cloud federation for a Database as a Service (DBaaS) application. It takes into account the storage system characteristics, customers I/O workloads and SLA. The proposed cost model considers both 1) Internal customers data placement cost including local placement, outsourcing, back-migration and penalty costs, and 2) External customers data placement cost including insourcing and geo-migration costs. It can be used to help in the decision-making process which aims to enhance customers QoS and reduce CSPs costs in a federation. Simulation results showed the relevance of the considered costs. We have shown that mis-considering some sub-costs may lead to a 95% cost error for external customers data placement and 80% for outsourcing customers. This may cause significant financial loss.**

***Index Terms*—Cloud, federation, hybrid storage, cost model.**

I. INTRODUCTION

CLOUD federation [1], [2], [3], [4] is a computing paradigm that consists in making several Cloud Service Providers (CSPs) cooperate by sharing resources. These CSPs insource and outsource their customers' data and services to provide continuous provisioning by exploiting temporal and spatial availability of resources [2] while reducing their cost. Cloud federation need was driven by novel applications such as mobile cloud, IoT, and big data. It came to address many limitations such as resource contention [5], [6], service interruption [7] and Quality of Service (QoS) degradation that may be due to the geographical distance to cloud resources.

DataBase as a Service (DBaaS) is one of the most important application processing Cloud service offered to cut the IT costs. For such a service, I/O performance and network latency are the two main metrics considered by the customers. Indeed, I/O system is one of the main system bottlenecks [8]. Moreover, it takes about 90% of the transaction execution time in some database queries [9].

Some cloud companies [10] already include latency guarantees in their Service Level Agreements (SLAs) and customers may pay extra charges for reducing I/O latencies.

In order to handle I/O bottlenecks, CSPs rely on the heterogeneity of storage devices. While Hard Disk Drives (HDDs) are used mainly to provide large storage volumes thanks to their low cost per GB, flash based Solid State Drives (SSDs) are integrated to reduce access latency and increase the I/O bandwidth [27], [28]. However, their higher cost does not allow for a massive use [29]. So, according to customers' QoS requirements, a CSP may use this heterogeneity to migrate or replicate their data between different storage classes or to other partner CSPs in the case of a Cloud federation. Dealing with such local and external storage system heterogeneity makes data placement strategies very challenging and needs an accurate cost estimation in order to make adequate decisions when placing data objects.

The cost of operating a Cloud is too significant to be ignored [30], as a matter of fact, using an accurate cost model is a critical issue. In effect, cost models are frequently used for optimization sake. Several state-of-the-art studies dealt with cost estimation of storage systems. We classified them in two categories. The first category is related to studies on centralized Clouds such as [12], [13], [16], [15], [11], [31]. They investigated several issues related to I/O efficiency in case of heterogeneous storage systems, see Table Ia. The second category is related to the cost estimation in the case of interconnected Clouds. These cost models dealt with different operations such as: initial placement of VMs [6], [17], [25], storage cost and geo-migration [21], [18], [32]. However, none of these studies took into account hybrid storage systems and I/O related cost, see Table Ib. They mainly used a fixed storage cost related to data volume without considering I/O operations cost (IOPS or latency).

In the case of a Cloud federation, for an accurate storage cost estimation, CSPs must consider both local and external storage costs. Indeed, it might be more cost effective to store a customer's data on a distant HDD than on a local SSD in case of customer mobility for instance. To consider this issue, the cost estimation must take into account both the details of the storage system and the properties of the Cloud federation. To the best of our knowledge these issues were not considered simultaneously in the existing work.

In this paper, we propose a model to evaluate the storage cost of object placement for DBaaS applications in a federated Cloud. We define an object as any logical entity of a database such as a table, a view, or an index. Our model consists of:

TABLE I: Related work classification

work	Hybrid storage system based cost models				
	Occupation	Energy	Wearout	Migration	Penalty
[11]			x		
[12]			x		
[13], [14]	x	x			
[15]		x	x	x	x
[16]	x	x	x	x	x

(a)

work	Interconnected cloud cost models				
	Fixed storage	Traffic	Geo-migration	Heterogeneity	Federation
[17]	x				x
[18], [19], [20]	x	x			
[21]	x	x	x	x	
[22], [23], [24]			x		
[6]				x	x
[25], [26], [24]		x		x	x

(b)

- Internal customers object placement cost: It takes into account (1) the cost of object placement for internal customers in the local infrastructure, (2) the cost of outsourcing local customers objects, that is buying resources from the other federation CSPs, (3) the cost of prospective back-migration that is the cost induced by bringing back previously outsourced internal customers objects and finally (4) the penalty cost related to the violation of customers SLAs.
- External customers placement cost: this is the cost of managing external customer objects. It consists of (1) the cost of insourcing external customers objects and (2) the cost of sending back these objects to their home or other clouds of the federation when requested.

We have evaluated the cost model and showed that the considered sub-costs are all relevant. With a simple use case, the lowest considered sub-cost part was 6%. Our cost model makes it possible to investigate whether outsourcing is relevant in a federation according to the resource cost and workload properties. It also provides the CSP with a mean to evaluate the penalty cost related to outsourcing some objects and to tune the resources to provide for the federation.

The paper is organized as follows. In Section 2, we discuss some related work, then, we define the system model and formulate the problem in Section 3. In Section 4, our cost model is presented and it is evaluated in Section 5. Finally, we conclude and give some perspectives in Section 6.

II. Related work

In this section we discuss existing studies about hybrid storage systems and interconnected clouds cost models.

A. Hybrid storage system cost models for the Cloud

Many efforts have been made to estimate the cost of data placement in hybrid storage systems. Existing work focused mainly on single clouds. In [13], [14] the storage system cost was estimated using occupation and energy costs. Authors of [11] considered only the wear out while migration cost has been modeled in [12]. Recently, in [15], a cost model has been proposed taking into account four factors: energy, wear out limit, migration between storage classes and penalties. The authors in [16] have extended the previous study with the storage occupation cost. This is summarized in Table Ia.

B. Interconnected cloud cost models

Many existing studies [18], [19], [20] have dealt with the cost of running big data applications, such as social networks, and scientific computing [33]. These studies consider storage, computation and bandwidth to evaluate the cost of data geo-migration and storage in distributed data-centers. All these cost models have used a fixed storage cost mainly related to the storage occupation. The running I/O workload has not been considered. Other studies investigated the cost of distributed clouds (eg: Amazon, Google, etc.). Cost models in this category addressed different aspects: Data center construction [34], energy cost [22], [23], [24], and bandwidth cost [24]. Finally, various cost models for federated clouds have been proposed. Toosi et al [6] designed a cost model which sets dynamically the cost of federated VMs according to the amount of idle resources of the CSPs, as well as a revenue model to maximize the CSP profit. In [17] a cost model, including insourcing and outsourcing costs of virtual resources was proposed. In [25], [26], the cost model consists in local, insourcing, outsourcing and network traffic costs.

Existing hybrid storage cost models are not applicable when it comes to interconnected clouds because data placement in this case implies other factors related to the distributed environment. Regarding federated clouds, the cost of I/O system has not been considered in detail. Indeed this cost has been generally considered as constant and was not related to application workloads.

In this paper, we propose a cost model for the storage system in a federated Cloud taking into account in one hand, the detailed cost of the storage system and in the other hand, the distributed nature and federation properties.

III. System model and Problem definition

In this section, we first describe the system model and then we formulate the cost modeling problem.

A. System model

We consider a federation F composed of D CSPs F= $\{CP^d, d \in [1..D]\}$ cooperating in a peer-to-peer intercloud fashion as depicted in Figure 1. This architecture is inspired from [35] and was used in several studies [6], [25], [17]. The figure emphasizes on the d cloud provider view.

Each CP^d has one datacenter and accommodates different types of storage classes $SC = \{sc_j, j \in [1..J]\}$. We limit our study to two storage classes in this work: SSDs and HDDs. Each sc_j has a limited capacity c_{sc_j}, a maximum IOPS per operation $iops_{op,sc_j}$, a wear out wo_{sc_j}, and a purchase cost p_{sc_j}. We assume that a CSP buys a bandwidth bw from an Internet provider with a purchase cost p_{bw}.

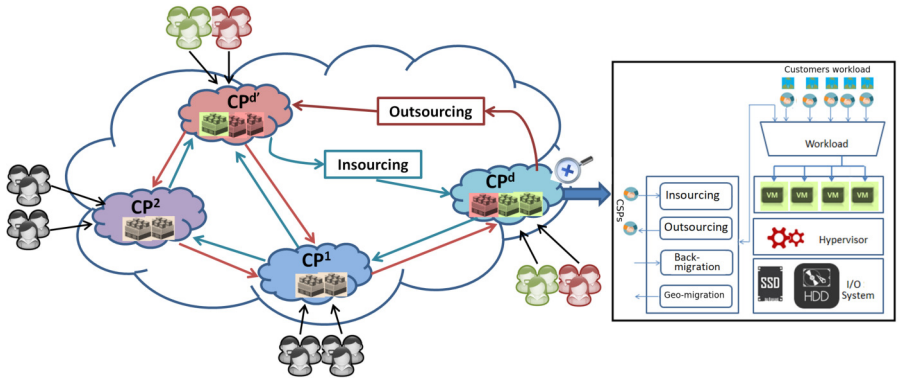


Fig. 1: Federated cloud based hybrid storage system

CSPs may outsource their customers objects to other federation members (called partners), for instance, when they cannot meet customers QoS requirements.

Therefore, each CP^d has two types of customers:

1) internal customers: that are the customers of $CP^d, U_{int} = \{u_k, k \in [1, K]\}$,
2) external customers: that are partner CSPs $\{CP^{d'}\}$ customers whose objects are outsourced to $CP^d, U_{ext} = \{u^{d'}_{k'}, k' \in [1, K']\}, CP^{d'} \in F - \{CP^d\}$. These customers' objects are insourced by CP^d.

We assume that each internal customer u_k:

1) has a set of objects $\{o_{i,k}, i \in [1..N]\}$. Each object has a size denoted $s_{o_{i,k}}$.
2) generates an I/O workload wl_k representing the I/O operations generated by the set of queries issued by the customer. According to the access pattern (sequential, random) and the operation type (read, write), We distinguish four I/O patterns as in [16]: sequential read (sr), sequential write (sw), random read (rr) and finally, random write (rw).
3) a requested QoS (SLA) in terms of IOPS and latency $iops_{sla,k}, ltc_{sla,k}$.
4) A penalty function pn_k composed of two parts, one related to storage performance : $pn_{iops,k}$ and the other to the latency: $pn_{ltc,k}$.

The same notation is used for the external customers.

Based on customers' SLA requirements and the amount of available storage resources, objects are served locally (placed and migrated between different storage classes), insourced, geo-migrated or geo-replicated to partner clouds. To clarify the system model, we introduce the following definitions:

Outsourcing [6], [36], [25], [3]: the ability that CSPs have to send some internal customers objects to other federation members. Outsourcing maybe achieved through either geo-migration or geo-replication.

Geo-migration [21], [18]: the process of moving objects to other clouds without caring about local copy synchronization.

Geo-replication [21], [37]: the process of maintaining multiple copies of objects on multiple sites (CSPs) for a better performance, availability, and reliability.

Insourcing [25], [26], [36], [3]: the opposite process of outsourcing. CSPs make available part of their unused resources to respond to requests from other members.

Inner migration [16], [15]: is the movement of some objects between different storage classes within the same infrastructure. A CSP CP^d migrates periodically some objects between the different storage classes. $O_{m,int}$, $O_{m,ext}$: represent the set of internal (external, respectively) customers objects to migrate between internal storage classes.

Back migration: This operation consists in bringing back previously outsourced objects to the local infrastructure.

The primary objective of a CSP is to reduce the used resources cost while meeting customers QoS needs. The cloud administrator has to take decisions about: (1) which objects need to be moved locally between different storage classes, (2) which ones need to be outsourced and to which partner CSP, i.e. which ones to be replicated or migrated, (3) which other CSP customer objects need to be insourced, (4) and finally, which previously outsourced internal customers objects, need to be brought back to the local infrastructure.

Periodically, the cloud administrator makes decisions about object placement. We note by T the time period during which monitoring is executed to extract objects I/O patterns and to compute the cost of outsourced objects in order to evaluate the overall placement cost as in [16], [15]. We assume that each cloud maintains two matrices, one for internal customers objects (A) and the other for external customers objects (B) such as $A[i, j](T) = 1$ when an internal object is placed in the cloud CP^j and $B[i, j](T) = 1$ when the object is an insourced object to cloud CP^j.

In our study, we assume that a given CSP charges its partners for insourcing actions. Reduced prices are used in order to foster cooperation within the federation. As in [6], each CP^d dynamically adjusts the price of its contributed storage resources according to the amount of idle resources. Let $Cap_{max_{rsc}}$ and $Cap_{idl_{rsc}}$ be the total and idle capacities for a given resource rsc of the provider CP^d. rsc maybe the storage occupation (occ) or performance in IOPS $(iops)$. If p_{rsc} is the price paid by internal customers for the resource rsc, its insourcing price F^d_{res} is obtained from the expression

in eq. (1) from [6]. For each time period T, CSPs will use eq.1 to adjust their insourcing prices.

$$F^d_{rsc} = \frac{Cap_{max_{rsc}} - Cap_{idl_{rsc}}}{Cap_{max_{rsc}}} * (p_{rsc}) \quad (1)$$

From the network resource point of view, only outgoing network (Internet) traffic is charged for customers in a given CSP, be it a local or an outsourced customer (in this case, the hosting CSP charges the external network traffic cost for outsourcing one).

In this paper, we assume that each CP^d dedicates a fixed part of storage and bandwidth resources for its internal customers $(sc_{j,int}, bw_{int})$ and another part for the external customers $sc_{j,ext}, bw_{ext}$ as noted in eq. 2.

$$\forall CP^d, \begin{cases} sc_j = sc_{j,int} + sc_{j,ext} \\ bw = bw_{int} + bw_{ext} \end{cases} \quad (2)$$

B. Problem formulation

In the following, we define our problem by giving system inputs and outputs.

Input: A CSP of the federation that has : (1) a set of storage classes, (2) an Internet bandwidth with maximum capacity and purchase cost, (3) a storage and outgoing network bandwidth costs for partner Clouds. (4) a set of internal customers, (5) a set of external customers, (6) two sets of objects $O_{m,int}$ and $O_{m,ext}$ to move between the local storage classes. (7) a monitoring period T, and (8) two matrices: $A(T)$ and $B(T)$.

Output: The monetary placement cost of all customers' objects for a given time period T.

IV. Storage cost model in a Cloud federation

A. Overview

We model the cost of object placement for a cloud CP^d, $d \in [1..D]$, belonging to a federation F for a given period of time T. The model is built hierarchically, see Figure 2.

The total object placement cost $Cost_{plc,T}$ is the sum of the placement cost of the internal customers ($Cost_{plc_{int},T}$) and the external customers ($Cost_{plc_{ext},T}$) as shown in eq. 3.

Note that non-recurring costs which do not depend on the objects placement like maintenance cost, human resources cost, air-conditioning costs are not considered in this paper. We do not also consider the cost related to data security.

$$Cost_{plc,T} = Cost_{plc_{int},T} + Cost_{plc_{ext},T} \quad (3)$$

B. Internal customers object placement cost (see (2), Figure 2)

This is the cost of placing internal customers objects. It includes the local placement cost $Cost_{lcl_{int},T}$, the outsourcing cost ($Cost_{out_{src},T}$), the back-migration cost ($Cost_{bck_{mgr},T}$) and the penalty cost $Cost_{pnt_{int},T}$ as shown in eq. 4.

The local placement cost is the storage cost of internal customers objects in the local infrastructure. The outsourcing cost is related to the placement of internal customers objects in partner CSPs. The back-migration cost represents the cost of bringing back the previously outsourced objects to the home infrastructure. Finally, the penalty cost represents the additional monetary cost caused by SLA violations.

$$Cost_{plc_{int},T} = Cost_{lcl_{int},T} + Cost_{out_{src},T} + Cost_{bck_{mgr},T} + Cost_{pnt_{int},T} \quad (4)$$

1) Local placement cost (see (4), Figure 2): It is obtained from the storage cost $Cost_{stg_{int},T}$ of internal customers objects and the inner migration cost (between storage classes within the CSP infrastructure) $Cost_{mgr_{int},T}$ of the set of objects $O_{m,int}$ as shown in eq. 5.

$$Cost_{lcl_{int},T} = Cost_{stg_{int},T} + Cost_{mgr_{int},T} \quad (5)$$

The storage cost: As in [16], $Cost_{stg_{int},T}$ is related to the occupation $Cost_{occ,T}$, the energy $Cost_{erg,T}$ and the wear out cost $Cost_{edr,T}$ due to I/O workload (see eq. 6).

$$Cost_{stg_{int},T} = max[Cost_{occ,T}, Cost_{edr,T}] + Cost_{erg,T} \quad (6)$$

The occupation cost is the amortized cost of the storage system over the period T. The energy cost is the energy consumed by the storage system to execute the I/O workload multiplied by the energy unitary price which we consider constant. Finally, the wear out cost is caused by the execution of the I/O workload which impacts the lifetime of the devices. For SSD, this cost is related to the amount of written data while for HDD, it depends of the amount of both read and written data. For more details see [16].

The migration cost: it consists in reading the set of objects $O_{m_{int}}$ from the source storage device and writing them to another destination storage device. It includes the energy and endurance costs.

2) Outsourcing cost (see (5), Figure 2): It is composed of geo-migrating and geo-replicating internal customers objects and the storage cost in the partner CSPs generated by these two geo-operations as in eq. 7.

$$Cost_{out_{src},T} = Cost_{geo_{mgr},T} + Cost_{geo_{rpl},T} + Cost_{ext_{plc},T} \quad (7)$$

The geo-migration cost (see (11), Figure 2): geo-migration of objects incurs local read operations $Cost_{rd,T}$ and internet bandwidth usage $Cost_{bw,T}$.

The set of objects concerned by the geo-migration $O_{g_{mgr}}$ is obtained from the difference between the matrix A at the period $T-1$ and the current period T where $o_i, A[i,d](T-1) = 1 \wedge A[i,d](T) = 0$. Geo-migration cost is shown in eq. 8.

$$Cost_{geo_{mgr},T} = Cost_{rd,T} + Cost_{bw,T} \quad (8)$$

The local read operation cost incurs energy $Cost_{rd_{erg},T}$ and wear out $Cost_{rd_{edr},T}$ costs.

$$Cost_{rd,T} = Cost_{rd_{erg},T} + Cost_{rd_{edr},T} \quad (9)$$

The Internet bandwidth cost is related to the bandwidth consumed by the geo-migration of objects to external clouds.

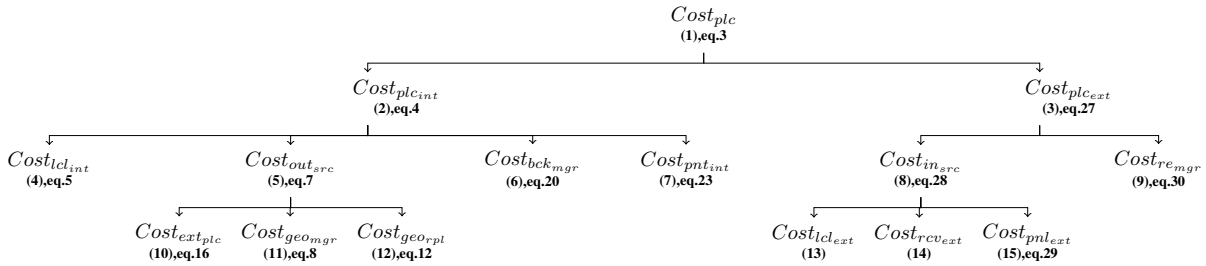


Fig. 2: The overall cost model

We calculate this cost by multiplying the size of all objects to be migrated by the amortized bandwidth over one unit of time $Cost_{bw_{amz},1}(bw_{int})$.

$$Cost_{bw,T} = \frac{\sum_{o_i \in O_{gmgr}} (S_{o_i}) * Cost_{bw_{amz},1}(bw_{int})}{bw_{int}} \quad (10)$$

We calculate the amortized bandwidth cost over one unit of time by distributing the purchase cost of bw_{int} over the subscription time period T_{sp}.

$$Cost_{bw_{amz},1}(bw_{int}) = \frac{p_{bw} * \frac{bw_{int}}{bw}}{T_{sp}} \quad (11)$$

The geo-replication cost (see (12), Figure 2): geo-replicating objects over multiple CSPs consumes supplementary storage and network costs. The geo-replication consists in adding, deleting and synchronizing replicas. So, its overall cost is the sum of the aforesaid operations costs as shown in eq. 12. Deleting replica cost is assumed to be nil.

$$Cost_{geo_{rpl},T} = Cost_{add_{rpl},T} + Cost_{syc_{rpl},T} \quad (12)$$

The cost of adding replicas is the same as the geo-migration cost without deleting the original copy and considering the set of objects $\{O^{d'}_{g_{rpl}}, d' \neq d\}$. They are obtained from matrix A.

$$Cost_{add_{rep},T} = \sum_{CP^{d'} \in F - \{CP^d\}} Cost_{geo_{mgr},T}(O^{d'}_{g_{rpl}}) \quad (13)$$

Concerning replica synchronization either CP^d forwards the synchronization or receives it. $O^{d'}_{s_{rpl}}$ is the list of objects sent to, or received from $CP^{d'}$ for synchronizing replicas. When CP^d forwards the synchronization, its cost is the sum of the read operations induced and the consumed Internet bandwidth costs, see eq. 14.

$$Cost_{syn_{rep},T} = \sum_{CP^{d'} \in F - \{CP^d\}} Cost_{geo_{mgr},T}(O^{d'}_{s_{rpl}}) \quad (14)$$

Whereas when CP^d receives the synchronization, this cost consists of writing operations, local Internet bandwidth use, and outgoing network costs (i.e back-migration cost, see later in eq. 20) as shown in the next equation 15.

$$Cost_{syn_{rep},T} = \sum_{CP^{d'} \in F - \{CP^d\}} Cost_{bck_{mgr},T}(O^{d'}_{s_{rpl}}) \quad (15)$$

The external placement cost (see (10), Figure 2): Once the objects have been migrated to other clouds, the CSP needs to pay partner clouds for hosting the objects. Therefore, the external storage cost includes the external occupation cost $Cost_{ext_{occ},T}$, the external workload execution cost $Cost_{ext_{wld},T}$ and the external penalty cost $Cost_{ext_{pnl},T}$. The penalty cost here is the charge paid by the partner clouds in case SLAs are violated for the outsourced objects.

The purchased resources are billed according to eq. 1.

$$Cost_{ext_{plc},T} = Cost_{ext_{occ},T} + Cost_{ext_{wld},T} - Cost_{ext_{pnl},T} \quad (16)$$

The external occupation cost: represents the cost of storing the migrated or replicated internal customers objects on partner clouds. We calculate this cost by using the federated occupation cost $F^{d'}_{occ}$ of $CP^{d'}$ as noted in eq. 17

$$Cost_{ext_{occ},T} = \sum_{d' | CP^{d'} \in F - \{CP^d\}} (\sum_i A[i,d'] * S(o_i)) * F^{d'}_{occ} \quad (17)$$

The cost of workload execution represents the amount of expected IOPS consumed by internal users. It is calculated the same way as in the eq. 18 where $F^{d'}_{iops}$ is the cost of one IOPS in the cloud $CP^{d'}$

$$Cost_{ext_{wld},T} = \sum_{d' | CP^{d'} \in F - \{CP^d\}} (\sum_i A[i,d'] * IOPS(o_i)) * F^{d'}_{iops} \quad (18)$$

The penalty cost is a percentage $\alpha\%$ of the total charges paid by C^p to the partner cloud, as in Amazon Cloud.

$$Cost_{ext_{pnl},T} = \alpha * (Cost_{ext_{occ},T} + Cost_{ext_{wld},T}) \quad (19)$$

3) Back migration cost (see (6), Figure 2): $Cost_{bck_{mgr},T}$ is the cost of bringing back the previously outsourced objects to the local infrastructure. It is composed of the cost of data transfer (outgoing network cost) from the partner clouds $Cost_{ntw_{out},T}$, the local Internet bandwidth cost $Cost_{bw,T}$

corresponding to the bandwidth consumed locally to receive the previously outsourced objects, and the cost $Cost_{wr,T}$ to write the received objects on the local storage, see eq. (20).

$$Cost_{bck_{mgr},T} = Cost_{wr,T} + Cost_{ntw_{out},T} + Cost_{bw,T} \quad (20)$$

The set of internal customers objects to be brought back $O_{b_{mgr}}$ is deduced from the matrix A.

The local write cost: is the cost of placing previously outsourced objects in the local storage system. This cost involves the cost of energy $Cost_{wr_{erg},T}$ consumed by the storage devices and the wear out cost of these devices $Cost_{wr_{edr},T}$ caused by the write operations. Some of these objects are placed on HDD and others on SSD, see eq. 21.

$$Cost_{wr,T} = Cost_{wr_{erg},T} + Cost_{wr_{edr},T} \quad (21)$$

Outgoing network cost: This cost is charged by partner CSPs when CP^d brings back their internal customer objects to the local infrastructure. We calculate this cost by multiplying the size of the objects to bring back by the cost of outgoing network $Cost_{out}(CP^{d'})$ of the partner.

$$Cost_{ntw_{out},T} = \sum_{CP^{d'} \in F - \{CP^d\}} (Cost_{out}(CP^{d'}) * (\sum_{o_i \in O_{b_{mgr}}} S_{o_i})) \quad (22)$$

Local Internet bandwidth cost is calculated with eq. 10.

4) Penalty cost (see (7), Figure 2): The violation of SLA terms by the CSP entails a penalty that should be paid. As in [16], we calculate the overall penalty as the sum of all internal customers penalties.

$$Cost_{pnt_{int},T} = \sum_{u_k \in U_{int}} (Cost_{pnt,T}(u_k)) \quad (23)$$

Internal customer penalty cost is composed of a penalty related to the IOPS $Cost_{pnt_{iops},T}(u_k)$ and another one related to the latency $Cost_{pnt_{ltc},T}(u_k)$.

$$Cost_{pnt,T}(u_k) = Cost_{pnt_{iops},T}(u_k) + Cost_{pnt_{ltc},T}(u_k) \quad (24)$$

The IOPS penalty cost is calculated as in [16], it is proportional to the ratio between the offered IOPS $iops_{offered}(u_k)$ and the requested one ($iops_{sla,k}$). The requested IOPS is defined in the SLA while we calculate the offered IOPS from the time needed to handle the I/O workload of the customer u_k and the total number of I/O requests issued to their objects. It is calculated as follows:

$$Cost_{pnt_{iops},T}(u_k) = pnt_{iops,k}(\frac{iops_{offered}(u_k)}{iops_{sla,k}}) \quad (25)$$

The latency penalty cost: In our model, we assume that customers pay extra charges to have a reduced network latency. If the latency offered $ltc_{offered}(u_k)$ to customer u_k is lower than the one requested $ltc_{sla,k}$, then a penalty is applied to the CSP. This penalty cost is proportional to the violation degree and the number of requests nb_{rqt}, as noted in Table VI. We calculate the latency penalty cost as follows:

$$Cost_{ltc_{iops},T}(u_k) = \sum_{i=1} nb_{rqt}(pn_{ltc,k}(\frac{ltc_{offered}(u_k)}{ltc_{sla,k}})) \quad (26)$$

TABLE II: Latency penalty per request

$ltc_{offered}$	Penalty (%)
$[0..B_0]$	0
$[B_i..B_{i+1}]$	x_i
$> B_n$	x_n

C. External customers placement cost (see (3), Figure 2)

This cost $Cost_{plc_{ext},T}$ is the sum of the cost of insourcing external customers objects $Cost_{in_{src},T}$ and the cost generated when some objects of the external customers are geo-migrated to other clouds or taken back by their home clouds (re-migration cost) re_{mgr}, see eq. 27.

$$Cost_{plc_{ext},T} = Cost_{in_{src},T} + Cost_{re_{mgr},T} \quad (27)$$

1) Insourcing cost (see (8), Figure 2): The insourcing cost is composed of receiving external customers objects cost $Cost_{rcv_{ext},T}$, the local placement cost $Cost_{lcl_{ext},T}$ and the penalty cost $Cost_{pnt_{ext},T}$.

$$Cost_{in_{src},T} = Cost_{rcv_{ext},T} + Cost_{lcl_{ext},T} + Cost_{pnt_{ext},T} \quad (28)$$

Receiving cost (see (14), Figure 2) of the external customers objects is the cost of writing these objects, calculated from eq. 21, and the cost of the consumed local Internet bandwidth, calculated with eq. 10. The set of external customers objects to be received $O_{e_{rcv}}$ is deduced from the matrix B

The local placement cost (see (13), Figure 2): of the external customers objects is calculated from eq. 5.

The penalty cost (see (15), Figure 2) is the sum of penalty costs of the external customers.

$$Cost_{pnt_{ext},T} = \sum_{u_{k'}^{d'} \in U_{ext}} (Cost_{pnt,T}(u_{k'}^{d'})) \quad (29)$$

In the federation, resource prices are dynamically set (see eq. 1), so the penalty cost is set accordingly and is related to the unused storage resources β which is calculated as follows:

$$\beta = \frac{Cap_{max_{rsc}} - Cap_{idl_{rsc}}}{Cap_{max_{rsc}}} \quad (30)$$

With rsc being a combination of the storage space and performance requested. The penalty cost of an external customer $u_{k'}^{d'}$ is thus given in eq. 31:

$$Cost_{pnt,T}(u_{k'}^{d'}) = \beta * Cost_{pnt_{iops},T}(u_{k'}^{d'}) \quad (31)$$

2) The re-migration cost (see (9), Figure 2): This cost is driven by the decision of home clouds to geo-migrate their customers back to their infrastructure or to another cloud. This incurs local read operations and Internet bandwidth costs as described in eq. 8. The set of objects concerned by the re-migration is deduced from the matrix B by the difference between the period $T-1$ and the current period T where for $o_i, B[i,d](T-1) = 1 \wedge B[i,d](T) = 0$.

V. Evaluation

This section presents an evaluation of the proposed cost model. Our aim is twofold: (i) validating the relevance of the sub-costs used in our cost model and comparing to state-of-the-art models, (ii) showing the flexibility of the cost model through the investigation of the impact of different parameters on the placement cost.

A. Experimental Settings

We used CloudSim simulator [38]. The simulated scenario is composed of a federation of 9 geographically distributed CSPs. Each CSP is composed of one datacenter running 1000 VMs as in [6]. Some of these VMs (set to 20% in our experiments) are devoted to the federation. Each CSP has an internet bandwidth of 1Gbps bought from an Internet service provider (the price was set to 1500$ per month). We used a standard VM configuration with 8 cores, 8 GB of RAM and a hybrid storage system. Characteristics of the storage system is provided in Table III.

TABLE III: Storage devices specifications

Characteristics	HDD	SSD($)
Price ($)	230	200
Size	1 TB	128 GB
Performance	Seek time:8.5 ms	rr:10000 IOPS, rw:40000 IOPS
	read/write: 9.5 ms	sr:540 MB/s, sw: 520 MB/s

Each CSP manages a set of databases (DBs) built using TPC-H and TPC-C benchmarks with different sizes and varied workload, we used the configuration given in [16]. For more details about the different DBs and storage system specifications see Table IV. The amount of storage penalty is set to 30 % of the total charges paid by the customer as in Amazon Cloud. The storage price is fixed according to Amazon gp2 price model (0.1$/GB/month) while the energy cost is set to 0.1$ per kWh as in [13].

Some of the customers are mobile and ask for a good latency by paying some extra charge. Their home CSP generally migrates their objects to the nearest CSP that meets the latency constraint otherwise the home CSP may undergo a penalty. Mobile customers are supposed to have a 1 week duration mobility. The price of latency is given in Table V [10] and its related penalty in Table VI. The evaluation is conducted over a 1 month with one-hour time period for dynamic resource price update. The insourcing prices are dynamically set by CSPs using equation (eq. 1) and changed each hour (period T). We assume that the outgoing network cost is set to 50% of amazon's which gives 0.09$/GB.

TABLE IV: Data bases specifications

Data base	Bench.	Reqs. nbr (op/h)	rr(op/h)	rw(op/h)	sr(op/h)	sw(op/h)
DB1(32GB/HDD)	TPC-C	432000	136800	46800	108000	32400
DB2(60GB/SSD)		28800000	601200	104400	5436000	147600
DB3(147GB/HDD)	TPC-H	331200	10800	216000	3600	18000
DB4(34GB/HDD)		355200	32400	216000	7200	10800
DB5(381GB/HDD)		15360	1080	8280	360	1080

TABLE V: Latency price/req.

Latency (ms)	Price by request ($)
< 200	0.0000001
< 400	0.00000005
< 600	0.00000002
> 600	0

TABLE VI: Penalty/req.

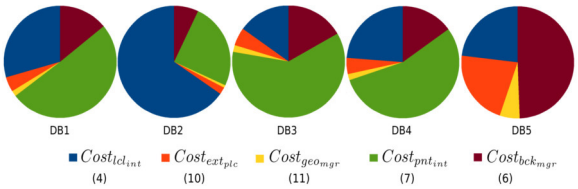
Latency (ms)	Penalty ($)
< 200	0
< 400	0.00000005
< 600	0.00000008
> 600	0.0000001

B. Evaluation results

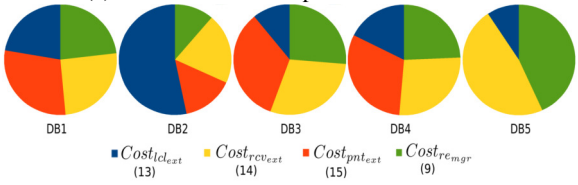
1) Relevance of the sub-costs: The first experiment concerns the relevance of the used sub-costs. In this scenario, the network latency between each pair of cloud infrastructures is assigned randomly between]200ms, 700ms]. First, we evaluated all different sub-costs for one CSP, then, we calculated the average placement cost of (1) insourced databases see (8), Figure 2 and (2) outsourced databases see (5), Figure 2, and compared the resulting costs with those of some state-of-the-art studies [18], [19], [21], [6], [25]. These costs were chosen as they are the higher ones.

Figure 3a and Figure 3b show the different sub-costs of our cost model (leaf nodes of Figure 2). In this evaluation, geo-replication costs were not considered (only geo-migration has been shown). We observe that all the modeled costs are relevant as each cost is high enough for at least one tested database. We observe that the local placement and penalty costs of internal customers are the highest ones. This is due to the fact that only a small part of customers are mobile (20%) and for small periods (1 week). Also, the database size affects directly the external placement cost because the storage in this case is bought from others CSPs. The database size affects likewise all the costs including geo-migration and back-migration, while the remaining costs are affected by the storage device type and workload patterns.

Figure 4a and Figure 4b show the cost of insourced and

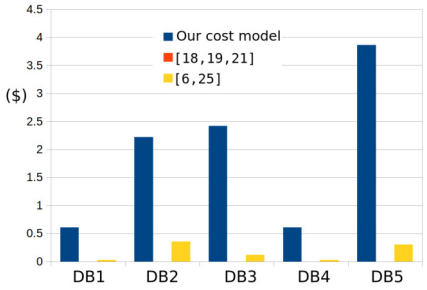


(a) Internal customers placement sub-costs

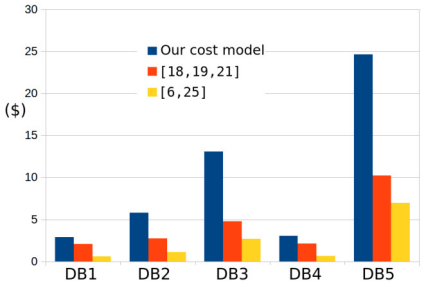


(b) External customers placement sub-costs

Fig. 3: Sub-costs of the simulated placement cost.



(a) Insourced database cost



(b) Outsourced database cost

Fig. 4: The average per week placement cost of the insourced and outsourced databases for a given CSP.

outsourced databases for a given CSP for our cost model as compared to state-of-the-art models. A first observation one may draw is that models from [18], [19], [21] do not consider the insourcing cost. In addition, the cost of storage services for outsourced objects is fixed to 1 GB/unit of time. In [6], [25], only the occupation cost is considered with no geo-migration cost which can lead, in the worse case, to a difference with our cost model of 95% for the external customers placement cost and 80% for the outsourcing cost. Furthermore, the penalty cost is ignored in those studies while in our work it can represent up to 61% of the placement cost of internal customers and 32% of that of external customers.

2) Spare resources pricing model evaluation: The objective of this part is to show how the proposed cost model makes it possible to easily evaluate the spare resource pricing model used according to the ran workload.

Figure 5.a shows, from one hand the cost of running insourced databases and from the other hand, the price billed to the partner clouds. As one can observe, the financial gain of insourcing objects highly depends on the nature of the ran workload. Insourcing large volumes of data generating a low number of queries (DB5) is of course more interesting.

Figure 5.b shows the outsourced objects cost when buying resources from partner clouds as compared to buying them from an external cloud. This Figure illustrates the ability of the model to highlight the gain obtained for a given CSP knowing its outsourced objects and the used federation pricing model. This may help to optimize the outsourcing decision.

3) Impact of latency and penalty, Federation vs DCC vs single Cloud: The designed cost model makes it possible to compare the cost of running a Cloud infrastructure within or out of a Cloud federation. In this part, we compare the average placement cost of mobile customers databases by varying the latency between clouds for three scenarios: (1) Geo-migration using a federation, (2) Geo-migration using a distributed cloud computing platform (DCC) and (3) Single Cloud without geo-migration (see Table 6). The aim of this part is to show the impact of the violation degree and the network latency on the average placement cost for the three configurations.

We notice, from Figure 6, that generally, the cost without outsourcing mobile customers increases with the increase of the network latency. For databases with heavy workloads (e.g. DB2) it is always interesting to outsource objects. In fact, outsourcing is interesting as long as the amount of outsourced data is small and the workload is heavy. This is not the case for data with small workload and large size. This is because geo and back-migrating these data implies a high network traffic cost (e.g. DB5). For some databases with large sizes and medium workloads, it is not cost-effective to outsource them as long as the violation of the latency is not high. However, when the degradation of network latency reaches a certain level, it becomes interesting to be outsourced. For instance, for DB3 the cost without migration (with penalty) is lower than the cost with migration when latency is $< 400ms$ but it is the opposite when the latency becomes $> 600ms$. Our cost model allows to make a trade-off between all these parameters for optimizing the overall cost of running a Cloud into a federation.

VI. Conclusion

A CSP can meet its customers QoS and minimize the cost of data placement by using either local or partner resources in a Federation. In this work, we have proposed a cost model for data placement on hybrid storage systems in a Cloud federation. Our model extends state-of-the-art work by considering geo-migration, penalty, back-migration, and geo-replication costs.

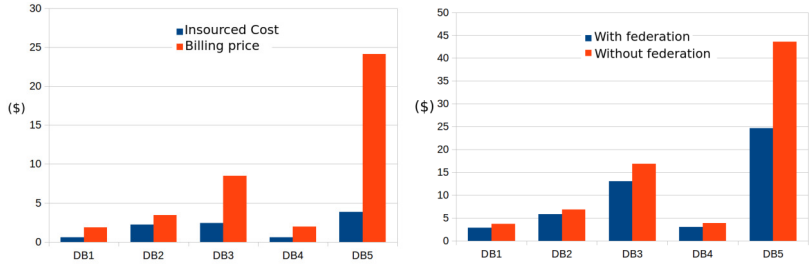
For future work, we will first investigate placement optimization strategies to find the optimal internal and/or external placement for CSP objects. We will also design pricing strategies for resources in a Federation.

The performed evaluations proved the relevance of the considered costs. It also pointed out that outsourcing and insourcing is a complex task that requires taking into account a large number of parameters. Even though our model is storage oriented, it can be used and integrated in a broader cost model considering other resources (e.g. CPU and memory).

References

[1] D. Villegas, N. Bobroff, I. Rodero, J. Delgado, Y. Liu, A. Devarakonda, L. Fong, S. M. Sadjadi, and M. Parashar, "Cloud federation in a layered service model," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1330–1344, 2012, http://dx.doi.org/10.1016/j.jcss.2011.12.017.

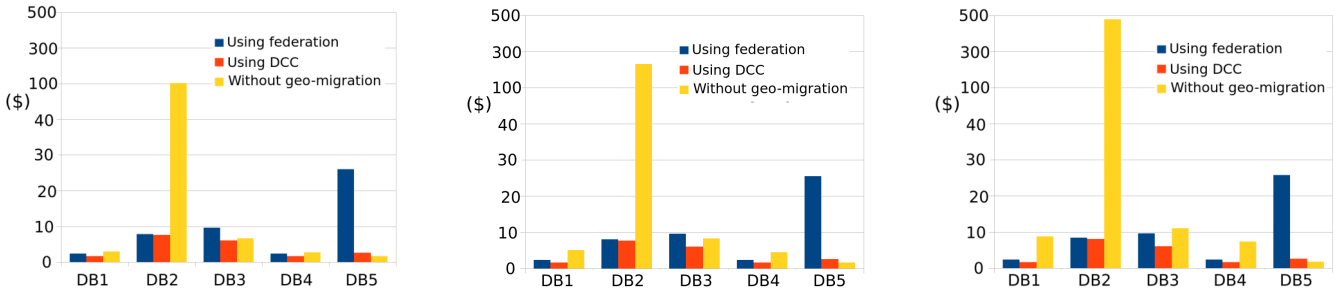
[2] H. Li, C. Wu, Z. Li, and F. C. Lau, "Profit-maximizing virtual machine trading in a federation of selfish clouds," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 25–29, http://dx.doi.org/10.1109/infcom.2013.6566728.



(a) Insourced cost and billed prices

(b) Outsourced cost with and without federation

Fig. 5: Federation impact on the cost



(a) $200 < Latence(ms) < 400$

(b) $400 < Latence(ms) < 600$

(c) $Latence(ms) > 600$

Fig. 6: Average cost with migration in federated cloud and distributed cloud VS average cost without migration

[3] M. R. Assis and L. F. Bittencourt, "A survey on cloud federation architectures: identifying functional and non-functional properties," *Journal of Network and Computer Applications*, vol. 72, pp. 51–71, 2016, http://dx.doi.org/10.1016/j.jnca.2016.06.014.

[4] R. Moreno-Vozmediano, E. Huedo, I. M. Llorente, R. S. Montero, P. Massonet, M. Villari, G. Merlino, A. Celesti, A. Levin, L. Schour *et al.*, "Beacon: a cloud network federation framework," in *Communications in Computer and Information Science*. Springer, 2016, pp. 325–337, http://dx.doi.org/10.1007/978-3-319-33313-7_25.

[5] M. Amiri and L. Mohammad-Khanli, "Survey on prediction models of applications for resources provisioning in cloud," *Journal of Network and Computer Applications*, vol. 82, pp. 93–113, 2017, http://dx.doi.org/10.1016/j.jnca.2017.01.016.

[6] A. N. Toosi, R. N. Calheiros, R. K. Thulasiram, and R. Buyya, "Resource provisioning policies to increase iaas provider's profit in a federated cloud environment," in *High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on*. IEEE, 2011, pp. 279–287, http://dx.doi.org/10.1109/hpcc.2011.44.

[7] Y. Gu, D. Wang, and C. Liu, "Dr-cloud: Multi-cloud based disaster recovery service," *Tsinghua Science and Technology*, vol. 19, no. 1, pp. 13–23, 2014, http://dx.doi.org/10.1109/tst.2014.6733204.

[8] E. Shriver, "Performance modeling for realistic storage devices," 1997, https://dl.acm.org/citation.cfm?id=269078.

[9] M. A. Sharaf, P. K. Chrysanthis, A. Labrinidis, and C. Amza, "Optimizing i/o-intensive transactions in highly interactive applications," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 785–798, http://dx.doi.org/10.1145/1559845.1559927.

[10] D. B. Terry, V. Prabhakaran, R. Kotla, M. Balakrishnan, M. K. Aguilera, and H. Abu-Libdeh, "Consistency-based service level agreements for cloud storage," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013, pp. 309–324, http://dx.doi.org/10.1145/2517349.2522731.

[11] Y. Cheng, M. S. Iqbal, A. Gupta, A. R. Butt, and V. Tech, "Pricing games for hybrid object stores in the cloud: Provider vs. tenant." in *HotStorage*, 2015, https://www.usenix.org/conference/hotcloud15/workshop-program/presentation/cheng.

[12] L. Lin, Y. Zhu, J. Yue, Z. Cai, and B. Segee, "Hot random off-loading: A hybrid storage system with dynamic data migration," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on*. IEEE, 2011, pp. 318–325, http://dx.doi.org/10.1109/mascots.2011.41.

[13] Y. Kim, A. Gupta, B. Urgaonkar, P. Berman, and A. Sivasubramaniam, "Hybridstore: A cost-efficient, high-performance storage system combining ssds and hdds," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on*. IEEE, 2011, pp. 227–236, http://dx.doi.org/10.1109/mascots.2011.64.

[14] N. Zhang, J. Tatemura, J. M. Patel, and H. Hacigümüş, "Towards cost-effective storage provisioning for dbmss," *Proceedings of the VLDB Endowment*, vol. 5, no. 4, pp. 274–285, 2011, http://dx.doi.org/10.14778/2095686.2095687.

[15] H. Ouarnoughi, J. Boukhobza, F. Singhoff, and S. Rubini, "A cost model for virtual machine storage in cloud iaas context," in *Parallel, Distributed, and Network-Based Processing (PDP), 2016 24th Euromicro International Conference on*. IEEE, 2016, pp. 664–671, http://dx.doi.org/10.1109/pdp.2016.119.

[16] D. Boukhelef, J. Boukhobza, and K. Boukhalfa, "A cost model for dbaas storage," in *International Conference on Database and Expert Systems Applications*. Springer, 2016, pp. 223–239, http://dx.doi.org/10.1007/978-3-319-44403-1_14.

[17] M. Hadji and D. Zeghlache, "Mathematical programming approach for revenue maximization in cloud federations," *IEEE transactions on cloud computing*, vol. 5, no. 1, pp. 99–111, 2017, http://dx.doi.org/10.1109/tcc.2015.2402674.

[18] Z. Wen, J. Cała, P. Watson, and A. Romanovsky, "Cost effective, reliable and secure workflow deployment over federated clouds," *IEEE Transactions on Services Computing*, vol. 10, no. 6, pp. 929–941, 2017, http://dx.doi.org/10.1109/tsc.2016.2543719.

[19] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. Lau, "Moving big data to the cloud: An online cost-minimizing approach," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 2710–2721, 2013, http://dx.doi.org/10.1109/jsac.2013.131211.

[20] W. Xiao, W. Bao, X. Zhu, and L. Liu, "Cost-aware big data processing across geo-distributed datacenters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 11, pp. 3114–3127, 2017, http://dx.doi.org/10.1109/tpds.2017.2708120.

[21] Y. Mansouri, A. N. Toosi, and R. Buyya, "Cost optimization for

dynamic replication and migration of data in cloud data centers," *IEEE Transactions on Cloud Computing*, 2017, http://dx.doi.org/10.1109/tcc.2017.2659728.

[22] A. Khosravi, L. L. Andrew, and R. Buyya, "Dynamic vm placement method for minimizing energy and carbon cost in geographically distributed cloud data centers," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 183–196, 2017, http://dx.doi.org/10.1109/tsusc.2017.2709980.

[23] K. Le, R. Bianchini, J. Zhang, Y. Jaluria, J. Meng, and T. D. Nguyen, "Reducing electricity cost through virtual machine placement in high performance computing clouds," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 22, http://dx.doi.org/10.1145/2063384.2063413.

[24] H. Yuan, J. Bi, W. Tan, and B. H. Li, "Cawsac: Cost-aware workload scheduling and admission control for distributed cloud data centers," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 976–985, 2016, http://dx.doi.org/10.1109/tase.2015.2427234.

[25] S. Rebai, M. Hadji, and D. Zeghlache, "Improving profit through cloud federation," in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*. IEEE, 2015, pp. 732–739, http://dx.doi.org/10.1109/ccnc.2015.7158069.

[26] M. Hadji, B. Aupetit, and D. Zeghlache, "Cost-efficient algorithms for critical resource allocation in cloud federations," in *Cloud Networking (Cloudnet), 2016 5th IEEE International Conference on*. IEEE, 2016, pp. 1–6, http://dx.doi.org/10.1109/cloudnet.2016.11.

[27] J. Boukhobza, "Flashing in the cloud: Shedding some light on nand flash memory storage systems," in *Data Intensive Storage Services for Cloud Environments*. IGI Global, 2013, pp. 241–266, http://dx.doi.org/10.4018/978-1-4666-3934-8.ch015.

[28] J. Boukhobza and P. Olivier, *Flash Memory Integration: Performance and Energy Issues*. Elsevier, 2017, https://www.elsevier.com/books/flash-memory-integration/boukhobza/978-1-78548-124-6.

[29] D. Lee, C. Min, and Y. I. Eom, "Effective flash-based ssd caching for high performance home cloud server," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 2, pp. 215–221, 2015, http://dx.doi.org/10.1109/tce.2015.7150596.

[30] C. Wu and R. Buyya, *Cloud Data Centers and Cost Modeling: A complete guide to planning, designing and building a cloud data center*. Morgan Kaufmann, 2015, https://www.elsevier.com/books/cloud-data-centers-and-cost-modeling/wu/978-0-12-801413-4.

[31] J. Tai, B. Sheng, Y. Yao, and N. Mi, "Live data migration for reducing sla violations in multi-tiered storage systems," in *Cloud Engineering (IC2E), 2014 IEEE International Conference on*. IEEE, 2014, pp. 361–366, http://dx.doi.org/10.1109/ic2e.2014.8.

[32] W. Xiao, X. Lei, R. Li, N. Park, and D. J. Lilja, "Pass: a hybrid storage system for performance-synchronization tradeoffs using ssds," in *Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on*. IEEE, 2012, pp. 403–410, http://dx.doi.org/10.1109/ispa.2012.59.

[33] Square kilometre array. https://www.skatelescope.org/.

[34] A. Simonet, A. Lebre, and A.-C. Orgerie, "Deploying distributed cloud infrastructures: Who and at what cost?" in *Cloud Engineering Workshop (IC2EW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 178–183, http://dx.doi.org/10.1109/ic2ew.2016.48.

[35] N. Grozev and R. Buyya, "Inter-cloud architectures and application brokering: taxonomy and survey," *Software: Practice and Experience*, vol. 44, no. 3, pp. 369–390, 2014, http://dx.doi.org/10.1002/spe.2168.

[36] A. N. Toosi, R. N. Calheiros, and R. Buyya, "Interconnected cloud computing environments: Challenges, taxonomy, and survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 7, 2014, http://dx.doi.org/10.1145/2593512.

[37] N. K. Gill and S. Singh, "A dynamic, cost-aware, optimized data replication strategy for heterogeneous cloud data centers," *Future Generation Computer Systems*, vol. 65, pp. 10–32, 2016, http://dx.doi.org/10.1016/j.future.2016.05.016.

[38] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and experience*, vol. 41, no. 1, pp. 23–50, 2011, http://dx.doi.org/10.1002/spe.995.

APPENDIX A: NOTATION TABLE

Symbol	Meaning
Federation	
F	Federation of a set of CPs
CP^d	cloud provider d
$Cap_{max_{res}}$, $Cap_{idl_{res}}$,p_{rsc}	The total and the idle capacities of storage resource res and its price.
F^d_{res}	The insourcing price of storage resource res of CP^d
Customers	
U_{int}, u_k	The set of internal customers and kth customer
U_{ext}, $u^{d'}_{k'}$	The set of external customers and k'th external customer of CP^d(internal customer of the cloud $CP^{d'}$)
wl_k	the workload of u_k
$iops_{sla,k}$, $ltc_{sla,k}$	The requested performance in terms of IOPS and latency
pn_k, $pn_{iops,k}$, $pn_{ltc,k}$	The penalty function, and its parts iops penalty and latency penalty
$iops_{offered}(u_k)$	The IOPS offered to customer u_k
$ltc_{offered}(u_k)$	The latency offered to customer u_k
Storage	
SC, sc_j	The set of storage classes, The jth storage class
$sc_{j,int}$, $sc_{j,ext}$	The internal and external customers storage parts
c_{sc_j}, p_{sc_j}, wo_{sc_j}	The capacity, price and wear out of sc_j
Bandwidth	
bw, p_{bw}	The bandwidth and its purchased cost
bw_{int},bw_{ext}	The internal and external customers bandwidths
Objects	
$o_{i,k}$, $s_{o_{i,k}}$	The ith object of customer u_k and its size
$req_{op,o_{i,k}}$	The average IOPS of type op issued to the object $o_{i,k}$
$O_{m,int}$, $O_{m,ext}$	The set of internal (external, respectively) customers objects to migrate between internal storage classes
$A[i,j], B[i,j]$	the internal and external customers objects placement matrices
General	
T	Period of time
T_{sp}	The internet subscription period
Costs	
$Cost_{plc,T}$	The total placement cost
$Cost_{plc_{int},T}$	The internal customers objects placement cost
$Cost_{plc_{ext},T}$	The external customers objects placement cost
$Cost_{tcl_{int},T}$	The local placement costs of the internal customers
$Cost_{tcl_{ext},T}$	The local placement costs of the external customers
$Cost_{pnt_{int},T}$	The penalty costs of the internal customers
$Cost_{pnt_{ext},T}$	The penalty costs of the external customers
$Cost_{pnt_{iops},T}$	IOPS penalty costs
$Cost_{pnt_{ltc},T}$	Latency penalty costs
$Cost_{stg_{int},T}$	The storage cost of the internal customers objects
$Cost_{stg_{ext},T}$	The storage cost of the external customers objects
$Cost_{mgr_{int},T}$	The inner migration costs of the internal customers objects
$Cost_{mgr_{ext},T}$	The inner migration costs of the external customers objects
$Cost_{out_{src},T}$	The outsourcing cost
$Cost_{in_{src},T}$	The insourcing cost
$Cost_{geo_{mgr},T}$	The geo-migration cost
$Cost_{bck_{mig},T}$	,The back-migration cost
$Cost_{re_{mgr},T}$	The re-migration cost
$Cost_{geo_{rpl},T}$	The geo-replication cost
$Cost_{add_{rep},T}$	The cost of adding a replica
$Cost_{syn_{rep},T}$	The cost of synchronizing replicas
$Cost_{rd,T}$	The local read cost
$Cost_{rd_{erg},T}$	The read energy cost
$Cost_{rd_{edr},T}$	The read wear out cost
$Cost_{wr,T}$	The local write cost
$Cost_{wr_{erg},T}$	The write energy cost
$Cost_{wr_{edr},T}$	The write wear out cost
$Cost_{bw,T}$	The local consumed internet bandwidth cost
$Cost_{bw_{amz},1}$	The amortized internet bandwidth over one unit of time
$Cost_{ntw_{out},T}$	The external outgoing network cost
$Cost_{ext_{occ},T}$	The external occupation cost
$Cost_{ext_{wld},T}$	The external workload cost
$Cost_{ext_{plt},T}$	The penalty cost

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1035–1042
DOI: 10.15439/2018F49
ISSN 2300-5963 ACSIS, Vol. 15

Interactive development of cyber physical systems using UETPN model

Attila O. Kilyen
Department of Automation
Technical University of Cluj-Napoca
Romania
Email: kilyen.attila.ors@gmail.com

Tiberiu S. Letia
Department of Automation
Technical University of Cluj-Napoca
Romania
Email: Tiberiu.Letia@aut.utcluj.ro

***Abstract*—This paper presents a novel approach to synthesise hybrid controllers. A two-phase multi-objective evolutive algorithm was used to generate Unified Enhanced Timed Petri Net (UETPN) models. These models combine capabilities of timed Petri-nets, fuzzy logic systems and simple arithmetic operators. They can handle both event-like and continuous inputs (and outputs). The first phase of the algorithm uses Koza style genetic programming combined with multi-objective methods such as NSGA-II and SPEA2 to obtain an initial model. The second phase improves the initial model with recombining the fuzzy rules with genetic algorithm GA. In order to generate UETPN models (with GP), an intermediate language was designed, called UETPN Lisp. Four example are presented to exemplify the potential of the proposed framework.**

***Index Terms*—hybrid control, Petri nets, genetic programming**

I. Introduction

HYBRID controllers have a substantial practical importance because almost every real application from simple temperature control to complex robotic agents can have both event-like inputs and outputs, and continuous ones. Well-known examples of problems solved by Genetic Programming (GP), such as the artificial ant and obstacle avoiding robot (presented by Koza at [1]), are formulated in a way that only one domain is involved. In this paper, a two-phased evolutionary algorithm is presented, which is capable of synthesising controllers for discrete event systems, discrete time systems and hybrid systems as well.

Unified Enhanced Timed Petri Net (UETPN) models are used as the target platform for the proposed evolutionary framework. They are based on Delayed Time Fuzzy Petri nets [2]. For an effortless expression of control algorithms, they were completed with mathematical operators. Their ability to competently model reactive applications is shown in [3]. They are fit for handling continuous (real number) variables and fuzzy logic variables and for performing simple arithmetical and logical operations. They are capable of modifying the execution (split, join, select or block) depending on some external or internal value.

The proposed platform generates a complete UETPN model with GP, and in the second phase, it tries to improve it by recombining the fuzzy rules with genetic algorithm (GA).

The overall proposed framework needs a fitness evaluator for the given problem as input. This fitness evaluator consists of a fitness function and a light-weigth simulator. The ouput is an UETPN model which can be employed to control the specified system.

In order to widen its applicability, the presented framework supports the usage of Pareto front-based multiobjective methods such as NSGA-II [4] and SPEA2 [5]. These methods can also help to reduce the bloat, an issue which typically affects GP [6]. The presented experiments highlight the general applicability of the framework. The proposed method is applied to four different problems, for which the fitness function has been changed.

II. Evolving Petri nets

Fuzzy Petri nets are applied in various fields: path-tracking control problems; adaptive task assignment; fault estimation, detection, and diagnosis for power systems; urban and rail traffic control and many more [2]. Despite the vast range of applications, surprisingly few attempts were made to generate them automatically. Wong in [7] presents a framework called LOGENPRO used to extract knowledge from databases. LOGENPRO uses GP applied to logical grammars, based on place-manipulation and transitions-manipulation operators (such as sequential or parallel division). These operators modify the predefined fuzzy Petri net (FPN). However, the overall process restricts the structure of the FPN. The overall result lacks any internal state, timing and inner loops. These restrictions limit the applicability of the proposed framework in control applications.

Nobile in [8] introduces a new type of PN called Resizable Petri Net. This have divided the places and transitions into two groups: hidden places (or transitions) and normal ones. In the proposed framework only the number of the hidden nodes varies, the resulted PNs resemble binary trees.

An entirely different approach is presented in [9], which evolves some parameters. Based on these parameters and a template the final PN is assembled. This approach allows the use of the traditional real-coded genetic algorithm. Nevertheless, it can be applied only in case the structure of the PN is known upfront. Another approach presented by [10] addresses a biological problem modelled by PNs. In this case,

TABLE I: Mapping table (i. e. MT) associated with *t4* of the first example (only the indices of the rules are marked)

$P9/P8$	X_{-2}	X_{-1}	X_0	X_1	X_2	ϕ
X_{-2}	ϕ	ϕ	ϕ	ϕ	ϕ	-2
X_{-1}	ϕ	ϕ	ϕ	ϕ	ϕ	-1
X_0	ϕ	ϕ	ϕ	ϕ	ϕ	0
X_1	ϕ	ϕ	ϕ	ϕ	ϕ	1
X_2	ϕ	ϕ	ϕ	ϕ	ϕ	2
ϕ	-2	-1	0	1	2	ϕ

the authors establish the number of the places based on the problem specification, traditional GA evolves only the arcs. The presented problem is solved, still, their approach is far too restrictive to be applied in a broader domain.

A more general idea is presented in [11], using traditional GA. The gene, a transition-place pair, is decoded as an arc linking them. They use their framework to synthesise PNs starting from a structural and behavioural specification. Following the evolutionary process, additional steps are needed to improve the result and to reduce its size. This approach has potential, yet we believe that GP is abler to discover and re-utilise building block in the case of a PN-like structure.

III. Unified Enhanced Timed Petri Net

The UETPN models ([12]) incorporate transitions that have associated mapping tables (MT) and optional arithmetic operators. The MT is an organised form of the fuzzy rules which determine the bahaviour of the transitions. The input token(s) can activate the rules and if there is no operator, these rules define the output of the transition and they decide whether a transition is executable. UETPN supports transitions with one or two input places and one or two output places.

The mapping of a transition can be defined as a function between the current marking of the pre-places and post-places.

MT can have different shapes based on the numbers of inputs and outputs. An MT with two inputs and one output is exemplified in Table I, while another with one input and two output in Table II. If the current marking of a place is marked with x_p, then one cell represents the following fuzzy rule:

$$IF\ x_{i0} is X_0 \wedge x_{i1} is X_{-2}\ THEN\ x_{o1} is X_1 \wedge x_{o2} is X_2 \quad (1)$$

Each place has an associated scale (s_k). The token t_k set in a place is always in $([-s_k, s_k] \cup \phi)$, where ϕ means *no information*. Petri nets express *no information* leaving the place empty, nevertheless, in the case of the UETPN a place always has a marking.

When a transition executes, the input tokens are fuzzified in the first step. The limits of the membership functions are defined based on the scale of the input place. Simple triangular membership functions are used. Secondly, the fuzzy rules in the MT are executed, the result is collected and defuzzified by the center-of-gravity method. The scale of the output place(s) determines the defuzzification intervals.

If an arithmetic operator is assigned to the transition, the following equation is applied:

$$map_i(x_{i1}, x_{i2}) = (x_{i1} \circ x_{i2}) \star FL_{MT}(x_{i1}, x_{i2}) \quad (2)$$

TABLE II: Mapping table (i. e. MT) associated with *t3* of the first example (only the indices of the rules are marked)

X_{-2}	X_{-1}	X_0	X_1	X_2	ϕ
$0, \phi$	$0, \phi$	ϕ, ϕ	$\phi, 0$	$\phi, 0$	ϕ, ϕ

where $\circ \in \{+, -, /, \times\}$, and $FL_{MT}(x_{i1}, x_{i2})$ stands for the result of the mapping table deffuzified in the interval $[-1, 1]$. The fuzzy rules can alter the original result, but, if all the conclusions are X_2, the result of the operator is unchanged. The obtained value is truncated based on the scale of the output place. Only the transitions with two input places can have operators.

The existence of every fuzzy rule is not compulsory for MT construction. ϕ signals the missing rules. The MT also has ϕ columns and rows, which supports the definition of rules even if one (or both) of the input tokens are ϕ.

Another role of the MT is to decide whether a transition is executable (referred to as enabledness). A transition is allowed to fire if there is at least one fuzzy rule with $non-\phi$ consequence (in the MT) which applies to the current input marking.

This definition of enabledness and the possibility to put ϕ in some cells of the MT facilitate the implementation of inhibitor arcs, reset arcs, and transitions which are always enabled or blocked.

A precise description of communication with the outside world is pivotal for the UETPN to model hybrid controllers. UETPN models represent input channels as input places. The tokens set in these places can emerge from the exterior world (environment) only. The output channels are represented as output transitions. The output transitions do not have post-places, they send the tokens outside the current component. Multiple UETPN components can be connected in the previously described way.

A. Definition of UETPN

[3] contains not only the complete definition of the UTPN models, but some significant examples and applications also. In this section, only a brief introduction is given. The examples do not resolve a real-life problem, however they illustrate some of the capabilities of UETPN models (and they exemplify the UETPN-Lisp introduced in the next section as well).

The definition of UETPN is:

$$UETPN = (P, T, pre, post, D, S, EFS, Map, Inp, Out, \alpha, \beta, \delta, \mathbf{M}, M^0)$$

where:

- P is the place set, T is the transition set ($P \cup T = \emptyset$), while $pre \subset (P \times T)$ contains the arcs from places to transitions, $post \subset (T \times P)$ includes the arcs from transitions to places. D is the delay set. δ is a mapping $\delta : T \to D$, which associates delays to transitions. Their meaning corresponds to the ones from classic Petri nets.
- $Inp \subset P$ are the input places (channels), $Out \subset T$ are the output transitions (channels).

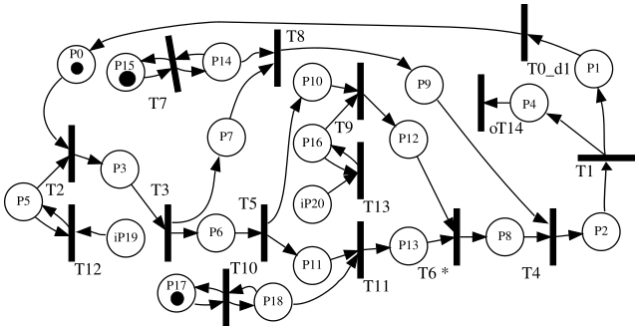


Fig. 1: Structure of the expression *(# (@ (@ i:br:0 (% (* i:br:1 c:2.0) c:0)) o:c:0) d:1)*

[t]

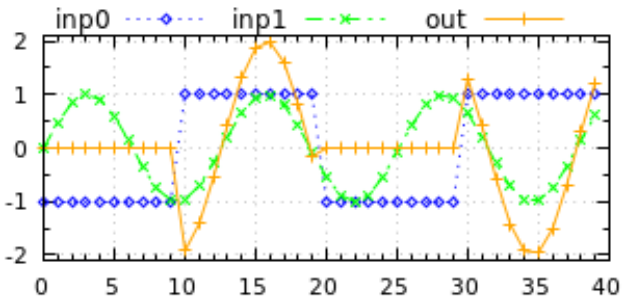


Fig. 2: Behavior of the first example

- $S = \{s_0, s_1, s_2 \ldots, s_m\}$ is a set of real numbers representing the scale factor set or bound set, α is a mapping such that $\alpha\colon P \to S$, it assigns a scale to each place from the set P.
- **M** is the marking vector, while M^0 is the initial marking vector.
- EFS is the extended fuzzy set $\{X_{-2}, X_{-1}, X_0, X_1, X_2\} \cup \phi$,
- $Map = \{map_i | i = 0, 1, 2, \cdots, n\}$ is the set of mappings. A mapping consists of one particular fuzzy logic rule set, expressed in a form of mapping table (MT) and an optional arithmetic operator. β is a mapping $\beta\colon T \to Map$, it assigns a mapping from the set Map to each transition

B. Example of a UETPN model

The first example illustrates how UETPN models can alter their execution based on the value of an (input) token. It demonstrates how mathematical operations can be implemented too.

Figure 1 represents the structure of the first example of a UETPN model. It has two input places and one output place. If the zeroth input channel has a positive non-ϕ token, the output equals two times the first input. Else, the output returns a constant zero token. The transition t3 has a key role in this behaviour. Its MT table returns a token to P6 if the value in P3 is positive, and to P7 if it is negative. Table II specifies the mentioned MT table. The constructions P17-t10-P18 and P15-t7-P14 are responsible for providing the constant values, P15 and P17 contain an initial token with the value of 2.0 and 0, respectively. T6 multiplies the input read from iP20 (via P12) with the constant from P17 (via P13). T4 has to have an MT table which enables if there is a non-ϕ token in P9 or P8, because only one main branch is executed simultaneously. Table I presents this MT. The transition oT14 fires the result to the outside world. Note, that t0 has one delay before the whole cycle starts over.

Figure 2 displays the behaviour of the described example. It can be observed that when the zeroth input is -1 (between the 0th and 10th time tick and between 20th and 30th time-tick), the output is zero. When the zeroth input is 1, the output doubles the sine wave of the first input.

The second example also alters its execution, however, in a completely different way. In contrast with the first example, the execution flow depends on whether a transition is executable. Note that the definition of the UETPN models does not prioritize the transitions, however, one can be sure that the executor tries to fire the transitions with the lowest delay at first. Another restriction is that in the case of conflicting transitions, the executor behaves in a deterministic way.

Figure 3 displays the structure of the second example. This example has two inputs and one output. All the places have 1.0 as scale. It contains an inner loop which executes until the outside environment inserts a token to iP11, which in turn is transmitted to P5. When a token appears at P4, t1 becomes executable, and the model stops its execution. The inner loop returns a token to the output oT9 if in the input place iP10 has a token which activates the X_0 rule of t5. Table III specifies the MT of t5. If t5 is not fireable, T4 starts firing. After T4 finished its firing (it has one tick delay) the whole cycle begins over via T3.

Figure 4 shows the behaviour of the second example. The output is activated when the zeroth input is in the interval $[-0.5, 0.5]$ which corresponds to the activation of rule X_0 after fuzzification. It is also observable that after an event appears in the first input (at the 30th time-tick), there is no more output regardless of the inputs.

Note, that the transitions t12 and t13 from the first example, t8 and t7 from the second example have the role in copying the new input token regardless of the existence or value of the previous inputs (stored in the buffer places after these transitions). These structures are not essential from the functional

TABLE III: Mapping table (i. e. MT) associated with *t5* of the second example (only the indices of the rules are marked)

$P2/P8$	X_{-2}	X_{-1}	X_0	X_1	X_2	ϕ
X_{-2}	ϕ	ϕ	-2	ϕ	ϕ	ϕ
X_{-1}	ϕ	ϕ	-1	ϕ	ϕ	ϕ
X_0	ϕ	ϕ	0	ϕ	ϕ	ϕ
X_1	ϕ	ϕ	1	ϕ	ϕ	ϕ
X_2	ϕ	ϕ	2	ϕ	ϕ	ϕ
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ

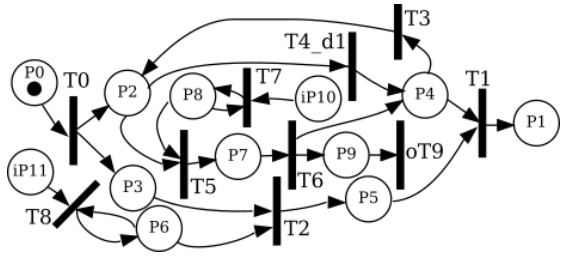


Fig. 3: Structure of *(& (#(?(@ i:eiz:0 o:c:0) d:1) d:0)i:enp:1)*

point of view, but since these models are direct correspondents to some UETPN Lisp expressions, it would be imprecise to omit them.

IV. Model construction

The semi-automatic way of fabricating a UETPN model has two phases. In the first one, an initial model is generated with GP and UETPN Lisp. It constructs a complete UETPN model, which is executable and evaluable with a fitness function. However, the available MT tables are predefined, which limits the possible behaviours. The result of the first phase is not only the UETPN model itself but also a meta-data about the context of transitions. Based on this metadata the MTs of some transitions can be re-trained with GA. This preselection of the transitions is necessary because in most cases the models yielded by the first phase have so many transitions that it is infeasible to optimize all of the MTs with GA.

There can be several reasons to employ the second phase. First of all, if the fitness value produced by the result of GP is not satisfactory, there is a chance that the second phase will improve it. Secondly, the fitness itself can be adjusted, taking into account other evaluation criteria. Thirdly, the parameters of the problem can also be modified. In this case, the re-training of the UETPN model achieved in the first phase may lead to significant improvement.

Although the entirely automatic synthesis is possible, it is highly recommended to analyse and evaluate the solutions manually after the first phase. It is recommended to ensure that the found solution has the desired behaviour, because the GP often can find a workaround to achieve a high fitness value without satisfying the real fitness criteria. The length of the

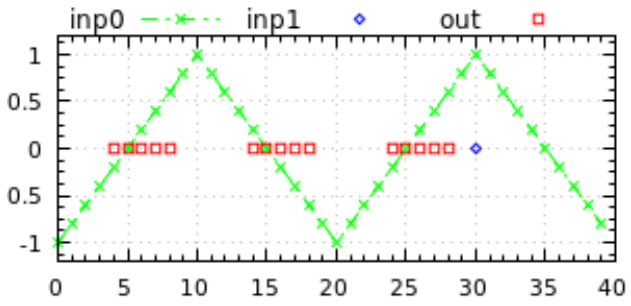


Fig. 4: Behavior of the second example

simulation for fitness evaluations is always a compromise, because the longer the simulation is, the more time it takes to run the algorithm. It is advised to execute the first phase several times and to choose a suitable solution before performing the second phase. Also, the types of transitions whose MTs are optimised have to be selected manually.

Finally, it has to be mentioned that in the case of problems which do not benefit from the effect of fuzzy rules, the second phase has no benefit. In these cases, the problem has to be solved in the first phase.

A. First phase:GP with UETPN Lisp

In the first phase, classical tree-based Koza style GP is used. The overall algorithm and the genetic-operators are not specified here, they are applied as in the literature ([13] and [1]). The focus of this section is on the UEPN Lisp, which is a small language used by GP framework. It presents the operators and the operands and exemplifies he conversation from UETPN Lisp to UETPN model.

UTEPN Lisp is transformed to UEPN model by breadth-first traversal. Every sub-expression is built up between two places. The algorithm starts by adding the two principal places. The root node is built between these. The first place has an initial token which begins the execution of the model. Both in case of the first example (Figure 1) and the second example (Figure 3), this starting token is placed in P0.

All of the operators of the UETPN Lisp have two operands. The sequence operator denoted by @ is the simplest one, it indicates that the two operands come after each other separated by a place. In the second example, the place P7 separates the transition T5 and T6, a structure which is the result of decoding a sequential sub-expression "*(@ i:eiz:0 c:c0)*". The selection operator (marked with *?*) builds up both of its children between its original starting and ending place. It can be noticed that either the transition t4 or the structure t5-P7-t6 executes from the second example. This is the result of a sub-expression of *(? (@ i:eiz:0 c:c0) d1)*. The loop operator (#) produces a similar structure, however, the direction of the second child is reversed. In the second example, T3 is the second child of a loop operator. T0 is the second child of a loop node in the first example.

The structures produced by the concurrency operator (&), positive-negative split operator (%), sum operator (+) and multiplication operator (*) are the same, the only difference are the MT tables and the mathematical operators associated with the transition. In the first example, the structure from P3 to P2 is the result of the sub-expression *(% (* i:br:1 c:2.0) c:0.))*, and the transition T3 and T4 are built as the part of the positive-negative split operator. The T3 has an MT which yields a token into one of its outputs only, based on the sign of the input. In the structure mentioned above, the fragment staring from P6 to P8 is built as the result of the sub-expression *(* i:br:1 c:2.0)*. The transitions T5 and T6 were added as a result of the multiplication operator.

In case of the second example, the root of the expression is a concurrency node. The transitions t0 and t1 are added as

part of this operator. As the name suggests, in this case, both branches are executed independently, and t1 becomes fireable if both finish their execution. If the sum or multiplication operator is used, T1 has the corresponding associated operator.

The most crucial operands are the input and the output operands. They connect the main flow of the model to the input place or output transition. They also have a type which specifies the MT tables used in the connection transition. For example, the input node *i:eiz:0* means that the zeroth input is read with a transition which is enabled if the input token activates the X_0 rule. In this case, the MT mentioned above belongs to t8 (second example), and iP10 is the zeroth input.

The P8 is a buffer place where each input node has a copy of the original input token, and t7 is an auxiliary transition, responsible for placing the new token to the buffer place. This role is essential in the case of models which use the same input multiple times. Other types of inputs are the blocking reader ("*br*"),non-blocking reader ("*nbr*"), enable if non-ϕ ("*enp*"), enable if phi ("*eip*"), enable if zero ("*eiz*"), enable if not-zero ("*enz*"). The type of the input is essential in case of the second example, where the zeroth input has " enable if zero" type. If the evolutive framework mutates it into "enable if not-zero" ("*enz*"), the behavior of the model is inverted. If changed to something else, the execution of the model would be dramatically different. Note that these input types help the framework to deal in a completely different way with event-like and continuous inputs . The ones with *"enable"* in their names block the token in the main flow without modifying its value. The *"reader"* ones copy the value of the input token into the main flow.

Similarly to input leaves, outputs also need auxiliary constructions. In the second example, P9 acts as a buffer place connecting T6 to the real output transitions oT9, which is the result of the leaf *o:c:0*. Outputs currently can be only *copy* type, however, defining new types is as easy as to define a new MT table. (This applies to the input types as well.)

Other operands are the delay nodes (*d:nr*), which insert transitions into the flow with a certain delay, the constant (*c:nr*) leafs that provides a mathematical constant, blocking leaf (*b*), whose role is to insert a transition which is never executable, negation leaf (*n*) which negates the sign of the tokens, inversion leaf (*v*), and memory leaf (*m:nr*) which delays the value of the token, but it does not block the execution of the model. Some of these (memory, inversion, constant) are implemented with complex structures, others are single transitions with unique MT tables and/or mathematical operators.

During the decoding, the created transitions are categorised as follows: input, output, split-starter, split-merger, auxiliary, others. These categories (one or more) can be selected for optimisation by the second phase. The early experiments showed that in the majority of the cases optimising the input, the output and the split-merger transitions yielded the same results as adding any other category to this group. What is more, widening the set of optimised transitions means larger search space for the second phase.

B. Second phase: GA for the fuzzy rules

Similarly to the previous section, the GA itself and the genetic operators are not presented here, they are used in the well-known way ([14]). Simple binary encoding is used. The user of the algorithm decides which type of transitions have to be optimised.

In order not to change the behaviour drastically, only the non-ϕ rules were marked to be re-trained by the presented GA, and they can trun to other non-phi rules. Three bits can represent the five possible rules. A three-bit unit was chosen as the gene, their sequence composes the chromosome for GA. Since three bits can represent eight values, three of them is not used. The crossover, the mutation and the creation of initial population were modified in order not to produce the unused combinations.

V. Experiments

A. Control of a first order system

A simple task was choosen to exemplify the working of the framework: control of a first order system. Firstly, a known structure is optimized. Secondly, the complete framework solves the problem.

The fitness functions used for these problems take into account the absolute error (e_a) and the steady state error(e_s) of the controlled system. All of the experiments use the same fitness function:

$$f(i) = 1/(1 + \alpha * e_a + \beta * e_s)$$

where α and β are constants set to 0.8 and 0.2.

Firstly, a proportional integral (PI) controller was manually defined in UETPN Lisp. The rules of the input, output and the split-merger transitions were optimised. After 50 runs, the average fitness was 19.46. One of the average results is presented on the upper part of the Figure 5.

Secondly, GP solves the same problem with the same fitness function. The after 50 runs, the average fitness of the results was 25.39. An average result is displayed in the middle part of in Figure 5. Not only that the second result has hogher fitness, but the difference is also conspicuous, the overshoot is better, and the overall error is smaller.

Thirdly, GA is applied to optimise the rules of the previous result. The average fitness in this case is 27.88, which means smaller increment, however, the differences (the lower part of the Figure 5) still can be noticed.

Additonally the result of the second step was re-trained with GA to control another modified first order discrete system. The solutions have superior performance than the orginal controller, however, they do not overperform a re-traned PI controller.

B. Artificial Ant Problem

The Artificial Ant is a classic GP problem used by Koza at [13] to illustrate GP. The previously presented UETPN Lisp has the disadvantage of being more general, in contrast to Koza's solution (and many other papers), which use specialised operators and operands. As it is anticipated, the

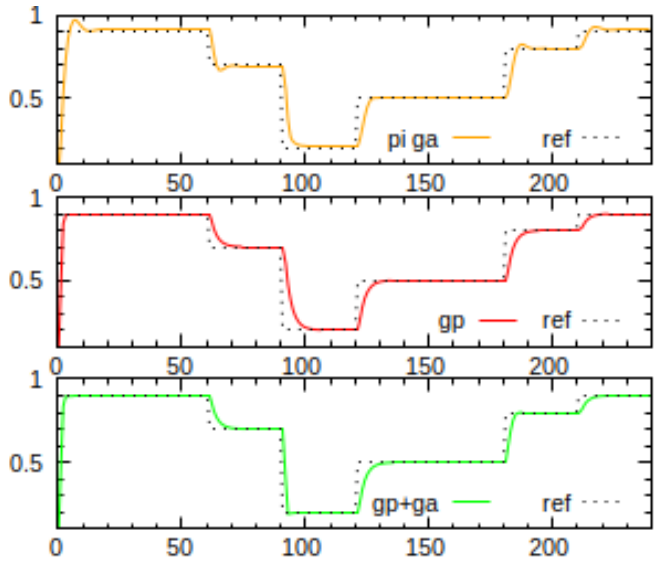


Fig. 5: Comparison of different results for the original system (PI optimized, GA only and GA+GP)

performance of the general framework does not match the specialised one.

An artificial ant is capable to turn left or right and to move forward. The single input is activated if food is ahead. The controller has three event-like outputs and one even-like input. The fitness function is the number of food units eaten. The ant is placed in the well-known Santa Fe trait.

The overall problem is known to be difficult to solve and it was used by several researchers to demonstrate features of their algorithm. A comprehenesive review of the problem can be found at [15]. The advantages of flexible genome show up in the case of this problem, because fixed length genome can represent only restricted number of program states.

Since this problem requires higher population number and iteration than the previous one, bloat presents a more serious complication. Bloat is the phenomenon when tree-sizes grow exponentially fast, wasting the computational time and filling up the overall genetic material of a population with useless fragments. The proposed framework implements traditional methods such as various form of Parsimony Pressure, which essentially means that the size of the individual is part of the fitness.

The problem with these methods are that it is hard to assess the size compared to the original fitness value. The demonstrated framework can also apply static and dynamic simplification over UETPN-Lisp. These methods replace the sub-expressions, which can be expressed simpler and delete the unused ones. Although they are effective on compacting the potential solutions, they cannot entirely solve the bloating phenomenon.

A robust method in the case of UEPN-Lips applied to the artificial ant problems is to use NSGA-II with two fitness functions. The first of them is the number of food eaten (f_f), the second one is the number of eaten multiplied by the size factor (f_s). This way the second objective favours candidates which are small and fit. This idea gives better results than the one presented [5] where the second objective evaluates only the size. The size factor (f_s) is calculated by:

$$f_s = (s_i - s_{prf})/(s_{max} - s_{prf})$$

where s_i is the size of the individual, s_{max} is the maximum allowed size, while s_{prf} is the preferred size. In these experiments $s_{max} = 500$ and $s_{prf} = 20$, outside these limits f_s is defined as 1 (if $s_i < s_{prf}$) or 0 (if $s_i > s_{max}$).

The original solution presented by Koza needs to evaluate $450 * 10^3$ individuals for an acceptable average solution ([15]), similar results can be obtained with $500 * 10^3$. The best algorithm proposed by [15] is equipped with problem-specific base language but also specialised crossover and mutation, and it has a far better success rate with $51 * 10^3$ evaluated individuals.

C. Room temperature control

This experiment is a hybrid application with two inputs: the reference temperature and the actual reading from the heat sensor in the room. There are two event-like outputs: one of them starts the heating, the other one stops it.

A discrete time system simulates the room temperature. It is based on the temperature differences:

- $\delta_{ht}[k] = t_{hw}[k] - t_r[k]$ is the difference between the temperature of the heating water (t_{hw}) and the room temperature (t_r).
- $\delta_o[k] = t_r[k] - t_o[k]$ is the difference between the outside temperature (t_o) and the room temperature.

The room temperature is simulated in the following way:

$$t_r[k+1] = t_r[k] + c_{ht} * \delta_{ht} - c_{wl} * \delta_o - c_{wi} * \delta_o$$

, where c_{ht} is the heating constant set to zero if the heating is turned off. c_{wl} is the wall constant, and c_{wi} is the window constant set to 0.0 zero of the window is closed. The opening and closing is the disturbance of the overall system. In this case, the temperature of the heating water is considered constant. Another important point is that the sensor reading is delayed compared to the simulated room temperature.

A elementary fitness function can be defined with the help of the sum of the error. However, this would lead to controllers which do not react to the input values but rather turn on/off the heating periodically. The problem can be solved by adding a lot of test scenarios with hectic temperature changes. This would lead to slower evaluation of a solution candidate, hence longer overall runtime. The approach used here evades the elongated run-time: the number of the minutes when the temperature is outside the interval of $t_{ref} \pm \delta$ is mesured, where δ was chosen to be 0.5. The behaviour of a solution found in the first phase (with GP) is presented in Figure 6. The controller turns the heater on only when it is necessary, and the heater stays on until a certain limit is reached. The desired behaviour is achieved, however the GA was not able to improve it.

In the second part of the experiment, the parameters of the room model were changed. The disturbance effect was higher

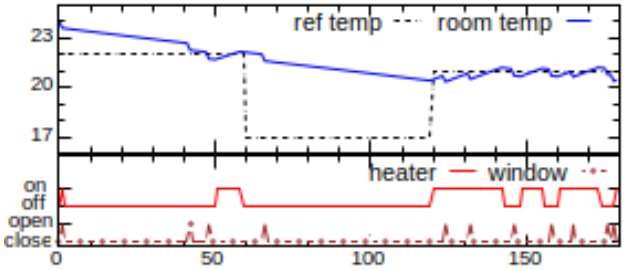


Fig. 6: The evolution of room temperature in case of GP

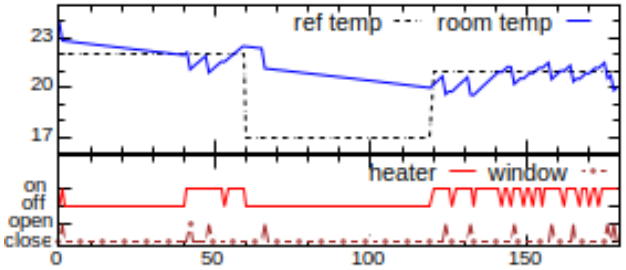


Fig. 7: The evolution of room temperature in case of the modified system

than in the previous experiment (presumably because the window is larger or the volume of the room is smaller compared to the window), while power of the heater was increased. In this case, the original controller performed poorer. The behaviour of a solution which has the same structure as the original one but it was re-trained with GA taking into consideration the new room model. The new behaviour of is shown in Figure 7. The presented solution was able to respond to the new challenge.

D. Room and Water Heater Temperature control

In this section, the desired controller has to control the temperature of the heating water, in contrast with the previous part where it was considered constant. This way the controllers have three outputs, first two of them have same functionality as in the previous section.The third has continuous behavior, aimed to control the water temperature in the heater tank. There are four inputs: the actual temperature of the room, the reference temperature of the room, the actual temperature of the water and the reference temperature of the water.

With the objective to simulate the temperature of the heating tank, the following intermediate variables are defined:

- $\delta_w[k] = t_{hw}[k] - t_pw[k]$ is the temperature difference between the heating water (t_{hw}) and the water from the pipe (t_{pw}).
- $\delta_{cmd}[k] = t_{max} - t_{hw}[k]$ is the difference between the maximum water temperature (t_{max}) and the current heating water.
- $\delta_t[k] = t_hw[k] - t_t[k]$ is the temperature difference between the heating water and the room.

The temperature of the heating water is given by the following discrete time equation:

$$t_{hw}[k+1] = t_{hw}[k] - c_h * \delta_w + c_{cmd} * u[k] * \delta_{cmd} - c_t * \delta_t$$

where c_h, c_{cmd}, c_t are constant, c_h is set to 0 if the heating is off. $u[k]$ is the output of the controller. The rest of the system is identical to the one mentioned before.

This system can be viewed as the combination of the first order system, and the setup presented in the previous section. Although it is self-explanatory that the framework can express the desired controller, it is hard to come up with the adequate fitness objective(s). The only known solution needs multi-objective optimization and far more evaluated individuals than any of the examples presented here.

The first objective is similar to the one in the previous section it counts the number of the minutes when the room temperature and the temperature of the water is off limit. The second fitness objective sums up the error and of the room temperature and the temperature of the water relative to the references. (The reference for the water temperature is constant 60). The third one is identical to the first one, except that it takes into account the size factor presented in the section related to the artificial ant. SPEA2 is known to handle better more than two objectives than NSGAII [5], hence it was chosen as the base algorithm for the experiment.

As it has turned out, the boiler is too weak compared to the room model. In practice, this means that when the heating is turned on in the room, the temperature of the water drops fast and constant temperature cannot be maintained for more than two or three minutes. In the real world, this would be a severe design problem, but the proposed framework overcame this flaw.

Figure 8 presents the behaviour of one solution found after GA is applied as well. This controller starts to heat up the water in the heating tank before the temperature of the room drops to a critical temperature. This way when the controller turns on the heating in the room, the water is already heated above the reference temperature, and the heating can remain on for a longer time. It is not the expected solution, but it is unarguably a creative one.

Figure 9 displays the operations of an another solution. This solution turns on the heating for a minute or two only in order to maintain the constant water temperature. This approach performs weaker in the first hour of the presented scenario than the previous one. However, in the rest of the time it has an adequate performance.

The first solution performs better from the perspective of the first fitness objective, while the second fitness objective favours the second solution. In this experiment, multiple fitness scenarios were needed in order to reduce the number of solutions which perform a cyclic behaviour, which fits the average cases, without reacting to the current situation.

The presented framework is capable of finding an acceptable solution in 10% of the cases. The population size was 3200 combined with 200 iteration. This took in average five hours, more than 50 experiments were performed.

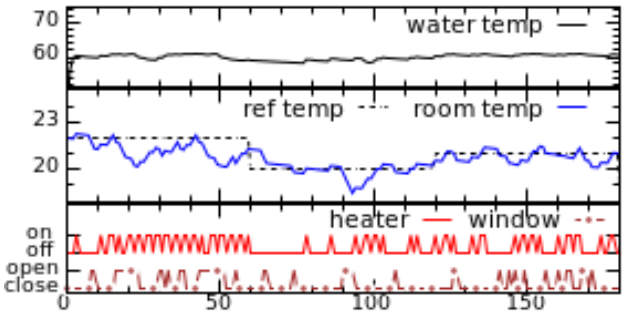


Fig. 9: The behavior of the second solution for Room and Water Heater Temperature control

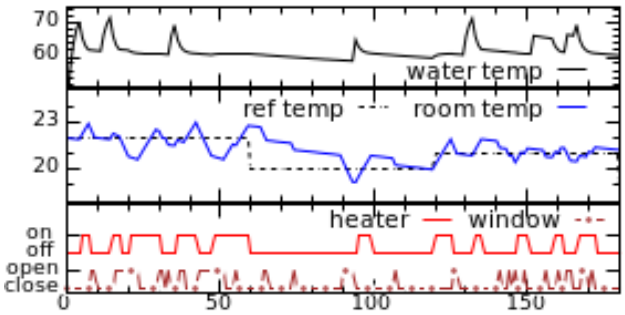


Fig. 8: The behavior of the first solution for Room and Water Heater Temperature control

VI. Conclusion

The current work focuses mostly on the capabilities of the framework and the effects of the two phases of the development. All of the presented experiments are reproducible with the code released [1] under an open-source license. Based on these experiments, it can be concluded that the given framework has the potential to generate controllers for hybrid systems. The presented experiments also highlight the fact that a complete base-language is not enough to tackle complex problems, the primary evolutive framework is at least that important.

The importance of multi-objective methods based on Pareto-fronts cannot be overstated. They have a strong focus on exploiting the known Pareto-front and on trying to improve in one way or another. This behavior is essential in case of controlling the bloat when shorter individuals are preferred, however, individuals with high fitness should not be deleted based on their size only. Yet, these algorithms have the disadvantage of being more disposed to early convergence to sub-optimal solutions. The explicit diversity control may be needed in the future.

Future development directions could focus on the compiling of the UETPN-model to machine code or Java Virtual Machine byte-code with the objective to make the evolution of a proposed solution faster. Another important direction is to conceive a method to deploy UEPN-models directly into micro-controllers with the aim to apply the presented result in real-life.

References

[1] J. R. Koza, "Genetic programming ii: Automatic discovery of reusable subprograms," *Cambridge, MA, USA*, 1994.
[2] K.-Q. Zhou and A. M. Zain, "Fuzzy petri nets and industrial applications: a review," *Artificial Intelligence Review*, vol. 45, no. 4, pp. 405–446, 2016.
[3] T. S. Letia and A. O. Kilyen, "Unified enhanced time petri net models for development of the reactive applications," in *2017 3rd International Conference on Event-Based Control, Communication and Signal Processing (EBCCSP)*, May 2017, pp. 1–8.
[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[1]https://github.com/AttilaOrs/FuzzP/tree/genetic

[5] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm," *TIK-report*, vol. 103, 2001.
[6] S. Bleuler, M. Brack, L. Thiele, and E. Zitzler, "Multiobjective genetic programming: Reducing bloat using spea2," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1. IEEE, 2001, pp. 536–543.
[7] M. L. Wong, "A flexible knowledge discovery system using genetic programming and logic grammars," *Decision Support Systems*, vol. 31, no. 4, pp. 405 – 428, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923601000926
[8] M. S. Nobile, D. Besozzi, P. Cazzaniga, G. Mauri *et al.*, "The foundation of evolutionary petri nets." in *BioPPN@ Petri Nets*. Citeseer, 2013, pp. 60–74.
[9] A. Gudelj, D. Kezić, and S. Vidačić, "Marine traffic optimization using petri net and genetic algorithm," *PROMET-Traffic&Transportation*, vol. 24, no. 6, pp. 469–478, 2012.
[10] J. Nummela and B. A. Julstrom, "Evolving petri nets to represent metabolic pathways," in *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '05. New York, NY, USA: ACM, 2005, pp. 2133–2139. [Online]. Available: http://doi.acm.org/10.1145/1068009.1068361
[11] T. Bourdeaud'huy and P. Yim, "Petri net controller synthesis using genetic search," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, Oct 2002, pp. 528–533.
[12] T. Murata, "Petri nets: Properties, analysis and applications," *Proceedings of the IEEE*, vol. 77, no. 4, pp. 541–580, 1989.
[13] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.
[14] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.
[15] D. Wilson and D. Kaur, "How santa fe ants evolve," *arXiv preprint arXiv:1312.1858*, 2013.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1043–1051
DOI: 10.15439/2018F173
ISSN 2300-5963 ACSIS, Vol. 15

The Evolution of a Healthcare Software Framework: Reuse, Evaluation and Lessons Learned

Alessandra A. Macedo, José A. Baranauskas
Universidade de São Paulo,
Departamento de Computação e Matemática,
Av. Bandeirantes, 3900 - Monte Alegre,
14040-901 Ribeirão Preto-SP, Brazil
Email: {ale.alaniz, augusto}@usp.br

Renato de F. Bulcão-Neto
Universidade Federal de Goiás,
Instituto de Informática,
Al. Palmeiras, Quadra D, Câmpus Samambaia,
74690-900 Goiânia-GO, Brazil
Email: renato@inf.ufg.br

***Abstract*—The literature describes examples of software frameworks providing developers with generic and reusable functionality for building healthcare applications. Using concepts and technologies from Information Retrieval, Machine Learning, and Semantic Web, we present a novel software framework called HSSF (Health Surveillance Software Framework) which aims to facilitate the development of applications to support health professionals in the prevention of chronic diseases. The main contribution of this paper includes lessons learned distilled from (i) the reuse and evolution of the HSSF components on the development of three new health surveillance applications, and (ii) a quantitative evaluation of the HSSF reusability in terms of time spent and artifacts reused on such development task. Lessons learned are summarized as advantages and drawbacks regarding HSSF reusability. The HSSF allows healthcare applications not only to relate scientific research evidences, exams and treatments, but also to incorporate them together into the clinical practice.**

I. Introduction

THE practice of software reuse can potentially make information technology products more efficient for clients and cheaper for the production. Reuse is not limited to the source code or to the machine code; in a broader sense, documents, coding styles, components, models, patterns and knowledge items may be reused.

One way to promote reuse is to create software frameworks as abstraction of functionalities to build and deploy applications. We consider software frameworks as reusable designs of all or part of a software system described by a set of codes, libraries, tools, APIs, and mainly by abstract classes and the way that instances of these classes collaborate [1]. Therefore, the main goal of a software framework is to create new software more efficiently by means of reuse.

As in other knowledge domains, there are various examples of software frameworks targeting the development of healthcare applications. For instance, OpenMRS [2] [3] describes data items such as clinical findings, laboratory test results or socioeconomic data that can be stored by medical systems. OpenMRS consists of a concept dictionary to avoid the need to modify the database structure to add new diseases. This framework supports the programming of new functions without the need to modify the core code when the focus is the reuse of concepts and/or database models. IBM has developed the SpatioTemporal Epidemiological Modeler (STEM), which models infectious and vector borne diseases. STEM provides developers with a plug and play software architecture for the development of simulations of disease spread, e.g. in bioterrorist crises [4].

In terms of reuse of services, the service-oriented architecture (SOA) paradigm is used as a software framework that aims at facilitating the support for clinical decision [5]. At the level of classes reuse, the virtual reality domain also offers frameworks such as (i) ViMeT [6], which focuses especially on the development of applications that simulate biopsy exams, and (ii) SOFA [7] [8], which targets research into medical simulation. Still considering medical simulation, TES [9] carries out computational simulations of transcranial electrical stimulation.

Despite of the large amount of software frameworks support in healthcare, institutions, professionals and patients are still burdened with the amount of information created due to the mass adoption of the Internet and to the availability of several types of documents, including health records, social websites for health, medical images, among others. Furthermore, the information contained in such documents is too complex and semantically rich for traditional search engines to make sense of. This scenario can benefit from health surveillance systems which can recommend information related to the medical records of a given patient [10], or to a similar user, for example, to provide informational and emotional support in on-line social websites for health [11].

This paper (i) presents the Health Surveillance Software Framework (HSSF) as an evolutionary and reusable design of software components to support the development of health surveillance applications [10], (ii) evaluates aspects of the HSSF reuse, and (iii) lists the lessons learned along the evolution of the HSSF architecture.

The HSSF diagram, an object-oriented application framework, allows distinct health surveillance applications to be created by instantiation of its abstract classes. HSSF has been developed by evolving our previous software which were

This work was supported by the São Paulo Research Foundation (FAPESP - Proc. 16/13206-4) and the National Council for Scientific and Technological Development (CNPq)

designed as a reusable set of functionalities for surveillance systems. Hence, we built HSSF by generalizing software components from three previous endeavors: the *Automatic-SL* [12], the *CISS* [13], and the *FREDS* [14].

The *Automatic-SL* system assists healthcare professionals in their decisions recommending Surveillance Levels (SLs) to identify patients' healthcare needs. It can be used to recommend pediatric procedures in primary healthcare, and it also identifies significant risk factors and protective factors associated with the patients and their families. The *CISS* establishes associations between chronic diseases (cardiovascular diseases, diabetes and obesity) reported in scientific papers and a given patient's clinical record with genetic and epigenetics[1] risk factors. Finally, the *FREDS* effort focuses on the definition of conceptual mappings between the content of medical images and the textual information contained in medical records, applied in a scenario of computer-aided diagnosis in thyroid cancer.

By reusing HSSF components, developers have created three new healthcare applications: the *CISS+* [16], the *CISS-SW* [17] and the *QASF* [18] [19]. They answered a questionnaire that helped us understand the reuse of the HSSF mainly in terms of the number of reused artifacts and the time spent on reuse activities. Finally, lessons learned were elaborated from this whole experience with the HSSF components.

As far as we know, there has been no related research to HSSF in terms of two perspectives: (i) the same underlying theory and technology, including Information Retrieval, Machine Learning, and Semantic Web; and (ii) the health surveillance support, i.e. aiding the prevention of chronic diseases by alerting healthcare workers about risk factors through retrieval of published scientific papers with information on epigenetic risk factors, or even classification of patients into risk groups. Most frameworks in the healthcare domain have proposed reusing services (mainly simulations, electronic health records and medical decision support) and data model (ontologies, dictionaries and databases).

This paper is structured as follows: Section 2 the development of the HSSF framework; Section 3 describes the evaluation carried out as an attempt to understand the HSSF reusability. Finally, Section 4 discusses the lessons we have learned along the process, and Section 5 brings final remarks and perspectives for future work.

II. The development of HSSF

Here we briefly present how the HSSF was developed and evolved in terms of its software components architecture and its core systems including the *Automatic-SL* [12], the *CISS* [13], and the *FREDS* [14] systems. Further details can be found elsewhere[2] [10].

[1]People exposed to risk factors (e.g. food shortage) at the beginning of life can have altered the gene expression, which can impact adult life by posing higher risk of developing chronic diseases. Epigenetics studies those changes in the gene expression [15].

[2]Source code with documentation, papers and reports with UML models are available at the official HSSF website – *http://dcm.ffclrp.usp.br/hssf/*.

A. The HSSF architecture

The architecture of the HSSF is comprised of three main layers — *Presentation*, *Business*, and *Storage* — so that each layer contains its own modules to process documents, as depicted in Fig. 1.

The *Business* layer consists of abstract classes and external packages of utilities, both two in the target domain, as well as two connector layers, called *Communication* layers, which provide the required communication components to the *Presentation* and *Storage* layers.

The *Presentation* layer presents different views and templates such as Graphical User Interfaces (GUIs), which allows access by two main types of users: (a) healthcare professionals, who can analyze risk groups automatically classified by means of surveillance services, or who can receive recommendations of papers related to a given patient's clinical record (e.g. during a medical appointment); and (b) researchers interested in investigating the relationship between risk factors, chronic diseases, risk groups and patients' records.

The two *Communication* layers are composed of connectors for tools, ontologies and knowledge sources. The upper *Communication* layer allows the presentation of recommendations to end users (healthcare professionals and researchers) via GUIs. The bottom *Communication* layer integrates the *Business* layer with features provided by the tools (e.g. classifiers) and knowledge sources (e.g. pre-processed collections of scientific papers and ontologies), and also comprises modules for communication with databases.

In the *Business* layer, the *Search For Papers* module interacts with public repositories of scientific papers. This module collects and updates a collection of papers. Currently, the repository crawler uses concepts from ontologies of target domain to focus the crawler on topics of interest. There is no crawler for clinical records because all clinical records of interest are considered to be associated with scientific papers.

Also in the *Business* layer, the *Textual Processing* module is composed of programming utilities and modules for *Paper Processing*, *Clinical Record Processing* and *Natural Language Processing*. *Textual Processing* module processes textual information from a set of clinical records and collected scientific papers, which are all stored in the *Storage* Layer. Each document (clinical record or paper) is processed, so *Paper Processing* and *Clinical Record Processing* modules can identify simple and complex terms. The *Natural Language Processing* module applies natural language processing, such as processing of n-grams, stemming, removal of stopwords and recognition of concepts. The recognized terms are statically weighted and stored. The processing of clinical records is similar to the processing of scientific papers.

The *Concept Recognition* module manipulates linguistic and knowledge resources, which in turn support the association between different lexical concepts. Clinical records can be manipulated in one language and papers can be processed in another language. For instance, the *Concept Recognition* module exploits classes and methods from the Unified Medical Language System (UMLS) [20] to identify concepts

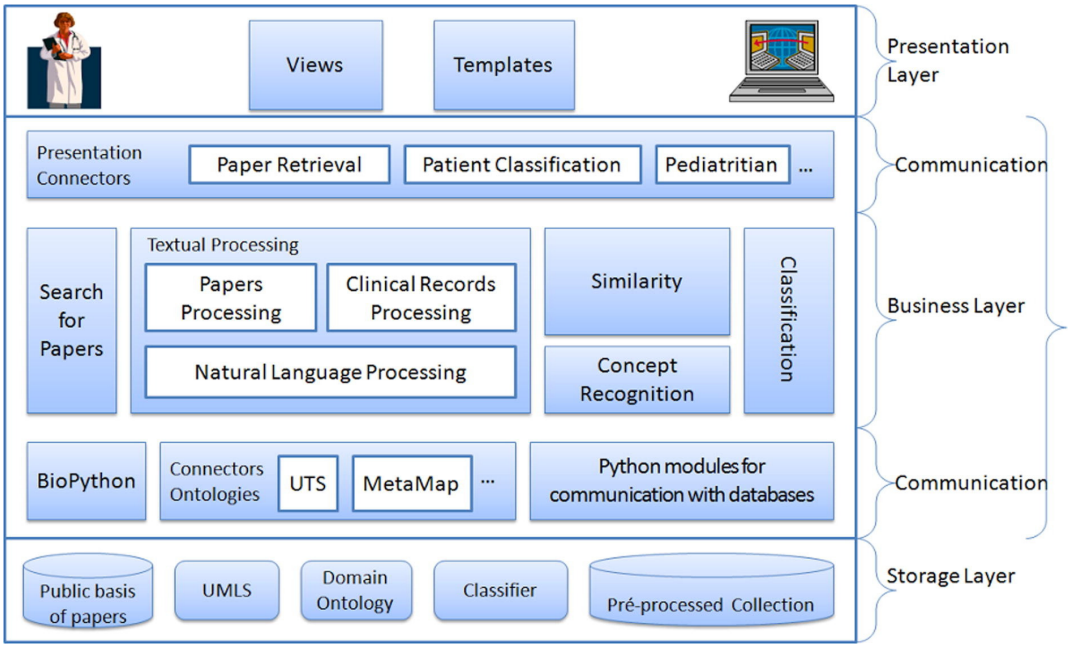


Fig. 1. The HSSF architecture [10].

from health topics such as epigenetics. The overall textual processing of the HSSF, supported by linguistic resources, includes the removal of stopwords, the processing of n-grams, the recognition of concepts, and the computation weights for concepts. The removal of stopwords for the papers and clinical records collections is based on lists of stopwords from programming utilities such as Snowball[3]. The processing of n-grams uses the open source Python NLTK[4] set of modules, linguistic data and documentation for research and development in natural language processing and text analytics.

The processing of clinical records is not identical to the processing of papers: after the processing of n-grams and the identification of concepts, a query array containing the remaining concepts is built for clinical records, whereas a weight matrix is composed of the scientific papers. Both the query array and the weight matrix are submitted to the *Similarity* module, which is in charge of computing similarities. The architecture of HSSF allows the *Similarity* module to calculate similarity measures between papers and clinical records, and it can also apply automatic relevance feedback.

B. The HSSF core systems

This section describes three health surveillance systems that we developed: the *Automatic-SL* [12], the *CISS* [13], and the *FREDS* [14] systems. Together these systems contribute with their main software components so as to build the HSSF architecture previously presented.

The *Automatic-SL* system aims to identify children with developmental problems and therefore assists healthcare professionals in their decisions and in reassessing recommendations as a multidisciplinary team [12]. After each medical appointment at a pediatric care center, this system collects patient information to automatically assign it a Surveillance Level (SL) measure which in turn indicates the type of healthcare procedure and service that a patient needs. Exploiting machine learning techniques, the *Automatic-SL* identifies significant risk and protective factors associated with patients and their families. Therefore, it provides surveillance indications that support preventive care to avoid diseases in adulthood.

As another surveillance service, the *CISS* system retrieves scientific papers that relate chronic diseases to genetic and epigenetic risk factors found in patients' clinical records [13]. From the PubMed repository, a *CISS* module routinely searches and retrieves new scientific papers in the domain of genetic and epigenetic risk factors for chronic diseases. Next, *CISS* processes textual information of that collection of papers for later retrieval of relevant papers according to a clinical record submitted by a healthcare professional.

To associate papers with a clinical record, *CISS* also processes its textual content and then calls a module which calculates the similarity among documents. In turn, this module accesses the pre-processed version of the collection of scientific papers to retrieve papers with the highest degrees of similarity to the clinical records. Selected papers are then presented to a healthcare professional with risk factors associated with the record previously submitted. By using this approach, healthcare professionals should be able to create a clinical routine with families and set up the best possible growing conditions.

Finally, the *FREDS* system was included during the design phase of the HSSF. Aiming to support decision making sys-

[3]A small string processing language designed for Information Retrieval purposes; documentation is available at http://snowballstem.org/.

[4]Documentation is available at https://www.nltk.org/.

tems in terms of diagnosis of diseases, the *FREDS* system establishes conceptual relationships or mappings between microscopic images content and textual information crawled from clinical records [14]. The main idea is to extract complementary information from exams that describe cell components similar to those identified in the microscopic image evaluated by a pathologist. The aim is to contribute with the reduction of the semantic gap between the computational retrieval of medical images and the human interpretation of their content. The *FREDS* system advocates that the semantic mapping can support the generation of knowledge.

The HSSF software framework has then emerged from the experience of developing the three aforementioned health surveillance systems. The orchestration of those systems as software components and their respective (required/provided) interfaces are summarized in Fig. 2.

The *CISS* component provides developers with classes supporting important functionalities: *Scientific Papers Searching*, *Medical Record Processing*, *Article Processing*, *Textual Processing*, *Natural Language Processing*, *Concept Recognition*, and *Similarity*. These functionalities are provided via *ClinicalRecordProcessing* to the *Automatic-SL* component so as it can classify patients according to surveillance level measures automatically computed. The *DocumentProcessing* is used by the *FREDS* component so as it can relate image reports to imaging exams on an automatic way.

The *FREDS* component offers classes to the *Image Feature Extraction*, the *Image Segmentation*, *Image Classification*, and the *Imaging Report Retrieval*. The *ImagingExamProcessing* provides these services. Finally, the *Automatic-SL* component serves a classification functionality to be reused by other applications via *SurveillanceLevelProcessing*.

III. The deployment of HSSF

In previous section, we described how HSSF was originally built by means of its fundamental software components which provide multiple services including clinical record processing, imaging exam processing, among others.

Throughout this section, we present how those components were reused to develop three new health surveillance applications as proofs of concept to the HSSF framework.

A. CISS+

As an evolution of the original *CISS* system, the first application developed by means of the reuse of HSSF components is called *CISS+* [10]. It augments the semantics of terms and concepts of scientific papers and clinical records by means of the use of the UMLS metathesaurus [20] and the MetaMap tool [21]. Experiments with UMLS and MetaMap demonstrated the effectiveness of the concept recognition task with a reduction of roughly 90% of terms.

Also as novelty, the *CISS+* employs automated techniques of relevance feedback to refine queries. The *Similarity* class of the CISS component calculates similarity measures among scientific papers and clinical records and runs automatic relevance feedback using three approaches: (i) using meta-information from the Medical Subject Headings (MeSH)[5] of scientific papers from the PubMed database; (ii) considering the whole set of documents with identified concepts after n-grams processing of medical records as relevant documents; and (iii) considering meta-information from the "Publication Type" field of PubMed papers.

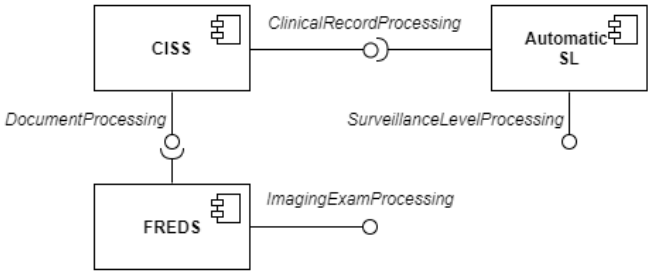


Fig. 2. The component diagram of HSSF.

Hence, the *CISS+* reuses mainly paper searching, textual processing, extended recognition of concepts, and mechanisms of query expansion in the *Similarity* and *Concept Recognition* classes of the CISS component.

B. CISS-SW

Using Semantic Web concepts and technologies, the *CISS-SW* is a search system that enables physicians to retrieve a scientific paper related to a patient's clinical record [17].

After the textual processing, the *CISS-SW* maps terms of papers into RDF triples[6] [22], stores them in a Triple Store[7], and composes a SPARQL [23] query by using the clinical records. With this query, the system retrieves form the triple store the papers related to the clinical records.

Therefore, the new functions of the *CISS-SW* include the processing and retrieval of scientific papers with Semantic Web support. In general, it reuses the classes related to the textual processing of clinical records provided by the HSSF.

C. QASF

A question-answering system returns short and direct answers to users. The *QASF* (Question Answering System in Chronic Diseases) application receives a question about chronic diseases and epigenetics information, and then looks for answers in collections of scientific papers [18] [19]. The aim is to help healthcare professionals to rapidly find focused related answers in the domain of chronic diseases.

The *QASF* architecture essentially consists of three modules: (i) *Question Processing*, (ii) *Answer Processing*, and (iii) *Document Processing* [18].

[5]The U.S. National Library of Medicine's hierarchically-organized terminology for indexing and cataloging of biomedical information.

[6]An RDF triple is a data entity composed of subject-predicate-object, like "John knows Steve" or "Steve is 42". RDF triples are the standard information exchange format in the Semantic Web.

[7]A triplestore is a purpose-built database for the storage and retrieval of RDF triples through semantic queries usually written in the SPARQL syntax.

The *Question Processing* module converts the question a user submits in natural language to a query that helps to search and select answers. This module applies pattern recognition and machine learning algorithms to handle the type and the content of the question, respectively. The *QASF* exploits linguistic and knowledge resources (e.g. WordNet and SNOMED) to support the processing of learning healthcare information.

As the only module reused from the HSSF architecture, the *Document Processing* module retrieves documents (e.g. scientific papers) that might have the answer to the user question. From each candidate document, this module also extracts excerpts that should be the answer.

Finally, the *Answer Processing* module processes all the potential right answers and classifies them according to a similarity value. Currently, the *QASF* considers the cosine between the user question and the candidate answer as the similarity value. Hence, the user is given the "n" first answers.

IV. Evaluation

Considering the software engineering literature, reusability is the degree to which an asset (e.g. a module or component) can be used in more than one software system, or in building other assets. In this section, we describe how we measured the reusability of the HSSF components in developing the systems described in the previous section.

Although the number of reused lines of code is a commonly used unit of measure, a software framework is more than lines of code. In the case of reuse, another well-known measure is the developer hour, which refers mainly to the time spent on searching, analysis and integration/modification. The time spent on these activities must be less than the time that is necessary to develop the artifact to be reused.

In order to evaluate the benefits of the HSSF framework, we requested each developer of the *CISS+* (one participant), *CISS-SW* (one participant) and *QASF* (two participants) applications to answer a questionnaire we elaborated, as illustrated in Fig. 3. The analysis of the corresponding answers was an attempt to understand the reuse of the HSSF components in terms of the time spent and the artifacts reused.

The *QASF* developers answered the questions together, then we have considered only one answer. We have disregarded the amount of time required to locate/search the HSSF as a tool to be reused because all the developers belong to the same group.

A. Questionnaire answers

Regarding the question 1, developers have considered classes, lines of code, and external packages as the software artifacts the most reused assets of the HSSF. From the answers to question 2, we identified that the most useful classes were the *Paper Search* and the *Paper Processing* classes, followed by *Natural Language Processing* and *Concept Recognition*. Regarding the question 3, despite analyzing from 20 to 25 classes, *QASF* developers reused from 1 to 4 classes without modification as well as from 1 to 4 classes with modifications. The *CISS+* developer analyzed between 15 and 19 classes and reused all of them without modifications. However, this person has actively participated in the creation of the HSSF. Finally, the *CISS-SW* developer analyzed from 5 to 9 classes and reused between 5 and 9 classes without modifications.

Considering the answers to question 4, as the *QASF* developers have not found proper documentation, they spent months to understand the model, to integrate the classes, and to understand and extend source code. The *CISS+* developer did not spend time on any of those activities because she is member of the *CISS* development team. Finally, the *CISS-SW* developer spent days reading the documentation and only weeks understanding the model and the classes, integrating these, and understanding and extending source code.

Another interesting result is the number of attempts that each developer made before reusing the HSSF components (question 5). Disregarding the *CISS+* developer, the other developers have only carried out between 1 and 3 attempts before the reuse. This small number of attempts is a good indicator of the HSSF reusability.

The last question of the questionnaire confirms this finding because the two developers have stated that they earned weeks and months regarding the development of their systems when they reused the HSSF. The developers of the *QASF* considered they did not earn much time, but we noticed they did not find any documentation (papers, reports, models, manuals, and help documentation) about the HSSF. Such documentation is available at the official HSSF website. *QASF* was recently created, hence the difficulty of reuse can be due to the need of updating many technologies exploited by the HSSF.

B. The average time spent on reusing HSSF classes

A look at the answers of the questionnaire has motivated us to find an ad hoc measure to quantify the time spent on the reuse of HSSF classes. To assess whether the reuse of classes yields a positive result besides the costs with the modification and integration of the reusable item of the HSSF in the current project, we have defined *the average time needed for classes reuse* as

$$T^{(a)} = \frac{T_c^{(u)}}{C} + \frac{T_c^{(m)}}{C_r} \tag{1}$$

in which:

- $T_c^{(u)}$ is the time associated with understanding of the classes;
- C is the number of classes analyzed;
- $T_c^{(m)}$ is the time spent integrating or extending the classes by source lines of code; and
- C_r is the total number of classes reused without modifications.

We ignored the risk of wasting time on the search for classes because the developers belong to the same team. Therefore, we only considered the questions 3.a, 3.b, 4.b, and 4.d (or 4.e) which are C, C_r, $T_c^{(u)}$, and $T_c^{(m)}$, respectively.

Questionnaire: An attempt to understand the HSSF reuse experience

1. What kind of artifacts did you reuse from CISS? (you can select more than one)

[ ] Models [ ] Classes
[ ] Tools [ ] Lines of code
[ ] Libraries [ ] API
[ ] External Packages [ ] Other:_____

2. If you **reused classes**, which components were more useful for your application? (you can select more than one)

[ ] Natural Language Processing [ ] Paper Processing
[ ] Paper Search [ ] Clinical Record Processing
[ ] Textual Processing [ ] Classification
[ ] Concept Recognition [ ] GUI
[ ] Similarity [ ] Other:_____

3. If you **reused classes**, please answer the following (a., b., c.) questions:

a. How many classes did you **analyze** to reuse from CISS?

[ ] 1 to 4 [ ] 10 to 14 [ ] 20 to 25
[ ] 5 to 9 [ ] 15 to 19 [ ] zero

b. How many classes (without modifications) did you **reuse** from CISS?

[ ] 1 to 4 [ ] 10 to 14 [ ] 20 to 25
[ ] 5 to 9 [ ] 15 to 19 [ ] zero

c. How many classes did you **extend/modify** from CISS to be reused?

[ ] 1 to 4 [ ] 10 to 14 [ ] 20 to 25
[ ] 5 to 9 [ ] 15 to 19 [ ] zero

4. How much **time** did you spend in the following activities to reuse CISS?

a. Reading documentation:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

b. Understanding/analyzing the model:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

c. Understanding/analyzing the classes:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

d. Integrating the classes:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

e. Understanding/analyzing lines of source code:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

f. Extending by coding:

[ ] hours [ ] months
[ ] days [ ] Other: _____
[ ] weeks

5. In average, how many **attempts** did you make **before** reusing an artifact from CISS?

[ ] 1 [ ] 4 to 5 [ ] More than 8
[ ] 2 to 3 [ ] 6 to 7 [ ] zero

6. Do you consider the reused artifacts **saved time** of the developing a new application?

[ ] Yes. How much ? Estimative:_____ hours. [ ] No.

Fig. 3. Questionnaire answered by developers who reused HSSF.

Using the answers given by the developers of the *CISS+*, *CISS-SW* and *QASF* applications, we have listed the results for this equation in Table I. This includes the average time (T^a) spent on reusing the classes of the HSSF by developers of the *CISS+*, *CISS-SW* and *QASF* applications. T^a, $T_c^{(u)}$ and $T_c^{(m)}$ are measured in terms of days. C and C_r, in turn, represent numbers of classes.

TABLE I
THE AVERAGE TIME (T^a) SPENT ON REUSING HSSF CLASSES

Developer	C	C_r	$T_c^{(u)}$	$T_c^{(m)}$	T^a
CISS+	19	19	7	6	**0.68**
CISS-SW	9	9	7	7	**1.55**
QASF	25	4	30	30	**8.7**

Table I also shows that the *CISS+* developer spent 0.68 days reusing each class of the HSSF. This was expected because she collaborated with the development of many HSSF classes before developing the *CISS+* application. The *CISS-SW* developer spent 1.55 days reusing each class of the HSSF, whereas the *QASF* developers spent much more time (8.7 days) than the other two developers. This was also expected because *QASF* developers did not find documentation about the HSSF and needed to update HSSF classes and packages.

Besides questions 3.a., 3.b., 4.b., and 4.d. (or 4.e.), some other quantitative answers to the questions can help to measure the HSSF reuse, but the reuse of the model, class, and source lines of code must be related because they consist of artifacts reused at different levels of design abstractions. In other words, our ad-hoc equation previously presented considers independent variables only.

In a near future, we intend to augment this initial effort including other time of reuse of other levels of design abstractions. Another situation to verify includes measures of the topology of the model, which includes connectivity, and between, among others, and should help us to infer coupling of classes, for example. Some literature works have presented result metrics in terms of reuse [24] [25].

V. ANALYSIS, LESSONS LEARNED AND RESULTS

When thinking about reuse, we must consider that the process of creating and updating of a *software framework* should never end. It is fundamental to reuse the framework while aiming to receive the developers' feedback on the development of new versions of the reused framework.

The creation of the HSSF framework has been a long process that has relied on collaborative work including ideas, software requirements, designs and development, augmentations, results and publications. This process started in 2007 with the initial development of the *Automatic-SL* system [12], which was followed by the creation of the *CISS* system [13] and *FREDS* [14]. In the end of 2014, we agreed on a version of the HSSF; and in 2015, we developed two new systems to validate the HSSF, the *CISS+* [16], and the *CISS-SW* [17] systems. Finally, in 2016, we included new classes in the HSSF after we developed the *QASF* system [18] [19].

The HSSF framework provides hot spots that are easy to manipulate such as:

- the insertion of other scientific information resources besides PubMed;
- the use of another domain ontology to create queries and to filter scientific papers from the desired information resources;
- the exploration of other ontologies and/or thesauri aiming at the recognition of medical and biomedical concepts besides UMLS;
- the manipulation of different types of clinical records or other documents in the healthcare domain; and
- the use of other classification and clustering techniques.

Given those hot spots, the less flexible one is the exploration of other ontologies and/or thesauri aiming at the recognition of medical and biomedical concepts besides UMLS. The HSSF carries a multilingual processing (mainly English and Portuguese), so it is necessary to apply a linguistic resource that can relate multilingual concepts and bring semantic relationships. In the healthcare domain, UMLS still represents the best option to recognize concepts, justifying the natural inflexibility. However, an extension of our framework to manipulate other linguistic resources besides UMLS has already been designed.

As frozen spots of the HSSF framework, we can mention abstract classes that allow each hot spot cited, for instance, an abstract class to illustrate ordinary attributes of different sets of documents and another abstract class to represent attributes of different document types. Other frozen spots consist of classes for textual processing (e.g. stopwords elimination and *n*-grams processing) as well as for the identification of relationships among documents.

In 2016, we presented the first version of HSSF as a software framework, which consisted of an architecture and a class diagram depicted elsewhere [10]. After we created the *QASF*, we were able to extrapolate and to update that diagram to integrate *Automatic-SL*, *CISS*, *FREDS* and *QASF*, as depicted in Fig. 4. By analyzing our own reuse experience, we have distilled some lessons learned and also classified these as advantages or drawbacks with the HSSF reuse.

In terms of **drawbacks** of the HSSF reuse, we can consider:

- the need to learn technical aspects of the programming languages, packages and ontologies to generate a steep learning curve that is necessary for the developers to know how the framework works before they can reuse it; and
- the cost in terms of demand of development expertise.

Considering these limitations, the HSSF faces problems in a specific abstraction domain. For years, some authors have advocate that the reuse of software frameworks degrade the performance/efficiency of the application and its security issues [26] [27]. However, HSSF users have not noticed any of these two points yet. Moreover, security is not an essential requirement of the *CISS+*, the *CISS-SW*, and the *QASF* applications.

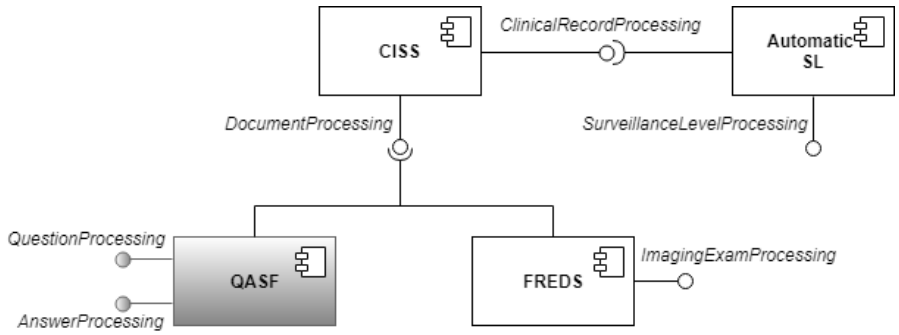


Fig. 4. The HSSF framework extended with two *QASF* services available to developers: *QuestionProcessing* and *AnswerProcessing*.

On the other side, the use of the HSSF offers the following **advantages**:

- the HSSF reduces the time and the energy spent on developing the *CISS+* and *CISS-SW* systems because the developers have the standard infrastructure and diagram of the HSSF providing the organization of modules and the classes of application;
- as a consequence from the previous advantage, developers can devote more time to requirements and user interfaces;
- the source codes of the three systems (*CISS+*, *CISS-SW* and *QASF*) are more organized and well documented because they follow the coding convention of the HSSF, which makes the source codes clean and easy to understand; and
- the use of the HSSF infrastructure affords three well-separated applications and defined business and logic layers from the user interface, making the code cleaner and extensible.

Finally, we believe that the HSSF improves the quality of the applications because the developers focus on the unique requirements of their application instead of spending time on infrastructure. Two of the users of the HSSF consider that their performance during the developing of their applications were improved reusing assets of the HSSF.

VI. Conclusion

The *Automatic-SL*, *CISS* and *FREDS* systems could be abstracted because they had common classes, purposes, and collaborations. For instance, the Relevance Feedback (RF-SL) classifier of the *Automatic-SL* system generates structured information from medical records and transforms it into bags of words; this system also eliminates stopwords and conducts stemming to produce a term-weight matrix. This matrix resembles the concept-weight matrix used by the *CISS* to compute similarity between scientific papers and medical records.

The main difference is that the matrix of the *CISS* uses UMLS concepts instead of simple terms. Nevertheless, the application recognition of concepts and the construction of the weight matrix according to these recognized concepts are perfectly applicable not only to the RF-SL module of the *Automatic-SL*, but also to other classifiers of the latter system. On the other hand, *FREDS* is composed of medical image processing classes that can also be related to textual processing.

Between 2007 and 2012, when the *Automatic-SL* and the CISS systems were created, the HSSF had its first classes designed for reuse with a focus on the processing and classification of healthcare-related information. Some years later, newcomers and specialists of our research group have used the HSSF, which has resulted in three new applications, the *CISS+*, the *CISS-SW* and the *QASF*. The developers of these systems have answered a questionnaire as an attempt to understand the reuse of the HSSF basically in terms of types of artifacts and time spent on reuse. The answers have allowed us to measure the average time that is necessary to reuse HSSF classes and to distill experiences such as lessons learned.

The HSSF is a software framework still under development. In a near future, it will be included a new set of class diagrams with more details for novel types of reuse, allowing much more general applicability. Besides this ongoing work, the results reported herein open new research directions that include:

- extension to the Chronic Disease Ontology (CDO) with knowledge obtained from scientific papers on epigenetics mechanisms and epigenetics risk factors for chronic diseases retrieved by the CISS;
- use of text entailment to map risk factors for chronic diseases;
- integration of new computational tools to map new concepts;
- modelling of an electronic medical record system coupled to *CISS* for use by the pediatric team; and
- investigation into the use of pediatric consensus by the HSSF.

By extending and reusing HSSF capabilities, our main goal is to allow healthcare applications developers to relate science research results, exams and treatments, which may be all incorporated into the clinical practice.

Acknowledgment

This work was supported by the São Paulo Research Foundation (FAPESP – Proc. 16/13206-4) and the National Council for Scientific and Technological Development (CNPq - Proc. 302031/2016-2 and 442533/2016-0).

REFERENCES

[1] D. Roberts and R. Johnson, "Evolving Frameworks: A Pattern Language for Developing Object-Oriented Frameworks," in *Pro. Conf. Pattern Languages and Programming*, vol. 3, 1996. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.8767

[2] B. Mamlin, P. Biondich, B. Wolfe, H. Fraser, D. Jazayeri, C. Allen, J. Miranda, and T. W, "Cooking up an open source emr for developing countries: Openmrs - a recipe for successful collaboration," in *AMIA Annual Symposium Proceedings*, 2006, pp. 529–533.

[3] L. P. M. P. Thompson A, Castle E, "Experience implementing OpenMRS to support maternal and reproductive health in northern nigeria," *Stud Health Technol Inform*, vol. 160, pp. 332–336, 2010.

[4] A. Falenski, M. Filter, C. T, A. A. Weiser, J.-F. Wigger, M. Davis, J. V. Douglas, S. Edlund, K. Hu, J. H. Kaufman, B. Appel, and A. K, "A generic open-source software framework supporting scenario simulations in bioterrorist crises," *Biosecurity and Bioterrorism: Biodefense Strategy*, September 2013.

[5] K. Kawamoto and D. F. Lobach, "Proposal for fulfilling strategic objectives of the u.s. roadmap for national action on decision support through a service-oriented architecture leveraging hl7 services," *J Am Med Inform Assoc*, vol. 14, no. 2, pp. 146–155, Mar-Apr 2007.

[6] A. C. M. T. G. de Oliveira and F. d. L. dos Santos Nunes, "Building a open source framework for virtual medical training," *Journal of Digital Imaging*, vol. 23, pp. 706–720, 2010.

[7] F. Faure, C. Duriez, H. Delingette, J. Allard, B. Gilles, S. Marchesseau, H. Talbot, H. Courtecuisse, G. Bousquet, I. Peterlik, and S. Cotin, "SOFA: A Multi-Model Framework for Interactive Physical Simulation," in *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, ser. Studies in Mechanobiology, Tissue Engineering and Biomaterials, Y. Payan, Ed. Springer, Jun. 2012, vol. 11, pp. 283–321. [Online]. Available: https://hal.inria.fr/hal-00681539

[8] H. Talbot, N. Haouchine, I. Peterlik, J. Dequidt, C. Duriez, H. Delingette, and S. Cotin, "Surgery Training, Planning and Guidance Using the SOFA Framework," in *Eurographics*, Zurich, Switzerland, May 2015. [Online]. Available: https://hal.inria.fr/hal-01160297

[9] E. T. Dougherty and J. C. Turner, "An object-oriented framework for versatile finite element based simulations of neurostimulation," *Journal of Computational Medicine*, vol. 2016, p. 15p., 2016.

[10] A. A. Macedo, J. Polettini, J. A. Baranauskas, and J. Chaves, "A health surveillance software framework to design the delivery of information on preventive healthcare strategies," *Journal of Biomedical Informatics*, vol. 62, pp. 159–170, August 2016.

[11] L. Jiang and C. C. Yang, "User recommendation in healthcare social media by assessing user similarity in heterogeneous network," *Artificial Intelligence in Medicine*, vol. 81, pp. 63 – 77, 2017, artificial Intelligence in Medicine AIME 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365717301185

[12] J. T. Pollettini, S. R. G. Panico, J. C. Daneluzzi, R. Tinós, J. A. Baranauskas, and A. A. Macedo, "Using Machine Learning Classifiers to Assist Healthcare-Related Decisions: Classification of Electronic Patient Records," *Journal of Medical Systems*, vol. 36, no. 6, pp. 3861–3874, 2012.

[13] J. T. Pollettini, J. A. Baranauskas, E. S. Ruiz, M. da Graça Pimentel, and A. A. Macedo, "Surveillance for the prevention of chronic diseases through information association." *BMC medical genomics*, vol. 7, no. 1, p. 7, Jan. 2014. [Online]. Available: http://www.biomedcentral.com/1755-8794/7/7

[14] H. C. Pessotti, L. O. M. Junior, E. G. Soares, and A. A. Macedo, "FREDS: Framework para redução da descontinuidade semântica em imagens médicas," in *Proceedings of the 11th Workshop on Medical Informatics - CSBC 2011*, 2011, pp. 1782–1791.

[15] D. J. P. Barker, "Fetal and infant origins of adult disease," *Monatsschrift Kinderheilkunde*, vol. 149, no. 13, pp. S2–S6, Jun 2001.

[16] A. A. Macedo, J. T. Pollettini, and E. V. Munson, "A chronic illness system using biomedical knowledge sources and relevance feedback," in *IEEE International Sympposium on Computer-Based Medical Systems*, 2015, pp. 244–249.

[17] J. Chaves, J. Pollettini, and A. Macedo, "Relating biomedical information using information mapping supported by semantic web," in *Proceedings of the 15th World Congress on Health and Biomedical Informatics*, ser. MEDINFO 2015, 2015, p. 1p.

[18] L. F. Almansa and A. A. Macedo, "Sistema de informação para perguntas e respostas em doenças crônicas," in *Proceedings of the 16h Medical Informatics Workshop - CSBC 2016*, Porto Alegre/RS - Brazil, July 2016, p. 10p.

[19] L. F. Almansa, G. Rubio, J. A. Baranauskas, and A. A. Macedo, "A question-answering architecture for surveillance information systems on chronic diseases based on knowledge from linguistic resources," *Submitted to Knowledge and Information Systems (KAIS)*, p. 25p., Feb 2018.

[20] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. Database-Issue, pp. 267–270, 2004. [Online]. Available: https://doi.org/10.1093/nar/gkh061

[21] *UMLS Reference Manual [Internet]*, National Library of Medicine (US)., 1999.

[22] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 concepts and abstract syntax," W3C, W3C Recommendation, Feb. 2014, http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[23] A. Seaborne and S. Harris, "SPARQL 1.1 query language," W3C, W3C Recommendation, Mar. 2013, http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[24] H. Koziolek, "Performance evaluation of component-based software systems: A survey," *Perform. Eval.*, vol. 67, no. 8, pp. 634–658, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2009.07.007

[25] M. Gupta, Chetna; Rathi, "A meta level data mining approach to predict software reusability," in *International Journal of Information Engineering and Electronic Business*, vol. 5, no. 6, Hong Kong, 2013, pp. 33–39.

[26] M. Fayad and D. C. Schmidt, "Object-oriented application frameworks," *Communications of the ACM, Special Issue on Object-Oriented Application Frameworks*, vol. 40, no. 10, oct 1997.

[27] A. A. Al-Baity, K. Faisal, and M. Ahmed, "Software reuse: the state of art," in *Proceedings of the International Conference on Software Engineering Research and Practice (SERP)*, Athens, 2013, pp. 1–7.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1053–1062
DOI: 10.15439/2018F51
ISSN 2300-5963 ACSIS, Vol. 15

Deep Object Comparison for Interface-based Regression Testing of Software Components

Tomas Potuzak
Department of Computer Science and Engineering/ NTIS – New Technologies for the Information Society, European Center of Excellence, Faculty of Applied Sciences, University of West Bohemia
Univerzitni 8, 30614 Plzen, Czech Republic
Email: tpotuzak@kiv.zcu.cz

Richard Lipka
NTIS – New Technologies for the Information Society/Department of Computer Science and Engineering, European Center of Excellence, Faculty of Applied Sciences, University of West Bohemia
Univerzitni 8, 30614 Plzen, Czech Republic
Email: lipka@kiv.zcu.cz

***Abstract*—In this paper, we describe the deep object comparison (DOC) algorithm, which is used for comparison of general objects in Java programming language based on their internal structures and values of primitive attributes. The DOC algorithm was designed to be utilized in our interface-based regression testing of software components, which enables to uncover subtle changes of the behavior of a component-based application under test with a newly installed version of a software component in comparison to its behavior with an old version of this component.**

I. Introduction

THE component-based software development is a part of software engineering for nearly two decades. It utilizes isolated reusable software parts called *software components*, which provide and/or require functionalities called *services*. The services are accessible using public interfaces of the components and the components are expected to interact solely using these interfaces. Specific details depend on the utilized *component model*, which defines the behavior, features, and interactions of software components and is implemented by a *component framework* [1].

Regardless the utilized component model, a common situation using the component-based software development is that a component can be used in different applications and an application consists of multiple components, which can originate by different manufacturers [1]. This underlines the necessity for the testing, not only of the individual components, but of the entire component-based application as well.

Additionally, many components exist in several versions, which can mutually differ by the internal behavior (i.e., there are different computations), by the external behavior (i.e., different interactions with other components), or by the public interface (i.e., different required and/or provided services). In theory, the change of the component's internal behavior should not affect its external behavior and therefore should not affect the behavior of the entire component-based application. Nevertheless, in reality, an unwanted error can be introduced into the new version of the component, a side effect of a method invocation can be added or removed, a computation can be prolonged leading to a time-out to expire, and so on. So, when installing a new version of a component to a functional component-based application, adequate regression testing is desirable even when there are no apparent external changes of the new version of the component in comparison to the old version [2].

During our previous research, we developed an approach for interface-based regression testing of software components, whose source code is not available (e.g., third-party software components). The approach is tailored for the situation when there is a new version of a component installed in a component-based application and we want to check if it exhibits the same behavior within the application as its old version [2].

The experimental implementation of the approach, which was described in [2] in detail, was designed for the OSGi [3] component model for Java programming language, but the ideas behind it can be used for other component models and programming languages as well. The overall process starts with the analysis of the services and their methods of the software components of the entire component-based applica-tion under test. For each method of each service of each component, a set of invocations is generated. Then, the invocations are performed in an iterative phase. In each iteration, all invocations are performed and their conse-quences are being observed and stored. New consequences of the same invocations can emerge, because the inner states of the components can change between iterations due to the invocation of the methods. Besides the consequences, new invocations can emerge during this phase as consequences of different invocations. Both the new consequences and new invocations are stored only if they are not already stored. This requires comparison of the already stored items with the newly created items. The process stops when no new consequences are created. The result is a testing scenario

This work was supported by Ministry of Education, Youth and Sports of the Czech Republic, project PUNTIS (LO1506) under the program NPU I and by European structural and investment funds (ESIF), project CZ.02.1.01/0.0/0.0/17_048/0007267.

with actions (i.e., invocations) and their consequences, which can be saved to a file. If this process is performed prior and after the installation of a new version of a component to the component-based application under test, the detailed comparison of these two saved scenarios can uncover changes in the behavior of the application caused by the new version of the component [2].

In our experimental implementation, we used the standard `equals()` method for the comparison of the objects associated with the consequences and the invocations (e.g., return values, values of the parameters). Although this can work in many cases, it cannot be used universally. Some objects do not implement the `equals()` method, which leaves them with the default implementation corresponding to the identity (only the same objects are considered equal). Even if the `equals()` method is implemented in the object, it can be implemented incorrectly, as pointed out in [4], [5], or [6]. And even if it is implemented correctly, it does not mean that it considers all primitive values of the object recursively [7]. So, it is possible that, although the `equals()` method returns `true` for a pair of objects, their internal structures can be different and/or contain some different primitive values. These subtle differences can be important for our approach, since they could mean different behavior, which we want to detect.

In order to mitigate this problem, in this paper, we describe the deep object comparison, which will replace the utilization of the `equals()` method. The deep object comparison enables to compare two objects based on the "shape" of their internal structures and all the corresponding primitive values. So, no changes in the objects are missed. This deep object comparison is suited to be used both during the generation of the scenario and during the comparison of two scenarios of our interface-based approach for regression testing of software components. The algorithm was implemented within our Interface Analysis Tool (InAnT). The deep object comparison was first tested as a stand-alone algorithm, before it will be incorporated into our approach. The description of the deep object comparison algorithm along with the description of the performed tests is the main contribution of this paper.

The paper is structured as follows. The interface-based regression testing of software components is briefly described in Section II. Related work is discussed in Section III. In Section IV, the deep object comparison is described in detail. The performed tests and results are described in Section V and the paper is concluded and the future work is discussed in Section VI.

II. Interface-based Regression Testing of Components

As it was mentioned in Section I, the interface-based regression testing of software components is designed to uncover any changes in a component-based application's behavior after the installation of a new version of a component [2]. The changes are detected during comparison of the testing scenario generated and stored from the application with the old version of the component and the scenario generated and stored from the application with the new version of the component [2].

A. Generation of the Testing Scenario

Our approach assumes that the entire component-based application is under test, because the components within it interact with each other. Their interactions are observed during the generation of the scenario in order to uncover the behavior of the particular components [2].

First step in the generation of the testing scenario is the determination of all methods of all services of the components of the application under test. This can be done by any method capable to retrieve complete method signature. We use standard OSGi methods and Java reflection [8] for this purpose in our experimental implementation [2]. The components, their services, and their method are inserted into a tree data structure, which forms the basis of the testing scenario (see Fig. 1a).

For each method of this structure, an initial set of invocations is generated and added into the structure. Each invocation contains a unique combination of values for all the parameters of the method.

The invocations are then successively performed (i.e., the methods are invoked with the parameters stored in the invocations in the tree data structure) in the iterative phase, one at a time, and the consequences of each invocation are observed (i.e., what happened when the method was invoked). The possible consequences are a thrown exception, a return value, a value change in "out" parameters of the method, a subsequent invocation of a service method of ano-

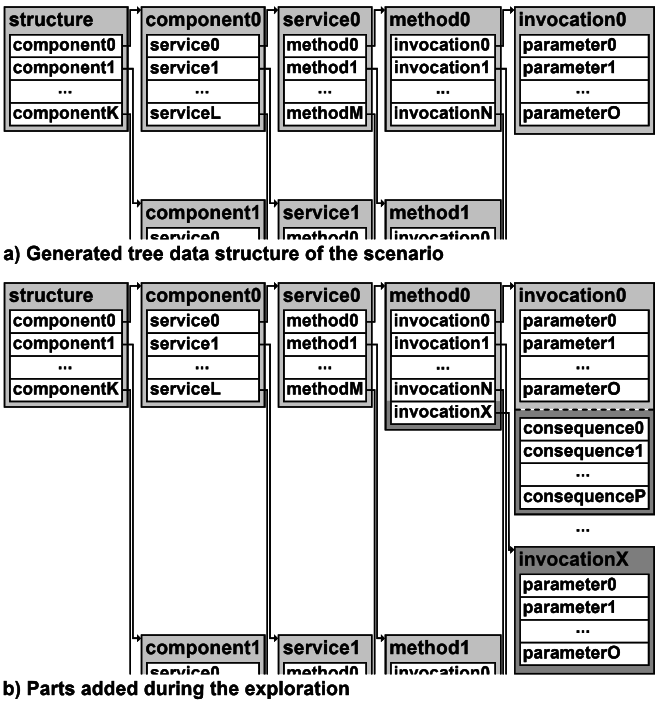


Fig. 1 The tree structure of the scenario

ther component, and a change of the inner state of the component. The last consequence differs from the others, since it is not easily observable from outside. So, it is not considered by our approach. There can be several consequences per method invocation. All the observed consequences are added to the tree data structure to the invocation, which caused them (see Fig. 1b), but only if they are not already present. Each type of consequence contains its type and type-dependent data (e.g., the return value, the instance of an exception, the changed value of an "out" parameter, etc.) [2]. Based on the type and the data, the consequences can be mutually compared, which is necessary for the determination, whether a new consequence is already present or not.

The most important consequences are the subsequent invocations. Each subsequent invocation is defined by the method, which it is invoking, and by the unique combination of its parameter values. When this consequence is observed, it is added to the tree data structure (if not already present) similarly to other types of consequences. Moreover, the invocation that the consequence represents is added to the invocations of the corresponding method into the tree data structure (again, only if not already present). These invocations are valuable, since their parameter values are genuine, originating in the internal logic of the component, which invoked the method [2].

The invocations contained in the tree data structure are performed several times in the iterative phase in order to exploit the subsequent invocations. The subsequent invocations generated in nth iteration can be performed in $(n + 1)$th iteration and their consequences can be thus observed. The iterative phase is stopped when no new consequences are generated in the current iteration. At this point, the testing scenario represented by the tree data structure is complete (see Fig. 1b). All invocations and consequences contain a number representing the iteration, in which they were added to the structure (starting with 1). The initial invocations created prior the iterative phase have this number set to 0. The generated scenario is saved to a XML file [2].

B. Comparison of the Testing Scenario

When a new version of a component is installed into the component-based application under test, the process described in Section II.A is repeated and a new scenario is created. The saved scenario is then loaded from the XML file and both tree data structures are compared. The comparison is performed on each level of the structures, starting from the component level [2].

On each level, it is checked, whether there are corresponding items (i.e., components, services, methods, invocations, consequences) in both tree data structures. If so, their subtree is expanded and the comparison continues on the lower level. If not so, the difference (item is missing in one or second tree data structure) is reported, this item is not expanded and its lower levels are not considered further [2]. The example of the result of the comparison is depicted in Fig. 2.

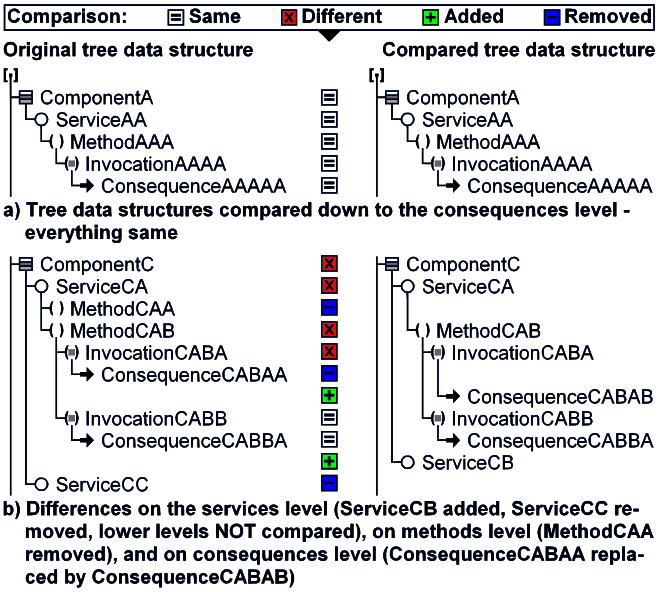


Fig. 2 Result of the comparison of two scenarios (tree data structures)

The most important differences are on the invocations and invocation consequences levels. These differences mean different behavior of the application under test with the old and the new version of the component. Differences on the methods or services levels imply that there are changes in the public interface of the component. Our approach of course detects these changes, but, unlike the changes in the behavior, these changes can be detected by other means as well, such as advanced static analysis methods (e.g., see [9]) [2].

C. Object Comparison Issues

Both during the generation of the scenario and during the comparison of two scenarios, we need to compare general objects, which is problematic.

During the generation of the scenario, the comparison of general objects is necessary in the iterative phase when new consequences and invocations are generated. They are added to the tree data structure only if they are not already contained, requiring their comparison to other consequences and invocations. It should be noted that due to the tree nature of the data structure, a newly generated consequence is compared only to the consequences of the corresponding invocation. Similarly, a newly generated invocation is compared only to the invocations of the corresponding method. So, the number of comparison is limited, it is not necessary to compare the consequence or invocation to all consequences or invocations.

Nevertheless, the comparison of two consequences lies in the comparison of their type and, if the type is the same, in the comparison of the type-related data. If the type of both compared consequences is the return value, then the associated return values are compared. The comparison is

trivial if the return values are of primitive types, but ambiguous if they are objects. In the experimental implementation of our approach, we use the standard `equals()` method for objects, which are not `null`. This can work in many cases, but cannot be used universally.

For example, some objects do not implement the `equals()` method, which leaves them with the default implementation corresponding to the identity (only the same objects are considered equal). An example of such object is in Fig. 3a. The `Point3D` class represents a point in space, but does not override the `equals()` method, leaving it with its default implementation (from the `Object` class). When a method returns a new instance of the `Point3D` in every invocation (see Fig. 3b), the comparison of this instance to another instance using `equals()` will always return `false`, even with the same values of their corresponding coordinates (see Fig. 3c). If an invocation of the method depicted in Fig. 3b were performed repeatedly during the iterative phase, its return value consequence would always seem different, because the return values would not be identical (based on the `equals()` method), although they would contain the same values of their corresponding coordinates. So, each newly generated consequence would be added to the tree data structure in each iteration. The iterative phase would not stop until an out-of-memory exception would occur. This problem can be mitigated (not solved) by introduction of the maximal number of iterations, but it is clear that this is not the intended behavior.

Moreover, there are further issues. Even if the `equals()` method is implemented in the object, it can be implemented

```
class Point3D {
  public int x;
  public int y;
  public int z;

  public Point3D(int x, int y, int z) {
    this.x = x;
    this.y = y;
    this.z = z;
  }
}
```

a) A class representing a point in space without overridden `equals()` method

```
...
public Point3D asPoint(int x, int y, int z) {
  return new Point3D(x, y, z);
}
...
```

b) A method returning a new instance of the `Point3D` class in every invocation

```
...
Point3D p1 = asPoint(1, 2, 3);
Point3D p2 = asPoint(1, 2, 3);
boolean comparison = p1 == p2; //comparison false
...
```

c) Two results of the method with the same primitive values compared

Fig. 3 Example of a class without overridden `equals()` method and of following problems

incorrectly, as pointed out in [4], [5], or [6]. And even if it is implemented correctly, it does not mean that it considers all primitive values of the object recursively [7]. So, it is possible that, although the `equals()` method returns `true` for a pair of objects, their internal structures can be different and/or contain some different primitive values. These subtle differences would not be detected using the `equals()` method.

The described problem is not limited to the return value consequences. The same problem is with the comparison of consequences representing a change in the "out" parameters of a method and with the comparison of invocations. Each invocation contains the combination of parameter values of a method and these values, which can be general objects, are compared during the comparison of the invocations.

During the comparison of two scenarios, the comparison of general objects is needed for the comparison of invocations and consequences, similarly to the generation of the scenario. Comparison of methods and higher levels of the tree data structure are based on data types and names, not general objects. The problem with the comparison of general objects is more pronounced here, though. The reason is that at least one of the compared scenarios is loaded from a XML file. In order to utilize the `equals()` method for the comparison, it would be necessary to recreate all the objects during the loading of the scenario. This would necessitate full-scale serialization of general objects during the saving of the scenario to the file. Hence, in the experimental implementation of our approach, the general objects contained in the invocations and consequences were compared only based on their classes and `null` values.

More specifically, the information stored to the XML file for an object was its real class or `null`. Hence, during the comparison of two scenarios, it was only checked, whether both compared objects are `null` or whether both compared objects are of the same class. In these two cases, the objects were considered equal. The exception was the instances of the `String` class, which were compared using their content. The reason is that the `String` instances can be easily saved and loaded to/from a file. It is clear that this significantly reduces abilities of our interface-based regression testing of software components. The entire approach works correctly and is able to detect changes in behavior of the application under test (see [2]). However, many subtle differences in the compared scenarios can remain hidden, because the information is lost during the saving of the scenario to a file. So, some changes in behavior of the application under test could remain undetected.

In order to solve all the described problems, we designed the deep object comparison (DOC) described in Section IV in detail. The DOC will be incorporated into our interface-based regression testing of software components where it will replace the `equals()` method during the generation of the scenario and the class-based comparison during the comparison of two scenarios.

III. Related Work

The issue of object equality in object-oriented languages is discussed in scientific literature mainly in relation to memory optimization and to object equality implementation. Both branches are discussed in following subsections. Although none of these branches is related to software testing, the algorithms described in the discussed papers solve problems similar to our deep object comparison.

We focused mainly on the papers regarding the Java programming language, since our current implementation is written in this language. However, the principles can be used in similar languages (e.g., C#) as well.

A. Memory Optimization

There are several papers focused on the optimization of memory management in languages, which utilize garbage collection (i.e., automatic disposal of objects, which are no longer in use by the program). The main idea behind these works is that there are a number of equivalent objects in the memory during the execution of a program, which can be replaced by a single instance while preserving the same behavior of the entire program (see [7], [10], [11], [12], etc.).

In [7], a tool enabling detection of equivalent objects in a Java application, which can be replaced by a single instance, is described. The investigated application is instrumented and, during its execution, all relevant heap activity is recorded. After the execution, the post-mortem analysis is performed. The objects of the application are separated into equivalence classes. Each equivalence class can be replaced by a single instance. No automatic optimization is performed. The tool only uncovers and reports the sites (i.e., positions in source code) of the program with the potential for an optimization by replacing more equivalent objects with a single instance [7]. In order to determine, whether an object can be replaced by another object without affecting the behavior of the application, it is necessary to compare these objects thoroughly. It is pointed out that a full comparison requires checking of two potentially cyclic labeled graphs for isomorphism. As a large number of comparisons is required in this approach, a hash value is calculated for each object, which is then used for faster comparison of the objects [7].

Similar approach is described in [10], although it is intended for a different programming language (Pharo). Again, the relevant heap activity is recorded during a run of the application under test and the post-mortem analysis is performed after the run. The main difference is that, in [10], the objects are divided into several types and only certain types of objects, which are susceptible to redundancy (e.g., instances of the `String` or `Point` classes), are considered as optimization opportunities. Directed vertex and edge labeled graph is used for the representation of the internal structure of objects for the comparison purposes. Based on it, a hash value is calculated for each considered object to speed up its comparison to other objects [10].

In [11], the conditions, which an object must satisfy to be considered for caching, are discussed. The paper is focused on immutability of objects, which is in some form often required by caching and similar techniques (including techniques described in [7] and [10]). It is pointed out that the border between construction and mutation of an object is not always clear [11].

In [13], an advanced method for the reduction of the duplication of strings is described. The comparison of objects is straightforward in this case, as the compared objects are instances of the `String` class.

In [12], a similar yet different technique to [7] and [10] is described. Similarly to the [7] and [10], the technique searches the sites (i.e., positions in source code) in an application, which have the potential to be optimized by reusing existing objects or data structures [12]. Unlike [7] and [10], the technique is focused not only on finding equivalent objects, but on finding reusable instances as well. The idea is that it is not necessary to create a new (possibly internally complex) instance when there is another instance of the same class available, which is no longer in use. In that case, it is only necessary to change settings of its primitive values, but it is not necessary to create new object and its entire internal structure. Moreover, since the unused object is reused, its garbage collection is saved. Thus, the techniques tend to utilize objects, which are relatively short-lived. So, unlike [7], [10], which are focused mainly on the reduction of the memory consumption, [12] is focused also on the reduction of computation time. For the representation of the internal structure of the objects, modified balanced parenthesis algorithm is used [12].

In contrast to [12], the technique described in [14] is focused on the long-lived objects, which are candidates for caching. For the speedup of the object comparison, a form of hash value called “fingerprint” of the object is used. In standard classes from the `java.lang` package, the `equals()` method is used [14].

In [15], the performance of the hash-consing (i.e., utilization of a global cache of objects) is discussed. The comparison of objects is based on the `equals()` method even when this method does not consider all attributes of the object. The authors also define weak immutability of an object based only on the values of the attributes, which are considered by the `equals()` method. Again, a hash value is used for the description of the objects [15].

B. Object Equality Implementation

The other group of scientific papers regarding the object equality is focused on the implementation of the `equals()` method. In majority of this papers, the required features of the `equals()` method are cited – the *reflexivity*, *symmetricity*, and *transitivity* [4]. In some works, it is pointed out that many textbooks contain flawed implementations of the `equals()` method (e.g., [4], [6]).

In [4], a checker of the `equals()` methods is described. The checker is focused on the features of the `equals()` methods under test and reports any violation of these features. However, it is not focused on comparison of the entire internal structure of complex objects [4]. In [6], the right design of the `equals()` method using design patterns is discussed.

A generator of the `equals()` methods for complex objects is described in [16]. In this work, the objects are compared on per-field basis. There are several "depths" of equality defined. Depth-0 corresponds to the referential equality, meaning that the objects are equal if both are the same object (corresponding to the comparison operator "=="). Depth-1 (shallow equality) means that, for all corresponding fields of two objects, the referential equality holds. The deep equality then means that, for all corresponding fields of two objects, the deep equality holds [16].

IV. Deep Object Comparison

As it was mentioned in Section II.C, the deep object comparison (DOC) is designed for our interface-based regression testing of software components. It will replace the comparison of general objects using the `equals()` method during the generation of the testing scenario and the comparison of general objects using their classes and `null` values during the comparison of two scenarios. Hence, the DOC algorithm has two phases – the forming of the graph representation of the object and the comparison of two graphs of the compared objects. The main advantage of this approach is that the graph representation of the object can be easily saved to and loaded from a file. So, no information will be lost during the saving of the scenario.

A. Object Equality

As arises from Section III, there are various views on the equality of two objects. Nevertheless, since our goal is to uncover any difference of the compared objects, however subtle, we have to adopt the view of the deep equality described in [16]. Similar definitions are described also in [7], [10], or [11]. Nevertheless, all the definitions of the equality described in these works are recursive. Since the DOC algorithm creates a graph representing the internal structure of the object, our definition is based on this graph. It is not recursive, but expresses similar conditions as the definition of the deep equality described in [16].

Suppose there are two compared objects `a` and `b` and the graph representations of their internal structures were created (see Section IV.B). Objects `a` and `b` are considered equal if and only if they are of the same class, their graph representations are isomorphic (i.e., have the same "shape"), and all values of the corresponding primitive attributes in the corresponding vertices of the graphs are of the same type and equal. Fig. 4 shows several examples of pairs of objects, which are considered equal or different.

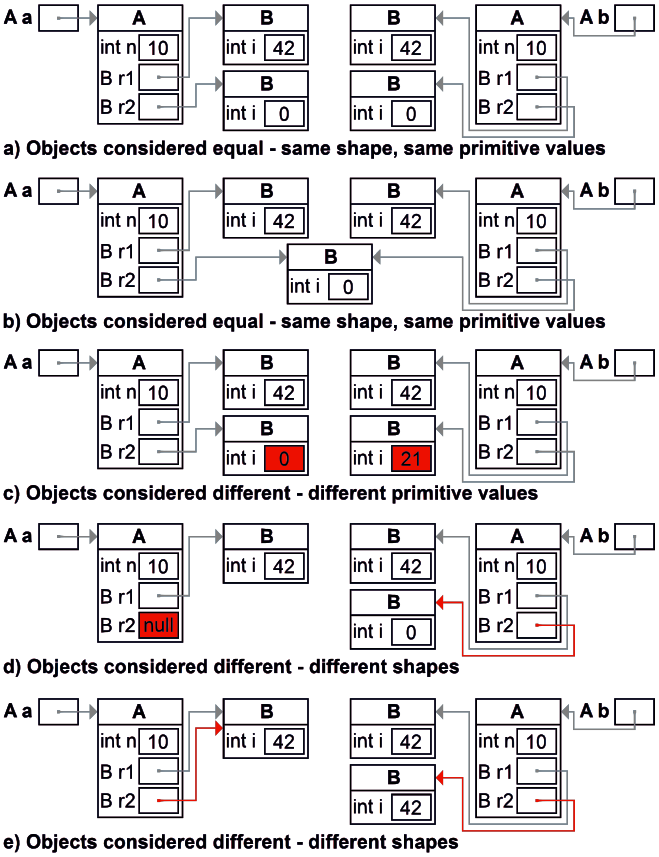


Fig. 4 Examples of objects considered equal and different

B. Forming of the Graph of an Object

When we compare two objects, it is checked whether both objects are not `null` and whether they are of the same class. If not so, the objects are considered different. If so, the DOC algorithm is used. Its first step is the creation of the graph representations of both compared objects. The representation is a directed vertex- and edge-labeled graph.

Each vertex of the graph corresponds to a single object of the internal structure of the object that we want to compare. The starting vertex of the graph corresponds to the object that we want to compare. Each vertex incorporates the ID, the state that is used during the comparison of two graphs (see Section IV.C) and the reference to the object that this vertex represents. It also incorporates a list containing types, names, and values of all primitive attributes of the object that this vertex represents.

Each directed edge of the graph represents a reference attribute of the object and points to the vertex, representing the object, to which the reference attribute is pointing. Each edge incorporates the type (i.e., the class) and the name of the reference attribute it represents.

The reference attributes that are arrays are treated specifically. If the attribute is an array of primitive values, each value is considered a primitive attribute with the name created from the name of the array attribute and the index of the value. If the attribute is an array of references, each value is considered a reference attribute, which means that it is

represented as an edge in the graph. The name is again constructed from the name of the array attribute and the index of the value. Similar approach is used for multi-dimensional arrays. As these arrays are represented as an array of arrays in Java, each inner dimension of the array is treated as an object (i.e., it is represented by a vertex in the graph) with attributes corresponding to the values on particular indices. Since these attributes do not have names, their indices are used instead. The reference attributes, which are `null`, and array attributes pointing to empty arrays are treated as primitive attributes.

The forming of the graph is based on the breadth-first search (BFS) algorithm starting in the object that we want to compare. There are a queue and a list of all vertices of the graph, which are empty at the start. First vertex is created from the object that we want to compare and is inserted to the list and to the queue. The forming of the graph then continues while the queue is not empty. First vertex is removed from the queue (current vertex) and all attributes of the object that is represented by this vertex are determined using reflection [8]. All primitive attributes are added to the list in this vertex. For each reference attribute, a new vertex is created. If this vertex is not in the list of all vertices, this new vertex is added to the queue and an edge is formed from the current vertex to the newly created vertex. If it is already present in the list, it is not added to the queue and a new edge is formed from the current vertex to the vertex from the list.

The presence of the vertex in the list is determined using the sequence searching and the comparison of objects that the vertices represent using the comparison operator "==". This way, it is ensured that already visited objects (e.g., due to cyclic references) are not visited again.

```
allVertices = List();
queue = Queue();
vertex = Vertex(comparedObject);
allVertices.add(vertex);
queue.add(vertex);
while (!queue.isEmpty()) {
  vertex = queue.remove();
  vertex.primitives=getPrimitives(vertex.object);
  references = getReferences(vertex.object);
  for (r: references) {
    neighbor = Vertex(r.object);
    index = allVertices.indexOf(neighbor);
    edge = Edge(r);
    if (index >= 0) {
      edge.vertex = allVertices[index];
    }
    else {
      edge.vertex = neighbor;
      queue.add(neighbor);
      allVertices.add(neighbor);
    }
    vertex.addEdge(edge);
  }
}
for (i = 0; i < allVertices.length; i++) {
  allVertices[i].ID = i;
}
```

Fig. 5 Pseudocode for the forming of the graph representation of an object

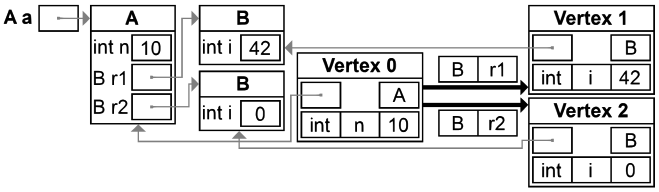


Fig. 6 The resulting graph (on the right) with references to the object (on the left)

When the queue is empty, the graph is fully formed. The last step is the assignment of the IDs to all vertices based on their indices in the list. These IDs are used during the saving of the graph to a file. The algorithm is described in Fig. 5 in pseudocode. The resulting graph for the object `a` from Fig. 4a is depicted in Fig. 6.

C. Comparison of Graphs

Two graphs created as described in Section IV.B can be easily compared without the references to the original objects, from which they were formed. The structure of each graph is represented by its vertices and edges, which also incorporate all necessary information – the values of primitive attributes and the names and types of all attributes. This is important, because the graph can be saved to a file and then loaded from this file and it is not necessary to recreate the original object, from which the graph was formed.

The graphs are compared using their parallel BFS exploration. Basically, in one loop, both graphs are explored.

```
v1 = graph1.rootVertex; v1.state = GRAY;
v2 = graph2.rootVertex; v2.state = GRAY;
queue1 = Queue();
queue2 = Queue();
queue1.add(v1);
queue2.add(v2);
equal = true;
while (!queue1.isEmpty()) {
  v1 = queue1.remove();
  v2 = queue2.remove();
  if (v2 == null) {
    equal = false;
    break;
  }
  if (!compare(v1.primitives, v2.primitives)) {
    equal = false;
    break;
  }
  for (e: v1.edges) {
    if (e.vertex.state == WHITE) {
      queue1.add(e.vertex);
      e.vertex.state = GRAY;
    }
  }
  for (e: v2.edges) {
    if (e.vertex.state == WHITE) {
      queue2.add(e.vertex);
      e.vertex.state = GRAY;
    }
  }
  v1.state = BLACK;
  v2.state = BLACK;
}
if (!queue2.isEmtpy())
  equal = false;
```

Fig. 7 Pseudocode for the comparison of two graphs

For each node, it is checked, whether they have the same count of primitive attributes with the same values. If a difference is found, the loop is ended prematurely and the objects are considered different. The objects are also considered different, when the graph of one of the objects is fully explored and the other is not. The algorithm is described in Fig. 7 in pseudocode.

V. Validation and Results

The described DOC algorithm was thoroughly tested using two sets of tests. In the first set, we focused on the correct functionality of the algorithm. The second set of tests was focused on the performance of the algorithm. All tests were performed on a standard notebook computer with dual-core Intel i5-6200U at 2.30 GHz with HyperThreading, 8 GB of RAM and 250GB SSD/500GB HDD. The software environment consisted of the Windows 7 SP1 (64 bit), Java 1.6 (32 bit), and Equinox OSGi framework.

A. Correct Functionality of the DOC Algorithm

The correct functionality of the DOC algorithm was tested by comparison of pairs of similar or equal objects. There were 5 pairs with variously complicated internal structures. The structures of all objects were created manually using the `A` class depicted in Fig. 8. The class and the objects were designed to test various situations, which can occur in internal structures of general objects – primitive attributes, reference attributes, arrays, and lists. Similarly, the changes introduced into one object of each pair of equal objects in order to create a similar but slightly different object represent various differences, which can occur in general objects – different lengths of an array, different values of an primitive attribute, different references, different elements of an array and/or a list. The internal structures of the objects were reasonably small (see Fig. 9) in order to enable controlled manual introduction of the changes and checking whether the results of the DOC algorithm are correct.

For each pair of objects, four tests were performed. In two tests, both objects of the pair were identical. In the remaining

```
import java.util.List;

public class A {
  private A parent;
  private int number;
  private String string;
  private List<A> list;
  private A[] array;

  public A(A parent, int number, String string,
      List<A> list, A[] array) {
    this.parent = parent;
    this.number = number;
    this.string = string;
    this.list = list;
    this.array = array;
  }
}
```

Fig. 8 The class `A` used for the testing of the functionality of the DOC algorithm

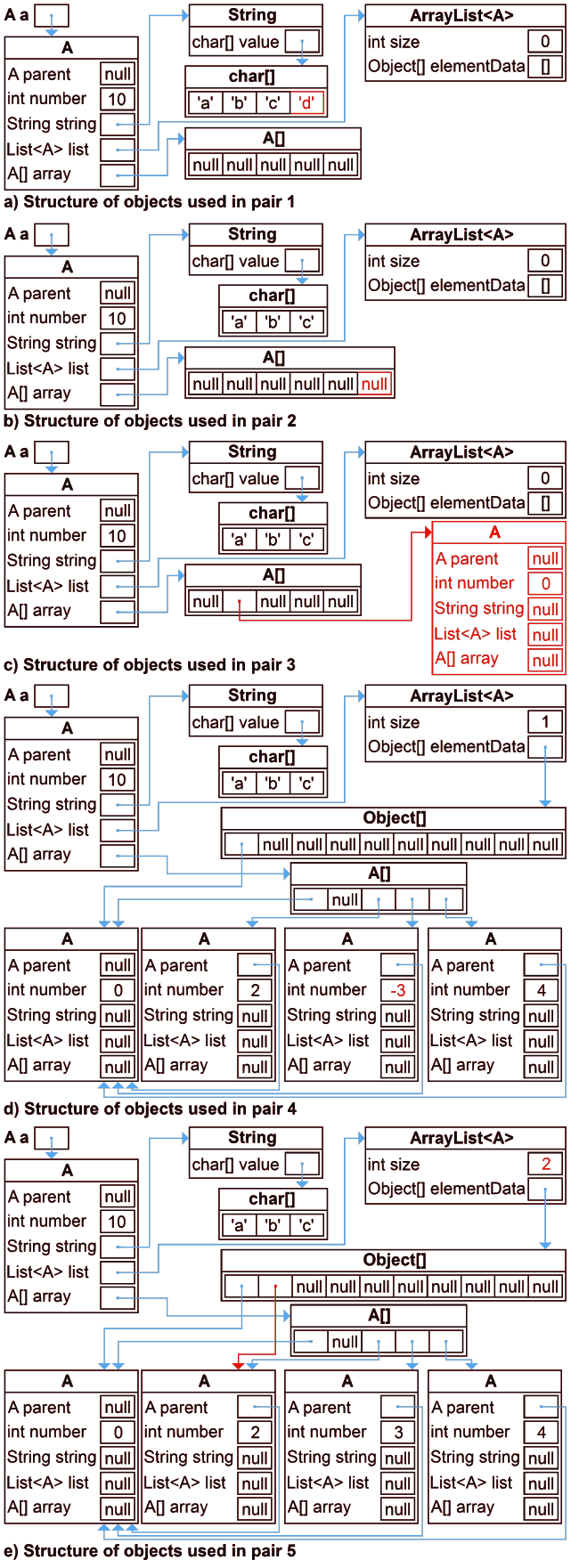


Fig. 9 Structures of objects used for the testing of the functionality of the DOC algorithm with highlighted differences

TABLE I THE RESULT OF THE COMPARISON OF PAIRS OF OBJECTS USING THE DOC ALGORITHM

Pair of objects	Both objects in memory		One object saved & loaded	
	Equal	Different	Equal	Different
1	`true`	`false`	`true`	`false`
2	`true`	`false`	`true`	`false`
3	`true`	`false`	`true`	`false`
4	`true`	`false`	`true`	`false`
5	`true`	`false`	`true`	`false`

two tests, one object was similar but slightly different. The differences are highlighted in Fig. 9. This way, both required results of the comparison (i.e., `true` and `false`) were tested. The graphs were created for both objects of the pair and then compared (once for identical objects and once for similar objects). Moreover, one of the graphs was saved to and loaded from a file and the comparison was performed again (once for identical objects and once for similar objects). This way, it was tested that the saving of the graph to a file will not affect the comparison.

For pair 1, the objects were instances with empty lists, arrays with 5 `null` elements, `null` parents, but with set values of numbers and strings. The difference introduced into one object was one character longer string (see Fig. 9a). For pair 2, the objects were the same, only the difference was that one array was one element longer (see Fig. 9b). For pair 3, the objects were again the same and the difference was that one array has one element (with index equal to 1) not `null`. Instead, it pointed to another instance of the `A` class with all attributes set to `null` or 0 (see Fig. 9c).

For pairs 4 and 5, the arrays had 5 elements, one element (with index equal to 1) was `null` and the remaining elements were instances of the `A` class with `null` strings, arrays and lists, but with parents set to the zeroth element of the array (with the exception of the zeroth element, which had the parent set to `null`) and the number set to the index in the array. The list contained one element corresponding to the zeroth element of the array. For pair 4, the difference was a different number value in the fourth element of the array (see Fig. 9d). For pair 5, the difference was one added element to the list (second element of the array – see Fig. 9e).

The results of the testing are summed in Table I. The algorithm returned the expected value (`true` or `false`) in every instance.

B. Performance of the DOC Algorithm

The performance of the DOC algorithm is important, because there is a significant number of object comparisons during both the generation of the testing scenarios and the comparison of two scenarios. Nevertheless, the number of comparisons is still far lower than it is necessary for the caching and similar techniques described for example in [7], [10], or [12]. For the testing, increasingly complex objects created using the `A` class were generated and compared. The compared objects were always equal. Different objects would cause premature ending of the comparison, because it is stopped on first uncovered difference (see Fig. 7). The results are summed in Table II.

TABLE II THE PERFORMANCE OF THE DOC ALGORITHM

Vertices count	Edges count	Graphs construction time [ms]	Graph comparison time [ms]
33	42	1.9	0.1
333	442	7.6	0.7
3 333	4 442	110.7	4.6
33 333	44 442	10 131.2	28.8

As can be seen in Table II, the graph construction phase is far more time-demanding than the graph comparison phase. There are several reasons for this. In the graph construction phase, the graphs of both the compared objects are created sequentially, which means that roughly half of the time is necessary to create the graph of a single object. More importantly, the sequential searching of the list of already visited objects is employed during this phase (see Fig. 5). Lastly, Java reflection is used during this phase, which is inherently slow [8].

The graph construction is not needed during the comparison of two testing scenarios, where the graphs are already constructed. However, it is necessary during the generation of the testing scenario for the comparison of invocations and consequences. Nevertheless, although these comparisons are numerous, the graph must be constructed only once per object. Multiple comparisons can then be performed using the graph representations of the objects only. So, the construction of the graph representations of the objects have a significant performance benefit in addition to the ability to easily save the graph representation to a file.

VI. CONCLUSION AND FUTURE WORK

In this work, we described the deep object comparison (DOC) algorithm. The algorithm was designed for our interface-based regression testing of software components for the comparison of general objects based on their internal structures and values of primitive attributes. Based on the performed tests, the algorithm works correctly for arbitrary objects, which can incorporate primitive and reference attributes, lists, and/or (single- or multi-dimensional) arrays. The performance of the algorithm seems to be reasonable, since the slower graph construction phase is expected to be performed by an order of magnitude less often than the faster graph comparison phase.

In our future work, we will incorporate the DOC algorithm to our interface-based regression testing of software components. We will then perform thorough testing focused on the resulting performance of the entire approach as well as on its increased ability to detect subtle changes in behavior of the component-based applications with new versions of components.

We will also consider the use of the DOC algorithm in another project focused on the generation of complex testing data (i.e., objects with complex internal structures). We plan

to employ particle swarm optimization [17] and/or combinatorial testing [18] to minimize the amount of generated data. The DOC algorithm can be utilized for the comparison of generated objects and/or their parts to remove any duplicity.

REFERENCES

[1] C. Szyperski, D. Gruntz, and S. Murer, Component Software – Beyond Object-Oriented Programming, ACM Press, New York, 2000.

[2] T. Potuzak, R. Lipka, and P. Brada, "Interface-based Semi-automated Testing of Software Components," Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, Prague, September 2017, pp. 1335-1344, http://dx.doi.org/10.15439/2017F139

[3] The OSGi Alliance, OSGi Service Platform Core Specification, release 4, version 4.2, 2009.

[4] C. R. Rupakheti and D. Hou, "An Abstraction-Oriented, Path-Based Approach for Analyzing Object Equality in Java," 2010 17th Working Conference on Reverse Engineering, Beverly, October 2010, pp. 205-214, http://dx.doi.org/10.1109/WCRE.2010.30

[5] C. R. Rupakheti and D. Hou, "EQ: Checking the Implementation of Equality in Java," 2011 27th IEEE International Conference on Software Maintenance (ICSM), Williamsburg, September 2011, pp. 590-593, http://dx.doi.org/10.1109/ICSM.2011.6080837

[6] D. E. Stevenson and A. T. Phillips, "Implementing Object Equivalence in Java Using the Template Method Design Pattern," Proceedings of the 34th SIGCSE technical symposium on Computer science education, Reno, January 2003, pp. 278-282, http://dx.doi.org/10.1145/611892.611987

[7] D. Marinov and R. O'Callahan, "Object Equality Profiling," Proceedings of the 18th annual ACM SIGPLAN conference on Object-oriented programing, systems, languages, and applications, Anaheim, October 2003, pp. 313-325, http://dx.doi.org/10.1145/949305.949333

[8] I. R. Forman, N. Forman, Java Reflection in Action, Manning Publications, 2004.

[9] K. Jezek, L. Holy, A. Slezacek, and P. Brada, "Software Components Compatibility Verification Based on Static Byte-Code Analysis," 39th Euromicro Conference Series on Software Engineering and Advanced Applications, Santander, September 2013, pp. 145-152, http://dx.doi.org/10.1109/SEAA.2013.58

[10] A. Infante, A. Bergel, "Object Equivalence: Revisiting Object Equality Profiling (An Experience Report)," Proceedings of the 13th ACM SIGPLAN International Symposium on Dynamic Languages, Vancouver, October 2017, pp. 27-38, http://dx.doi.org/10.1145/3170472.3133844

[11] A. Infante, "Identifying Caching Opportunities, Effortlessly," Companion Proceedings of the 36th International Conference on Software Engineering, Hyderabad, May 2014, pp. 730-732, http://dx.doi.org/10.1145/2591062.2591198

[12] G. Xu, "Finding Reusable Data Structures," Proceedings of the ACM international conference on Object oriented programming systems languages and applications, Tuscon, October 2012, http://dx.doi.org/10.1145/2398857.2384690

[13] K. Nasartschuk, M. Dombrowski, K. B. Kent, A. Micic, D. Henshall, and C. Gracie, "String Deduplication During Garbage Collection in Virtual Machines," Proceedings of the 26th Annual International Conference on Computer Science and Software Engineering, October 2016, pp. 250-256

[14] G. M. Rama and R. Komondoor, "A Dynamic Analysis to Support Object-Sharing Code Refactorings,", Proceedings of the 29th ACM/IEEE international conference on Automated software engineering, Vasteras, September 2014, pp. 713-723, http://dx.doi.org/10.1145/2642937.2642992

[15] M. J. Steindorfer and J. J. Vinju, "Performance Modeling of Maximal Sharing," Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering, Delft, March 2016, http://dx.doi.org/10.1145/2851553.2851566

[16] N. Grech, J. Rathke, and B. Fischer, "JEqualityGen: Generating Equality and Hashing Methods," Proceedings of the ninth international conference on Generative programming and component engineering, Eindhoven, October 2010, pp. 177-186, http://dx.doi.org/10.1145/1942788.1868320

[17] B. S. Ahmed, L. M. Gambardella, W. Afzal, and K. Z. Zamli, "Handling constraints in combinatorial interaction testing in the presence of multi objective particle swarm and multithreading," Information and Software Technology, vol. 86, pp. 20–36, 2017 http://dx.doi.org/10.1016/j.infsof.2017.02.004

[18] M. Bures and B. S. Ahmed, "On the effectiveness of combinatorial interaction testing: A case study," 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), July 2017, pp. 69–76, http://dx.doi.org/10.1109/QRS-C.2017.20

5th Doctoral Symposium on Recent Advances in Information Technology

THE aim of this meeting is to provide a platform for exchange of ideas between early-stage researchers, in Computer Science and Information Systems, PhD students in particular. Furthermore, the symposium will provide all participants an opportunity to get feedback on their studies from experienced members of the IT research community invited to chair all DS-RAIT thematic sessions. Therefore, submission of research proposals with limited preliminary results is strongly encouraged.

Besides receiving specific advice for their contributions all participants will be invited to attend plenary lectures on conducting high-quality research studies, excellence in scientific writing and issues related to intellectual property in IT research. Authors of the two most outstanding submissions will have a possibility to present their papers in a form of short plenary lecture.

TOPICS

- Automatic Control and Robotics
- Bioinformatics
- Cloud, GPU and Parallel Computing
- Cognitive Science
- Computer Networks
- Computational Intelligence
- Cryptography
- Data Mining and Data Visualization
- Database Management Systems
- Expert Systems
- Image Processing and Computer Animation
- Information Theory
- Machine Learning
- Natural Language Processing
- Numerical Analysis
- Operating Systems
- Pattern Recognition
- Scientific Computing
- Software Engineering

EVENT CHAIRS

- **Kowalski, Piotr Andrzej,** Systems Research Institute, Polish Academy of Sciences; AGH University of Science and Technology, Poland
- **Lukasik, Szymon,** Systems Research Institute, Polish Academy of Sciences, AGH University of Science and Technology, Poland

PROGRAM COMMITTEE

- **Arabas, Jaroslaw,** Warsaw University of Technology, Poland
- **Atanassov, Krassimir T.,** Bulgarian Academy of Sciences, Bulgaria
- **Balazs, Krisztian,** Budapest University of Technology and Economics, Hungary
- **Bronselaer, Antoon,** Department of Telecommunications and Information at Ghent University, Belgium
- **Castrillon-Santana, Modesto,** University of Las Palmas de Gran Canaria, Spain
- **Charytanowicz, Malgorzata,** Catholic University of Lublin, Poland
- **Corpetti, Thomas,** University of Rennes, France
- **Courty, Nicolas,** University of Bretagne Sud, France
- **De Tré, Guy,** Faculty of Engineering and Architecture at Ghent University, Belgium
- **Fonseca, José Manuel,** UNINOVA, Portugal
- **Fournier-Viger, Philippe,** University of Moncton, Canada
- **Gil, David,** University of Alicante, Spain
- **Herrera Viedma, Enrique,** University of Granada, Spain
- **Hu, Bao-Gang,** Institute of Automation, Chinese Academy of Sciences, China
- **Koczy, Laszlo,** Szechenyi Istvan University, Hungary
- **Kokosinski, Zbigniew,** Cracow University of Technology, Poland
- **Krawiec, Krzysztof,** Poznan University of Technology, Poland
- **Kulczycki, Piotr,** Systems Research Institute, Polish Academy of Sciences, Poland
- **Kusy, Maciej,** Rzeszow University of Technology, Poland
- **Lilik, Ferenc,** Szechenyi Istvan University, Hungary
- **Lovassy, Rita,** Obuda University, Hungary
- **Malecki, Piotr,** Institute of Nuclear Physics PAN, Poland
- **Mesiar, Radko,** Slovak University of Technology, Slovakia
- **Mora, André Damas,** UNINOVA, Portugal
- **Noguera i Clofent, Carles,** Institute of Information Theory and Automation (UTIA), Academy of Sciences of the Czech Republic, Czech Republic
- **Pamin, Jerzy,** Institute for Computational Civil Engineering, Cracow University of Technology, Poland
- **Petrik, Milan,** Czech University of Life Sciences Prague, Faculty of Engineering, Department of Mathematics, Czech Republic
- **Ribeiro, Rita A.,** UNINOVA, Portugal

- **Sachenko, Anatoly,** Ternopil State Economic University, Ukraine
- **Samotyy, Volodymyr,** Lviv State University of Life Safety, Ukraine
- **Szafran, Bartlomiej,** Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Poland
- **Tormasi, Alex,** Szechenyi Istvan University, Hungary
- **Wei, Wei,** School of Computer science and engineering, Xi'an University of Technology, China
- **Wysocki, Marian,** Rzeszow University of Technology, Poland
- **Yang, Yujiu,** Tsinghua University, China
- **Zadrozny, Slawomir,** Systems Research Institute, Poland
- **Zajac, Mieczyslaw,** Cracow University of Technology, Poland

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1065–1068
DOI: 10.15439/2018F247
ISSN 2300-5963 ACSIS, Vol. 15

ECG Signal Analysis for Troponin Level Assessment and Coronary Artery Disease Detection: the NEEDED Study 2014

Dominika Długosz, Aleksandra Królak
Łódź University of Technology,
Faculty of Electrical, Electronic, Computer,
and Control Engineering, Institute of Electronics,
ul. Wólczańska 211/215, 90-924 Łódź, Poland,
Email: dominika.a.m.dlugosz@gmail.com,
aleksandra.krolak@p.lodz.pl

Trygve Eftestøl, Stein Ørn, Tomasz Wiktorski,
Kay Raymond Jenssen Oskal, Martin Nygård
University of Stavanger,
Faculty of Science and Technology,
Department of Electrical and Computer Engineering,
4036 Stavanger, Norway,
Email: {trygve.eftestol, stein.orn, tomasz.wiktorski}@uis.no,
{martin.a.nygard, kay.oskal}@gmail.com

***Abstract*—Physical exercise is widely recognized as beneficial to the cardiovascular system. However, intense exercise may also carry fatal risk. Investigation of this phenomenon is one of the primary purposes of the North Sea Race Endurance Exercise Study (NEEDED). This paper describes analysis of electrocardiograms (ECG) and heart rate signals collected from amateur athletes, participants of the race, to facilitate non-invasive estimation of the level of cardiac troponin I (cardiovascular risk biomarker) and detection of coronary artery disease (CAD). It was demonstrated that the combination of ECG and heart rate parameters can predict CAD with high specificity (up to 98%) and relatively good sensitivity. Moreover, while troponin level assessment is unlikely to be reliably performed using regression techniques, it might be possible using a new, probabilistic classification-based model. Further evaluation of the latter requires the use of additional data, which is one of possible directions for the future work.**

I. Introduction

CARDIAC troponins T and I (cTnT, cTnI) are protein subunits involved in contraction of the heart muscle. Their increased blood levels are widely associated with an occurrence of damage to the myocardium of diverse etiologies, including coronary artery disease (CAD) [1], [2], [3]. Nonetheless, recent studies have reported increase in cTnT and cTnI levels incident to prolonged, high-intensity physical exercise in presumably healthy individuals, predominantly recreational athletes [4]-[9].

This phenomenon was observed also in the frame of the North Sea Race Endurance Exercise Study (NEEDED) conducted at the University of Stavanger and Stavanger University Hospital in Norway [10], [11]. Aiming to explore the impact of long-term physical effort on the physiology of the cardiovascular system, the study recruited its participants from recreational cyclists competing in Nordsjørittet (the North Sea Race) - an annual cycling competition, organized in Rogaland, Norway. In 2014, over a thousand study subjects were examined i.a. for blood levels of cardiovascular biomarkers (including cTnI) and electrocardiograms (ECG). Supplementary data was retrieved from some of the participants' sports watches. Herein described research work was concerned with analysis of the abovementioned data. The main hypothesis stated that blood level of cTnI and presence of CAD can be predicted based on parameters of ECG and heart rate (HR). In particular, the prediction might be guided by physical effort-induced changes in not explicitly pathological ECG.

II. Study population and the dataset

The investigated population comprised 160 presumably healthy individuals. A total of 53 individuals were assessed by coronary computed tomography angiography, 6 of whom were diagnosed with CAD. In these 6 and further 14 cases, at least one cTnI level was elevated (>190 ng/l [10]). The analyzed data included:

- blood levels of cTnI, measured at: 24 h before, 3 h after, and 24 h after the race;
- 10-second 12-lead ECG recordings, collected at the same measuring time points as the cTnI data;
- clinical data: participants' age and BMI;
- formerly processed [12] HR data from sports watches.

III. Data preparation

A. ECG preprocessing and segmentation

The ECG was preprocessed by: filtering (band-pass Butterworth infinite impulse response filter, order: 5, cut-off frequencies: 1 and 40 Hz) and baseline offset removal. Subsequently, the signal was segmented into heartbeat templates using tools implemented in BioSSPy toolbox. Key points of the ECG (vertices, onsets, and endpoints of P, S, and T waves) were determined based on averaged beat templates from lead I similarly to our previous work [13]. Onsets of positive waves (P, T) were searched for within a window preceding the maximal ascending slope of the wave (n_{ms}): $[n_{ms}–w,\ n_{ms}–s]$, with w - window size, s - additional spacing to compensate for a possible 'M pattern'. Onset candidates were points $n \in [n_{ms}-w,\ n_{ms}-s]$ satisfying conditions:

$$N_c = \{n:\ |y'(n)| \leqslant |k \cdot y'(n_{ms})| \wedge y''(n) \geqslant 0\} \quad (1)$$

TABLE I: Parameters of detection of M and W patterns.

Wave	Pattern	Lead	k	l
P	M	I, II	0.2	3
R	M and W	V1, V5, V6	0.2	2
S	W	I, V6	0.3	2

where: y - the signal; k - a threshold factor. Selection criterion, depending on point type, was of form:

1) $arg \min_{n \in N_c} g(n),\ g(n) = |y(n) \cdot y'(n)|$, or
2) $arg \max_{n \in N_c} y''(n)$

Detection P and T waves endpoints was performed analogically on the descending slopes of the waves. For S wave (negative wave) endpoint, the respective segment was inverted. The parameters: w, s, and k were tuned separately for each point type based on standard durations of the waveforms.

IV. Feature extraction

A. ECG signal

ECG features described below were defined in three areas: time domain, frequency domain, and correlation analysis.

1) Features from lead-I signal measurements: basic time-domain features, including:

- heart rate,
- QT interval duration corrected for HR (according to the formula of Fridericia [14]),
- ST segment duration,
- ST elevation,
- P wave shape coefficient, i.e. the ratio of width and amplitude of the wave.

2) M and W patterns: notches in P, R, and S waves[1](see Fig. 1). An M pattern was deemed to be present in the signal if a sufficient number l of samples n of given wave satisfied the following criteria:

$$\begin{cases} y'(n) < 0 \ \wedge \ y(n) > k \cdot y(n_{max}) & for\ n_{on} < n < n_{max} \\ y'(n) > 0 \wedge y(n) > k \cdot y(n_{max}) & for\ n_{max} < n \leqslant n_{end} \end{cases} \tag{2}$$

where: n_{on}, n_{peak}, n_{end} - indices of onset, peak, and endpoint of the given waveform; y - signal values; k - threshold factor: minimal portion of the peak amplitude. For detection of W patterns, the signal was inverted. Parameters k and l were determined empirically for each wave type, as summarized in table I.

3) Heart axis: direction[2] estimated using a pair of perpendicular leads: I and aVF:

$$\theta = arctan\left(\frac{V_{net,aVF}}{V_{net,I}}\right) \tag{3}$$

[1]Depending on wave and lead of their manifestation, M and W patterns might be symptoms of atrial hypertrophy, left or right bundle branch block, and other conditions [15].

[2]The mean direction of the electric field vector (in the coronal plane) throughout an ECG cycle. Its significant deviation may be a symptom of disorders affecting the conduction system (e.g. bundle branch block) [16]

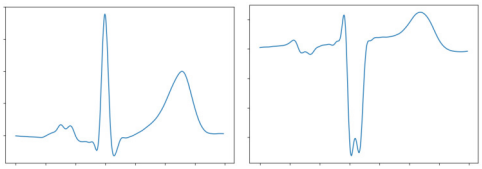


Fig. 1: Exemplary averaged beat templates: (left) M pattern in a lead-I P wave; (right) W pattern in a lead-V1 R wave.

V_{net} denotes net QRS potential in a lead, calculated as:

$$V_{net} = \begin{cases} max(V_{QRS}) & if\ min(V_{QRS}) \geqslant 0 \\ max(V_{QRS}) - |min(V_{QRS})| & otherwise \end{cases} \tag{4}$$

where $V_{QRS}(t)$ - voltage of the QRS complex [17].

4) Lead-I QRS templates correlation: used to assess internal morphological consistency of a recording. The parameter was calculated as the mean value of the upper-triangle elements (excluding the diagonal) of correlation matrix of QRS templates from a single lead-I signal.

5) Frequency-domain features: derived from power spectra of the QRS complexes. QRS-only signal was extracted by a mask: QRS selection windows (rectangular with half-Gaussian slopes), replicated at locations of the R peaks. Parameters derived from power spectrum of the masked signal included:

- two slopes characteristic for log-transformed low-frequency moiety of the signal, calculated over frequency ranges: [3 Hz, 8 Hz] and [15 Hz, 19 Hz] (as in [18]),
- the mean of power spectral density signal over frequency range [150 Hz, 300 Hz] (upper half of the range).

B. Heart rate

HR signal from sports watches, resampled and processed in the previous works, was a base for calculation of two parameters:

1) HRp99: mean of 99th percentile HR samples:

$$HR_{p99} = \frac{\overline{\{hr \in HR \ : \ hr \geqslant P_{99}(HR)\}}}{HR_{max}} \tag{5}$$

where: HR_{max} - maximal predicted age-dependent HR: $HR_{max} = 208 - 0.7 \cdot Age$ [19].

2) HR90 time: a portion of the race time in which participant's HR was above 90% their individual HR reserve:

$$t_{HR90} = \frac{\sum_{i=1}^{n} i\ [HR_i > (HR_{rest} + 0.9 \cdot HR_{reserve})]}{\sum_{i=1}^{n} i} \tag{6}$$

where: n - the number of samples, HR_{rest} - resting HR (from ECG prior to the race), and $HR_{reserve}$ - individual HR reserve: $HR_{reserve} = HR_{max} - HR_{rest}$ [4].

C. Clinical data and parameters from the previous work

- participants' age and BMI,
- max and mean HR from Tinghaug hill segment (a major steep climb of the race, associated with substantial effort and strain to the heart) [12], normalized by HR_{max}.

Following the hypothesis on the importance of changes versus momentary state, parameters from days 2 and 3 were

expressed as ratio (for HR) or difference with respect to day 1. To sum up, there were: 14 parameters from the ECG (each day), 4 from the HR, and 2 from the clinical data - in total, 48 features. A general assumption was to use data gathered prior to, or simultaneously with collection of blood for troponin assay. Thus, there were three sets of input parameters for cTnI level assessment, with 16, 34, and 48 features. CAD detection was performed on two sets: one with the 48 features and the second, additionally including cTnI levels (51 features).

V. Experiments setup

The data analysis had two main objectives: cTnI level estimation (separately for each day) and CAD prediction - together, four problems. The small size and strongly nonuniform distribution of the data motivated the use of leave-one-out cross-validation (LOOCV) approach.

A. CAD detection

Two methods were applied for this binary classification task:

1) Automatically optimized classification: a search for the best classifier launched with the use of TPOT - genetic programming-based software providing tools for model selection and tuning [20]. The number of generations and population size were set to 7 and 70 respectively.

2) Grid-search-optimized decision tree classifier: the classifier was optimized in terms of i.a. class weights and maximal depth. The best estimator instance was passed to LOOCV loop.

The results were evaluated using two types of metrics: confusion matrix and the area under the receiver operating characteristic (ROC) curve, weighed by counts of true positive (TP) and true negative (TN) observations.

B. cTnI level estimation

Due to a strong bias towards the lowest values and vast dispersion of the highest readings, the cTnI values were log-transformed before analysis. Estimation of the cTnI level - as a continuous variable - was approached using two techniques:

1) Automatically optimized regression approach: determination of an optimal regression model with the TPOT software (7 generations and population size of 70).

2) Probabilistic classification-based approach: an intermediate class definition was obtained by stratification of the log-transformed cTnI levels for a given day into 2 to 10 layers of equal breadth. Test samples were classified using a logistic regressor with probabilistic classification output. Next, in the LOOCV scheme, the cTnI level for each test sample was estimated by computing a weighted average of the class centers (medians, $\mathbf{c_{cTnI}}$) with class probabilities vector $\mathbf{p}$ as weights:

$$cTnI \;=\; 10^{(\mathbf{p^T}\,\cdot\,\mathbf{c_{cTnI}})} \tag{7}$$

VI. Results and discussion

A. CAD detection

The results of CAD prediction are summarized in Table II. Results were influenced by a random component (randomly seeded decision tree classifier; obtaining a deterministic output is possible, but might introduce a bias to the results). The information on the true cTnI levels significantly influenced the results, in particular in terms of the false positive (FP) detections. When the cTnI data was included, all FP results were from the elevated cTnI group; otherwise, all FPs belonged to the low cTnI class.

TABLE II: Summary of the results of CAD prediction; TP = true positive, TN = true negative, FP = false positive, FN = false negative (given in counts of observations).

Dataset	Approach	Model specification	Confusion matrix	ROC area
Without cTnI data	TPOT optimization	gradient boost. classifier	TN 154, FP 0, FN 6, TP 0	0.42
	Decision tree + grid search	number of features: 5	TN: 152, FP: 2, FN: 3, TP: 3	0.74
With cTnI data	TPOT optimization	gradient boost. classifier	TN 154, FP 0, FN 6, TP 0	0.83
	Decision tree + grid search	number of features: 5	TN: 151, FP: 3, FN: 1, TP: 5	0.91

Formerly [12], CAD detection based on the HR and clinical data with the best true positive rate (TPR) of 0.86 (6/7 cases) and the same true negative rate (TNR). Herein described experiments were performed with a greater scope of parameters (from the ECG signal), but with reduced number of CAD-positive observations (by one sample, i.e. 14%). TPR decreased to 0.5 (3/6 cases). However, TNR was noticeably higher, reaching 0.98 (152/154 cases). With the cTnI data, TPR increased to 0.83 (a single undetected case). Most significant factor contributing to TPR decrease is the depletion of the positive class strength.

The results do not unambiguously prove that features derived from the ECG signal improve CAD prediction rate (compared to HR only). However, they clearly demonstrate the value of ECG in excluding otherwise suspected CAD in healthy individuals.

B. Troponin level determination

1) Regression approach: this approach failed to establish a reliable prediction model. In none of the cases, the scores reached positive values (while 1 denotes a perfect prediction, and 0 - a constant model). One of possible explanations lies in the physiology of the circulatory system: the differences in ECG and HR parameters between the individuals did not sufficiently correlate with differences in cTnI level over the range of the values. Indeed, it should not be assumed that e.g. the duration of the QT segment would increase proportionally to the increase in the extent of CAD, associated with elevation in cTnI. The negative conclusion was confirmed by an independent team at the University of Stavanger [21].

2) Probabilistic classification-based approach: exemplary results (for day 3) are presented in Fig. 2. The highest true value was excluded from the plot to provide a better plot scale for analysis of the results. The estimation was strongly biased towards the lowest level, containing majority of the observations. Additionally, there are some false high predictions. However, there is a group of observations actually

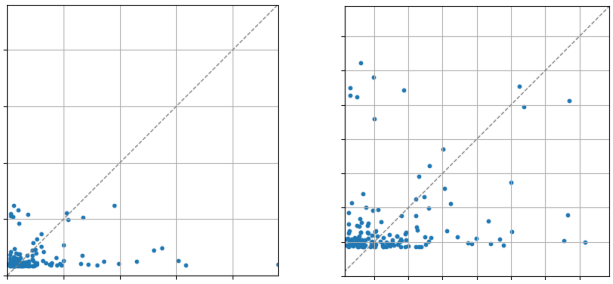


Fig. 2: cTnI level determination with probabilistic classification: results for day 3; horizontal: true, vertical: predicted values; (left) full set, (right) zoom at the lower range.

following the diagonal of the plot. Although they constituted minority of the set, these predictions were fairly accurate.

There were multiple factors hindering the prediction. The number of samples (160) was low. This effect was escalated by data imbalance. In day-1 data, 39% of the samples had the same cTnI level of 1.6 ng/l, and 86% did not exceed the level of 5 ng/l (mean $\pm$ standard deviation: 2.2 $\pm$ 0.8). In the remaining 14% of the dataset, the values spread between 4.9 and 284.8 ng/l (mean $\pm$ standard deviation: 26.9 $\pm$ 55.8). On the remaining days, there were additional substantial, but isolated peaks (5025.9 ng/l for day 2 and 7918.5 ng/l for day 3 - the latter being more than 33-fold larger than any other in the set). Excluding a single high-cTnI sample (in LOOCV) further amplified the imbalance.

VII. Conclusions

The herein reported project was concerned with analysis of data from the NEEDED research program with two main objectives. First of all, the level of circulating cTnI following prolonged, strenuous exercise was to be estimated using information derived from ECG and HR signals. The main conclusion to this problem was negative, though meaningful for the future research: the correlation between cTnI and parameters derived from HR and ECG was not found sufficient to establish a continuous, regression-based model for estimation of the level of the former. However, it is possible that the alternative classification-based approach is more likely to correspond to the underlying physiological mechanisms governing the exercise-induced cTnI response. Nevertheless, both conclusions need to be validated on a greater dataset.

The second goal was improvement of the rate of detection of CAD compared to previously achieved results. ECG was not found to increase the sensitivity of the prediction. However, this refers to a decrease in TP detections by 1 observation with lowered strength of the positive class. On the other hand, the ECG features notably enhanced the detection in terms of its specificity - from 86% to 98%. Best results of prediction of CAD could be achieved by including the cTnI level information into the analysis. The two types of information appear to be complementary - cTnI improves sensitivity of the model (a single undetected case), while ECG and HR data promote its specificity by effectively excluding non-CAD cases of cTnI elevation.

References

[1] J. Sarko and C. V. Pollack, "Cardiac troponins," *Am J Emerg Med*, vol. 23, no. 1, pp. 57–65, Jul. 2002.

[2] L. Babuin and A. S. Jaffe, "Troponin: the biomarker of choice for the detection of cardiac injury," *CMAJ*, vol. 173, no. 10, pp. 1191–1202, Nov. 2005.

[3] T. Omland, M. A. Pfeffer, S. D. Solomon, J. A. de Lemos, H. Røsjø *et al.*, "Prognostic Value of Cardiac Troponin I Measured With a Highly Sensitive Assay in Patients With Stable Coronary Artery Disease," *J Am Coll Cardiol*, vol. 61, no. 12, pp. 1240–1249, Mar. 2013.

[4] P. Aagaard, A. Sahlén, L. Bergfeldt, and F. Braunschweig, "Heart Rate and Its Variability in Response to Running—Associations with Troponin," *Med Sci Sports Exerc*, vol. 46, no. 8, p. 1624, Aug. 2014.

[5] R. Shave, A. Baggish, K. George, M. Wood, J. Scharhag *et al.*, "Exercise-Induced Cardiac Troponin Elevation: Evidence, Mechanisms, and Implications," *J Am Coll Cardiol*, vol. 56, no. 3, pp. 169–176, Jul. 2010.

[6] T. M. H. Eijsvogels, M. D. Hoogerwerf, M. F. H. Maessen, J. P. H. Seeger, K. P. George *et al.*, "Predictors of cardiac troponin release after a marathon," *J Sci Med Sport*, vol. 18, no. 1, pp. 88–92, Jan. 2015.

[7] A. Legaz-Arrese, K. George, L. E. Carranza-García, D. Munguía-Izquierdo, T. Moros-García *et al.*, "The impact of exercise intensity on the release of cardiac biomarkers in marathon runners," *Eur J Appl Physiol*, vol. 111, no. 12, pp. 2961–2967, Dec. 2011.

[8] S. Regwan, E. A. Hulten, S. Martinho, J. Slim, T. C. Villines *et al.*, "Marathon Running as a Cause of Troponin Elevation: A Systematic Review and Meta-Analysis," *J Interv Cardiol*, vol. 23, no. 5, pp. 443–450, Oct. 2010.

[9] E. Serrano-Ostáriz, A. Legaz-Arrese, J. L. Terreros-Blanco, M. López-Ramón, D. Cremades-Arroyos *et al.*, "Cardiac Biomarkers and Exercise Duration and Intensity During a Cycle-touring Event," *Clin J Sport Med*, vol. 19, no. 4, pp. 293–299, Jul. 2009.

[10] Ø. Skadberg, Ø. Kleiven, M. Bjørkavoll-Bergseth, T. Melberg, R. Bergseth *et al.*, "Highly increased Troponin I levels following high-intensity endurance cycling may detect subclinical coronary artery disease in presumably healthy leisure sport cyclists," *Eur J Prev Cardiol*, vol. 24, no. 8, pp. 885–894, May 2017.

[11] Ø. Kleiven, M. Bjoerkavoll-Bergseth, Ø. Skadberg, T. Melberg, B. Auestad *et al.*, "P3242prolonged release of cardiac troponin I after endurance exercise could indicate silent coronary artery disease in recreational athletes," *Eur Heart J*, vol. 38, no. suppl_1, Aug. 2017.

[12] K. Oskal, "Myocardial damage during mountain bike race - an analysis of data from Nordsjørittet 2014 (NEEDED study)," Master's thesis, University of Stavanger, Jun. 2016.

[13] D. Długosz, T. Eftestøl, S. Ørn, T. Wiktorski, and A. Królak, "The North Sea Bicycle Race ECG Project: Time-Domain Analysis," vol. 11. Prague: Annals of Computer Science and Information Systems, Sep. 2017, pp. 1353–1356.

[14] B. Vandenberk, E. Vandael, T. Robyns, J. Vandenberghe, C. Garweg *et al.*, "Which QT Correction Formulae to Use for QT Monitoring?" *Journal of the American Heart Association*, vol. 5, no. 6, Jun. 2016.

[15] Thomas B. Garcia and Neil E. Holtz, *EKG - sztuka interpretacji*, 1st ed. Warszawa: MediPage, 2007.

[16] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*, 11th ed. Pennsylvania: ELSEVIER SAUNDERS, 2006.

[17] E. Pietka and J. Kawa, *Information Technologies in Biomedicine*. Springer Science & Business Media, May 2010.

[18] M. L. Talbi and A. Charef, "PVC discrimination using the QRS power spectrum and self-organizing maps," *Comput Methods Programs Biomed*, vol. 94, no. 3, pp. 223–231, Jun. 2009.

[19] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *J Am Coll Cardiol*, vol. 37, no. 1, pp. 153–156, Jan. 2001.

[20] R. Olson, R. Urbanowicz, P. Andrews, N. Lavender, L. Kidd *et al.*, "Automating Biomedical Data Science Through Tree-Based Pipeline Optimization," in *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*, ser. Lecture Notes in Computer Science. Springer, Cham, Mar. 2016, pp. 123–137.

[21] M. Bjorkavoll-Bergseth, O. Kleiven, T. Melberg, O. Skadberg, B. Uestad *et al.*, "Increased heart rate does not explain the cardiac troponin increase following strenous exercise - the needed 2014 advanced heart rate monitor substudy," *Europrevent*, p. 296, 2018.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1069–1072
DOI: 10.15439/2018F191
ISSN 2300-5963 ACSIS, Vol. 15

The Design of Digital Filter System used in Stimulation with Tomatis Method

Krzysztof Jóźwiak
Lodz University of Technology
Institute of Electronics
ul. Wólczańska 211/215, 90-924 Łódź, Poland
Email: krzysztof.jozwiak@edu.p.lodz.pl

Michał Bujacz, Aleksandra Królak
Lodz University of Technology
Institute of Electronics
ul. Wólczańska 211/215, 90-924 Łódź, Poland
Email: {aleksandra.krolak, michal.bujacz}@p.lodz.pl

***Abstract*—The Tomatis Method is a rehabilitation technique used in psychology, the main aim of which is stimulating the cochlea in the inner ear by filtered air-conducted and bone-conducted sounds. The system of electronic filters and amplifiers used for this therapy is called the Electronic Ear. Commonly, it is an analog device, that is expensive and after a few years its functionality declines. In this paper, we introduce a digital Electronic Ear system using an STM32F4 family micro-controller and ADC/DAC integrated circuits. The design of digital sound filters allows to adjust more parameters and overcomes some of the constraints of analog systems. In this paper, we provide a short review of the Tomatis Method, the main functions of the Electronic Ear and we describe the designed system with comparison measurements to the analog one.**

I. Introduction

MUSIC therapy has numerous applications in psychology and rehabilitation. It can be used to help patients manage stress, concentration or depression. However, in many cases these methods do not give visible effects, when they are compared with a control group [1-4].

The Tomatis Method (TM) has been shown to help people with various psychological disorders (autism, dyslexia, ADD and more) [10], [11], [12], [13], [14], [15], [16], [17], [18], [20] though in some controlled trials the results were also inconclusive [8],[19]. It uses music, but it is not a music therapy, as it is based on brain and inner ear stimulation by selective sound filtering. The electronic device that implements the filter system for TM is called an Electronic Ear (EE). Most EEs that are commonly used are analog systems, that have their own constraints and high prices. We propose a new, digital Electronic Ear (DEE) system designed for cheaper implementation of TM and its further studies.

II. Tomatis Audio-Psycho-Phonology

Tomatis Laws cover the theory behind the relationship of hearing and speaking. Alfred Tomatis observed that problems with singing or speaking in some particular spectrum are connected with difficulties with hearing in the same spectrum and proposed that the improvement of vocal and voice range can be provided by opening the hearing ability to a specific frequency spectrum. This principle is called the ***Tomatis effect***. Over the years his ideas were implemented in other forms of sound therapy and a generic Electronic Ear tool was developed.

The Electronic Ear is based on the concept of stimulating the basilar membrane (BM) in the cochlea in two ways - by air and bone conduction. ***Air conduction*** is based on the typical ear pathway, with sound vibrations picked up by the eardrum, transferred by the ossicles to the oval window of the cochlea. ***Bone conduction*** is the perception of vibrations travelling through the bone and stimulating the BM either through pressure changes in the cochlear fluid or vibrations of the cochlear wall [5]. Bone conducted sounds reach the brain faster due to the higher density of the medium; however, large part of the acoustic signal is reflected due to the air-bone impedance mismatch. That is why bone conduction mostly favors our own voice. The conclusion drawn by Tomatis from this observation was that air conduction is used to "communicate with the outer world" and bone conduction is used to "listen to oneself". [6],[7].

The next feature that is important for TM is human earedness, i.e. which ear is dominant in the hearing process. Left ear dominance is problematic, because the left ear has the longer path to the left hemisphere. The delay between left and right ear is about 0.4-0.9 ms. People, who are left-eared have a lot of listening problems that can cause various psychological disease, such as dyslexia [6]. One of the main aims of Tomatis Rehabilitation is to increase the role of the right ear for the people who are left-eared. For this reason there are separate delays and adjustable volume levels for left and right channels.

Fig. 1 shows the flow chart of the EE. It is an audio system, the main function of which is amplifying, gating and filtering sound going to the three output channels - air left, air right and bone conduction. Gating is the process of changing the filter type depending on signal level. Below a certain amplitude level the C1 filter is on and above a certain level the filter is switched to C2 filter. The sudden change gives an impulse that theoretically stimulates the brain. The C1 and C2 filters are shelving filters with 1 kHz cross-band frequency. The boost/attenuation level is adjusted by the specialist administering the TM therapy and can be set from -5 to 5 dB. The EE also contains a second filter - a High Pass filter with an adjustable cut-off frequency. It is supposed function is to imitate the medium of a mother's womb. In some cases there is a need to increase or decrease a role of bone conduction in comparison to air conduction. For this reason

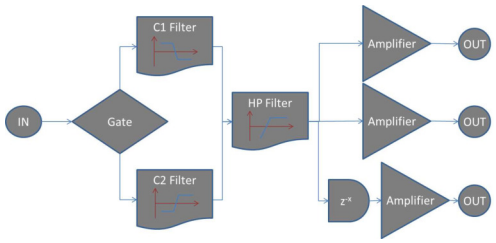


Fig. 1. Flow chart of the Electronic Ear (EE). The Gate depending on the signal level diverts it to the shelving filters C1 or C2. The HP filter cuts off low frequencies. The Delay allows to the air and bone conducted sounds to arrive at different times and the Amplifiers control the level of the three outputs.

there is an adjustable delay between the air and bone channel called ***precession***.

TM is used for many purposes. One of the popular aims is in therapy of autistic children. According to Tomatis theory autism is connected with problems with listening to the external world - in most autistic children bone conduction sensitivity is higher than air conduction. The therapy's aim is to lower this sensitivity in order to open the child up to communication with others. There were many studies carried out to verify the effectiveness of this rehabilitation. One of the most controversial was [8]. It revealed that there was no clear difference between the effects on a control group and an experimental group put under TM rehabilitation. For response to this article there were two follow-up studies [9], [10]. They showed the opposite view and pointed out missteps in the previous research. These two and another one [11] show a significant improvement after the TM rehabilitation measured in the Children's Autism Rating Scale (CARS) or in Gilliam Autism Rating Scale (GARS).

An overview of over 30 TM studies by Gerritsen showed that the majority of studies demonstrated positive effects of TM therapy. Some of the better documented TM treatments include:

- reduction of the Attention Deficit Disorder (ADD) by, similarly to autism, increasing the focus on air conducted sounds versus bone conduction [6],[12],[13].
- treating dyslexia [6],[13],[20] (that is theoretically connected with left-earedness) by increasing the sensitivity of the right ear.
- supporting language learning [14], by setting the C1/C2 filters to frequencies most commonly appearing in a given language. TM rehabilitation has also been successfully used for:
- Epilepsy - visible effect in more than 50
- Cerebral palsy: significant improvement [16]
- Stuttering and hoarseness [13]
- Emotional problems including depression: high effect, compared with a control group [17]
- Music skills improvement: results are divided [18],[19]

Most of tests have shown a positive effect of TM, but there are also those, that did not report any differences when compared to a placebo treatment. We need to consider that negative results can stem from inappropriateness in investigation methods or in the way the EE device was used [9]. This shows that the TM requires further rigorous studies.

III. Electronic Ear Stimulator

In terms of the electronic design main functions of Electronic Ear are: Converting audio signal to digital data (ADC); Pre-processing; Gating; Filtering; Digital-Analog converting (DAC); Amplifying. The device can be divided into four main modules: audio input, audio output, micro-controller and power supply. Three voltage levels are needed:

- 5V - external power supply. It is used as an input for the step-down converter and as an analog power supply for the input section. The transformer power source is used in order to decrease the distortions in the low frequency audio signals.
- 3.3V - it is given by Low Dropout Positive Regulator TC2117 from MicroChip. It is used for digital power supply for μC and for the digital part (communication interfaces) of the ADC/ DAC converters.
- 2.5V - it is necessary for the output analog power supply and it is provided by LP2985-N

In this project external 24-bit stand-alone integrated circuits were used. For input a PCM4201 from Texas Instruments and for output CS432L22 from Cirrus Logic were used. In both cases communication between the audio converter and μC was provided by an I2S interface, the audio data word length is 24 bits and is Left-Justified in the frame. CS43L22 supports I2S and it was working in the slave mode by being connected to four pins: MCLK, LRCK, SCLK and SDO. It has its own transmission interface that has 3 communication ports - BCK (equivalent to SCLK), DATA (equivalent to SDO), FSYNC (equivalent to LRCK) and one system clock port, that is necessary for proper work. The serial interface is very similar to I2S and after a few modifications it can work in high performance mode communicating in master mode. The main difference is that the system clock needs to be 512 times larger than the sampling frequency (256 times larger in I2S mode). To get that result an extra I2S interface was used with a 2 times larger sampling frequency and the MCLK pin connected to the system clock pin in the audio converter. That allows a clean transmission without errors.

The CS43L22 is a DAC converter, but also a headphone amplifier. It allows to prevent from using two independent chips (one for converting, one for amplifying) and minimalizes space on the PCB board. In the project there are two output channels - air and bone channel. In order to support three channels (that are required for the EE) there is used an external adapter that splits the air channel to left and right channels with a build-in balance regulator. CS43L22 has two serial interfaces and except of I2S for data transmission it has also I2C interface for control. It allows to regulate all functions of the chip, for example volume, that is continuously sent to the control register.

For digital signal processing the STM32F407VGT6 microcontroller was used. Data was stored in a float buffer. Its

size was set in order to provide maximum 500 milliseconds delay, which for 48kHz frequency sample gave (24000 + filter order) size. User interface gives possibility to adjust the main parameters of the EE - C1, C2 levels value, frequency, precession and volume and the advanced options as gate upper level and four options for filter order. Changing filter order affects its selectivity and as a result the intensity of stimulation.

There were used filter orders in range of 51 to 101 (only odd orders are used). Every filter was designed as FIR (Finite Impulse Response) filter using the window method. The window that was chosen for this purpose was modified Tukey window (tapered cosine window). It is combination of a rectangular window and a Hamming window with an α coefficient that determines what part of the window should be flat. It makes cosine lobe of width $\alpha/2$ * N (N is filter order) and rectangular window of width (1 - $\alpha/2$) * N at the center of filter function. With α = 0.8 it gives the best results - sufficient filter selectivity providing expected filter band levels even for low cross-frequency (500 Hz is a minimum used value) and low ripple frequency response.

In this project two types of filters are required - simple HP/LP filter and a shelving filter for the C1 and C2 filtering. To obtain a two-level shelving filter it was necessary to use a parallel connection of filters - one for left part (LP filter) and one for right part (HP filter). The cross-frequency is set where the characteristic crosses the 0 dB line. For that reason there was a need to modify one frequency by a coefficient dependent on filter order and C1 and C2 levels.

The big challenge was to appropriately design the filter in case when both shelving filter levels have the same sign. Normally it should not cross the zero decibels line. But measurements of the original EE show, that with this case it should be artificially modified in order to give 0 dB gain at the cross-frequency point. A cascade connection of two filters with similar cut-off frequency was used, but modified regarding to the purpose. For the positive shelving filter levels the gain near the cross-frequency was decreased (frequencies of filters diverge) and for negative it was increased (frequencies converge). Then, obtained filter is serially connected with the normal shelving filter.

In this section the frequency-responses of filters in the designed DEE are showed and compared with those measured for the original, analog EE device.

Fig. 2 and 3 show frequency-responses for 101-order filters for designed device. As it can be seen, passband levels have correct values, remain stable and C1 and C2 filter lines cross near the 0 dB level. In the first case the 500 Hz (the minimal needed value for this project) and in the second 1 kHz (frequency of analog EE) cross-band frequency was set. In figure 3 the filter with two positive levels was shown to provide appropriate gain modification near cross-frequency.

IV. Results

In Fig. 4 and 5 frequency responses measured for the DEE and analog EE are shown for 1 kHz cross-band frequency and [±5, ±5] and [±5, ±1] levels respectively. The unstable

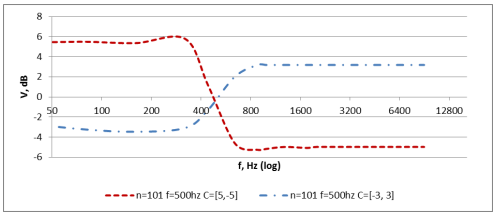


Fig. 2. Frequency-response for n=101, f=500 Hz and passband levels set.

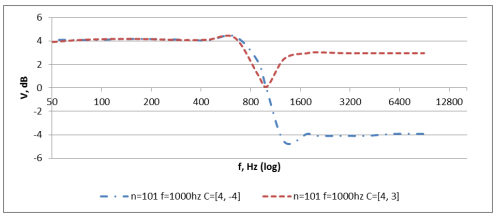


Fig. 3. Frequency-response for n=101, f=1000 Hz and passband levels set.

and exceeded passband levels during the whole frequency bandwidth for analog devices can be seen. In Fig. 5 one can notice the ringing artifacts for the bandwidth above 1 kHz. The cross-point for analog device also has an incorrect value i.e. 850Hz for first case and 865 Hz for the second case.

V. Conclusions

The DEE, which was the topic of this paper was designed properly and it provides all the required functions. It services 2 channels, that with external splitting adapter allow to support three channels. If the right/left channel delay was needed, there would be a possibility to use second, same DAC integrated device for extra output channels. The DEE uses 24-bit audio data with 48kHz sampling frequency. The included filters give the desired effects - the shelving filter levels and cross-frequency value are as expected. It monitors changes at air and bone channels and gates them to C1 or C2 filters depending on the signal level and precession value.

Compared to the analog EE, the designed device has all the same functions, but it has also extra - features allowing to remove some previous constraints. First it is a digital system, so the filter parameters are stable and do not depend on

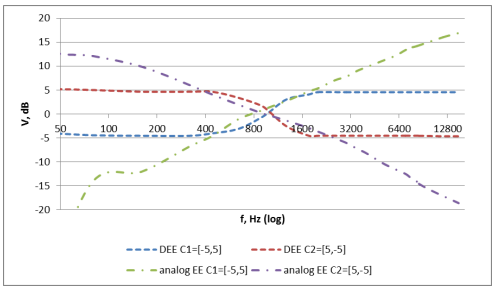


Fig. 4. Frequency-response of C1/C2 filters measured for digital and analog device. Theoretically, the analog EE's pass bands should be flat, but the measurements show they clearly have a linear drop-off.

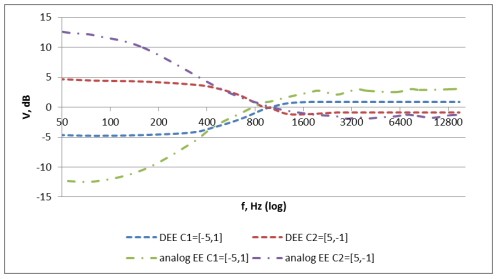


Fig. 5. Frequency-response of C1/C2 filters measured for the proposed digital and the original analog TM device.

the duration of operation or component aging. In Figures 4 and 5 there is visible effect of this occurrence for the analog device - the filter levels are different than expected and are unstable. We can only speculate about the discrepancies between the theoretical spectra and those measured for the commercial EE analog device. One theory is component aging (e.g. degradation of dielectric in the capacitors) or it could simply be a design flaw. Another advantage is that the cost of production of the DEE (less than 200EUR) is much lower than the analog TM device (approximately 10,000EUR). The fact that filters in the designed system are digital also gives an opportunity to develop it in the future, change the parameters or to upgrade it depending on the requirements.

The next advantage is that digitalization allows more parameters to be adjusted by the user. It is possible to set the cross-frequency of the shelving filter. In analog EE C1 and C2 filters have always the same cross-frequency and it is 1 kHz. It has also only one type of filter selectivity. In the designed device four cases for filter order allow to change the intensity of stimulation and pick the best one for the individual patient. The changeable filter frequency allows to modify frequency band, that impacts the most to the listener.

To sum up, a modernized digital version of the device used for the rehabilitation by the Tomatis method was designed and tested. Its performance was measured to be the same or better than the original equipment, while allowing much more control over the functionality.

REFERENCES

[1] L. Hohmann, J. Bradt, T. Stegemann, S. Koelsch, "Effects of music therapy and music-based interventions in the treatment of substance use disorders: A systematic review," *PLOS,* 2017, doi: 10.1371/journal.pone.0187363.

[2] M. Bodner, R. P. Turner, J. Schwacke, C. Bowers, C. Norment, "Reduction of Seizure Occurrence from Exposure to Auditory Stimulation in Individuals with Neurological Handicaps: A Randomized Controlled Trial," *PLOS,* 2012, doi: 10.1371/journal.pone.0045303.

[3] M. Bucur, A. L. Marian, "The impact of the Mozart effect on creativity myth or reality," *Creativity,* 2016.

[4] L. C. Lin, W. T. Lee, H. C. Wu, C. L. Tsai, R. C. Wei, H. K. Mok, "The long-term effect of listening to Mozart K.448 decreases epileptiform discharges in children with epilepsy," *Epilepsy Behav.* vol. 21(4), 2011, pp. 420–424, doi: 10.1016/j.yebeh.2011.05.015.

[5] P. Henry, T. Letowski, *Bone conduction: Anatomy, physiology, and communication,* Army Research Laboratory, 2007.

[6] P. Sollier, *Listening for wellness: An Introduction to the Tomatis Method,* The Mozart Center Press, 2005.

[7] N. Doige, *A bridge of sound in: The brain's way of healing,* Ed. Mińska Barbara, Vital, 2016.

[8] B. A. Corbett, K. Shickman, E. Ferrer, "Brief Report: The Effects of Tomatis Sound Therapy on Language in Children with Autism," *J. Autism Dev Disord.,* vol. 38(3), 2007, pp. 562–566, doi: 10.1007/s10803-007-0413-1

[9] J. Gerritsen, "Response to Brief Report: The Effects of Tomatis Sound Therapy on Language in Children with Autism," *J. Autism Dev Disord.,* vol. 38(3), 2008, pp. 567, doi: 10.1007/s10803-007-0471-4

[10] M. AbediKoupaeia, K. Poushanehb, A. Z. Mohammadic, N. Siampour, "Sound Therapy: an Experimental Study with Autistic Children," *Procedia - Social and Behav. Sciences,* vol. 84, 2013, pp. 626–630

[11] J. M. Neysmith-Roy, "The Tomatis Method with severely autistic boys: Individual case studies of behavioral changes," *South African J. Of Psychology,* vol. 31(1), 2001.

[12] L. Sacarin, *Early Effects of the Tomatis Listening Method in Children with Attention Deficit,* AURA - Antioch University Repository and Archive, 2013.

[13] J. Gerritsen, *A Review of research done on Tomatis Auditory Stimulation,* Mozart Center, 2009.

[14] I. M. du Toit, W. F. du Plessis, D. K. Kirsten, "Tomatis Method Stimulation: Effects on Student Educational Interpreters," *J. of Psych. in Africa,* vol. 21(2), 2011, doi: 10.1080/14330237.2011.10820454.

[15] G. Coppola, A. Toro, F. F. Operto, G. Ferrarioli, S. Pisano, A. Viggiano et al., "Mozart's music in children with drug refractory epileptic encephalopathies," *Epilepsy Behav.,* vol. 50, 2015, pp. 18–22, doi: 10.1016/j.yebeh.2015.05.038.

[16] I. Przybek-Czuchrowska, E. Mojs, E. Urna-Bzdęga, "Opis przypadku dziecka z organicznym uszkodzeniem w obrębie ośrodkowego układu nerwowego leczonego metodą treningu słuchowego Tomatisa," *Neuropsychiatria i Neuropsychologia,* vol. 10(1), 2015, pp. 40–45.

[17] J. O. Coetzee, *The effect of the Tomatis Method on depressed young adults,* Potchefstroom University, 2001.

[18] W. du Plessis, S. Burger, M. Munro, D. Wissing, W. Nel, "Multimodal enhancement of culturally diverse, young adult musicians: a pilot study involving the Tomatis method," *South African J. of Psych.,* vol. 31(3), 2001.

[19] A. Vercueil, H. Taljaard, W. du Plessis, "The effect of the tomatis method on the psychological well-being and piano performance of student pianists: an exploratory study," *SAMUS,* vol. 31, 2011.

[20] T. Gilmor, "The efficacy of the Tomatis method for children with learning and communication disorders: A meta-analysis," *Int. J. of Listening,* vol. 13(1), 1999, doi: 10.1080/10904018.1999.10499024.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1073–1077
DOI: 10.15439/2018F308
ISSN 2300-5963 ACSIS, Vol. 15

New Grid for Particle Filtering of Multivariable Nonlinear Objects

Piotr Kozierski[1,2)], Jacek Michalski[2)], Talar Sadalla[2)], Wojciech Giernacki[2)], Joanna Ziętkiewicz[2)], Szymon Drgas[1)]
Poznan University of Technology, Piotrowo street 3a, 60-965 Poznan, Poland[1,2)]
Faculty of Computing, Institute of Automation and Robotics, Division of Signal Processing and Electronic Systems[1)]
Faculty of Electrical Engineering, Institute of Control, Robotics and Information Engineering, Division of Control and Robotics[2)]
Email: piotr.kozierski@gmail.com, jacek.michalski95@wp.pl

***Abstract*—In the paper a new grid (potentially linear, nonlinear and even semi-Markovian jump system) was presented. All transition and measurement functions were proposed. Moreover, the transition functions of two types were considered – dependent on one and many different state variables. Also 10 types of measurements were proposed for both nodal and branch cases. Based on the obtained results one can see, which measurement functions are "easy", and which are "hard" for state estimation task.**

I. Introduction

PARTICLE filter (PF) is potentially very good estimation method because is based on the optimal solution – Bayes filter. The biggest disadvantage of PFs is their need for computational power – number of calculations grows exponentially with a system variables number [1]. This is the reason why PF methods are usually used only for very small plants.

Some solutions to this problem are hybrid filters, e.g. Rao-Blackwellized PF (RBPF) [2]-[3], or Marginalized PF [4], in which state variables are divided into two groups – one group is estimated using PF method, and the second group – using Kalman Filter (KF) methods (linear, extended or unscented).

Another solution was proposed in [5] – all state variables are divided into groups; however, the disadvantage of this method is loss of information contained in measurements, which uses state variables from two or more groups.

Dispersed Particle Filter (DPF) was proposed by the authors in one of the previous works [6] – this method assumes that dependences between state variables and measurements are relatively sparse (transition and measurement models depend on relatively small number of different state variables). Unfortunately, previously used plants – the power systems – have number of state variables two times higher than number of nodes. Additionally, based on studies from [7], the reasonable number of particles should be about 4 to the power of plant variables number. This caused reduction of plant dimension and simultaneously increase the number of nodes is needed.

This work was not supported by any organization

For this reason the authors proposed a new grid for further research. In this network, one node is associated with one state variable – thanks to this, considerations on the network structure will be possible for systems with relatively small state vector length. Proposed grid is very general and one can model both linear and highly nonlinear models (for both transitions and measurements).

In the second section particle filter algorithm is described. In Section III the proposed grid is presented. Prepared simulations and obtained results are shown in the fourth section. The last section contains drawn conclusions.

II. Particle Filter

An object in state space can be written as

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{f}\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}, k\right) + \mathbf{v}^{(k)} \\ \mathbf{z}^{(k)} = \mathbf{h}\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}, k\right) + \mathbf{n}^{(k)} \end{cases}, \tag{1}$$

where $\mathbf{x}^{(k)}$ is a state vector, $\mathbf{u}^{(k)}$ is input vector, $\mathbf{z}^{(k)}$ is output vector, and vectors $\mathbf{v}^{(k)}$ and $\mathbf{n}^{(k)}$ are internal and measurement noises, respectively – all at k-th time step. The main task of particle filter is to estimate state vector based on the measurements and input signals.

The particle filters operation principle is based on the recursive Bayesian filtering [8]

$$\overbrace{p\left(\mathbf{x}^{(k)} \mid \mathbf{Z}^{(k)}\right)}^{\text{posterior}} = \frac{\overbrace{p\left(\mathbf{z}^{(k)} \mid \mathbf{x}^{(k)}\right)}^{\text{likelihood}} \cdot \overbrace{p\left(\mathbf{x}^{(k)} \mid \mathbf{Z}^{(k-1)}\right)}^{\text{prior}}}{\underbrace{p\left(\mathbf{z}^{(k)} \mid \mathbf{Z}^{(k-1)}\right)}_{\text{evidence}}}, \tag{2}$$

where $\mathbf{Z}^{(k)}$ is a set of measurement vectors from the first to k-th time step, $p(\mathbf{x}^{(k)}|\mathbf{Z}^{(k)})$ is a posterior Probability Density Function (PDF), $p(\mathbf{x}^{(k)}|\mathbf{Z}^{(k-1)})$ is a prior PDF, $p(\mathbf{z}^{(k)}|\mathbf{x}^{(k)})$ is a likelihood, and $p(\mathbf{z}^{(k)}|\mathbf{Z}^{(k-1)})$ is an evidence.

The key idea in PF is to implement the posterior PDF as a set of particles. The i-th particle is represented by a pair $\{\mathbf{x}^{i,(k)}, q^{i,(k)}\}$ – value (state vector) and weight. Higher weight increases probability that the value $\mathbf{x}^{i,(k)}$ is close to the real state vector. If the number of particles, N, is high enough, the information about the posterior PDF contained in particles set is the same as in continuous function.

The first particle filter was proposed in 1993 by Gordon, Salmond and Smith [9] and was called Bootstrap Filter. Operation principle of this PF is presented in Algorithm 1.

Algorithm 1 – Bootstrap Filter

1. Initialization. Draw N initial values $\mathbf{x}^{i,(0)}$ from initial PDF $p(\mathbf{x}^{(0)})$, set time step k=1.
2. Prediction. Draw N new particles from the transition model $\mathbf{x}^{i,(k)} \sim p(\mathbf{x}^{(k)}|\mathbf{x}^{i,(k-1)})$.
3. Update. Compute the particle weights based on the measurement model $q^{i,(k)} = p(\mathbf{z}^{(k)}|\mathbf{x}^{i,(k)})$.
4. Normalization. Normalize weights so that their sum be equal to 1.
5. Resampling. Draw N new particles using the posterior PDF obtained in previous steps (the chance that particle will be drawn is equal to normalized weight).
6. End of iteration. Calculate the estimated state vector, increase the time step k=k+1, go to the second step.

PF can be used for both non-Gaussian distributions and complex transition models. Example of such complex system model can be semi-Markovian jump system [10], in which the whole system model can switch itself (with some probability) into other structures (other equations), and also can go back to the previous "system states".

For more information about PFs, references [11]-[15] are recommended.

III. Proposed Network

The authors proposed grid, which on the one hand can be easily prepared, but on the other hand provides wide possibilities in creation of new plants. This is why there are many complex and maybe even illegible options presented below; however, the most common networks will be presented in a very simple way. Moreover, if one needs to add any dependence, which was not specified, still there is a possibility to present this on the scheme.

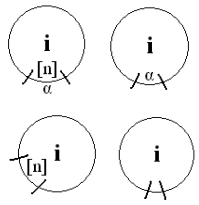
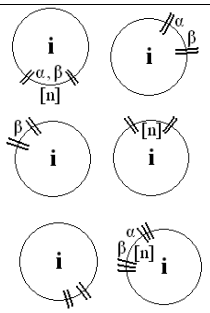
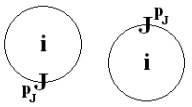
The proposed network is composed of nodes and lines (branches). Two different nodes can be connected by line. With every i-th node exactly one state variable x_i is associated with. Nodes can be represented in two ways – by circles or by squares. Transition function f_i depends on the shape of i-th node.

Filled figures are associated with measurements that when placed on lines (branches), refer to state variables that are associated to those lines (branch measurements), and when placed near the nodes (or on the nodal branches), refer to state variables in those specific nodes (nodal measurement).

It was assumed that networks are autonomous and thus any designation for input signal is not presented.

Expressions for different transition functions of the first type (circles) are presented below, and their connections with scheme designations are described in Table I.

TABLE I.
EXPLANATION OF DESIGNATIONS – NODES, PART 1, FOR TRANSITION FUNCTIONS WHICH ARE BASED ONLY ON I-TH STATE VARIABLE

Eqn.	Designation of node	Explanation
(3)		α and n should be given on the scheme. However, if one parameter is omitted, it is assumed that this value is equal to 1. Also both parameters can be omitted (α=1 and n=1).
(4)-(5)		When all three values are given, one must take into account that α should be written before β (above or on the left side of β). If any parameter (α, β or n) is not presented, it is assumed that it is equal to 1; however, one should keep in mind that β can be omitted only if α is also omitted. Designations for equation (5) are the same, but triple lines (through circles) should be used.
(6)		p_J should be written outside of the circle as J's sub- or super-script (also from the left). If p_J is omitted, it is assumed that p_J = 0.5.
-		To use another function (which must be explain in a text) a double circle should be used.

$$x_i^{(k+1)} = \begin{cases} \alpha\left(x_i^{(k)}\right)^n + v_i^{(k)} & n \in \mathcal{Z} \\ \alpha\left|x_i^{(k)}\right|^n + v_i^{(k)} & n \notin \mathcal{Z} \end{cases} \tag{3}$$

$$x_i^{(k+1)} = \begin{cases} \dfrac{\alpha\left(x_i^{(k)}\right)^n}{\beta + \left(x_i^{(k)}\right)^2} + v_i^{(k)} & n \in \mathcal{Z} \\ \dfrac{\alpha\left|x_i^{(k)}\right|^n}{\beta + \left(x_i^{(k)}\right)^2} + v_i^{(k)} & n \notin \mathcal{Z} \end{cases} \tag{4}$$

$$x_i^{(k+1)} = \begin{cases} \alpha\left(x_i^{(k)}\right)^n \cdot \cos(\beta k) + v_i^{(k)} & n \in \mathcal{Z} \\ \alpha\left|x_i^{(k)}\right|^n \cdot \cos(\beta k) + v_i^{(k)} & n \notin \mathcal{Z} \end{cases} \tag{5}$$

$$x_i^{(k+1)} = \begin{cases} f_i\left(x_i^{(k)}, \mathbf{u}^{(k)}, k\right) + v_i^{(k)} & \text{with prob. } p_J \\ -f_i\left(x_i^{(k)}, \mathbf{u}^{(k)}, k\right) + v_i^{(k)} & \text{with prob. } (1 - p_J) \end{cases} \tag{6}$$

For the second type of transition functions, connections between nodes matter. Branch between i-th and j-th nodes has a value $\mu_{i,j}$=$\mu_{j,i}$≠0, whereas if there is no connection between i-th and j-th nodes, branch value $\mu_{i,j}$=$\mu_{j,i}$=0. It is also assumed that $\mu_{i,i}$=1.

There are specified three types of lines, which differ in line functions $fl_{i,j}$ (where i is the number of node from which line function "was called"). These functions are written

below, and their connections with scheme designations are described in Table II.

$$fl_{i,j}^{(k)} = \begin{cases} \left(x_j^{(k)}\right)^m \cdot \mu_{i,j} & m \in \mathcal{Z} \\ \left|x_j^{(k)}\right|^m \cdot \mu_{i,j} & m \notin \mathcal{Z} \end{cases} \tag{7}$$

$$fl_{i,j}^{(k)} = \begin{cases} \sin\left(\left(x_j^{(k)}\right)^m - x_i^{(k)}\right) \cdot \mu_{i,j} & m \in \mathcal{Z} \\ \sin\left(\left|x_j^{(k)}\right|^m - x_i^{(k)}\right) \cdot \mu_{i,j} & m \notin \mathcal{Z} \end{cases} \tag{8}$$

$$fl_{i,j}^{(k)} = \ln\left(0.001 + \left|x_j^{(k)}\right|^m\right) \cdot \mu_{i,j} \tag{9}$$

Line functions $fl_{i,j}$ can be used in both, transition and measurement functions. Proposed types of transition functions, which use values of other state variables, are presented below, and their designations are described in Table III.

$$x_i^{(k+1)} = \begin{cases} \alpha\left(x_i^{(k)}\right)^n \cdot \sum_{\substack{j=1 \\ j\neq i}}^{N_x} fl_{i,j}^{(k)} + v_i^{(k)} & n \in \mathcal{Z} \\ \alpha\left|x_i^{(k)}\right|^n \cdot \sum_{\substack{j=1 \\ j\neq i}}^{N_x} fl_{i,j}^{(k)} + v_i^{(k)} & n \notin \mathcal{Z} \end{cases} \tag{10}$$

$$x_i^{(k+1)} = \alpha \cdot \sum_{\substack{j=1 \\ j\neq i}}^{N_x} \frac{fl_{i,j}^{(k)}}{\beta + \left|fl_{i,j}^{(k)}\right|^n} + v_i^{(k)} \tag{11}$$

$$x_i^{(k+1)} = \alpha\left(\cos\left(x_i^{(k)} \cdot \beta k^n\right) + \sum_{\substack{j=1 \\ j\neq i}}^{N_x} \cos\left(fl_{i,j}^{(k)} \cdot \beta k^n\right)\right) + v_i^{(k)} \tag{12}$$

$$x_i^{(k+1)} = \begin{cases} f_i\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}, k\right) + v_i^{(k)} & \text{with prob. } p_J \\ -f_i\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}, k\right) + v_i^{(k)} & \text{with prob.} \left(1 - p_J\right) \end{cases} \tag{13}$$

The type of internal noise one can describe in a text or mark on the scheme. PDF type with parameters should be connected with specific node by dashed line. Examples have been presented in Fig. 1.

Measurements are marked by filled figures on the scheme. Measurement designations on branches have different meaning than designations associated with nodes. Moreover, one should keep in mind that measurement location matters (the first index indicates near which node the measurement is located), because measurement functions generally are not symmetric. Possible nodal measurements in i-th node (P_i, Q_i, R_i, S_i, T_i) are described by equations (14)-(18), whereas possible branch measurements between i-th (at this node measurement is placed) and j-th nodes ($P_{i,j}$, $Q_{i,j}$, $R_{i,j}$, $S_{i,j}$, $T_{i,j}$) are described by equations (19)-(23). Designations of specific measurements are presented in Table IV. To use another measurement function one can simply use new filled figure on the scheme (function should be explained in text).

TABLE II.
EXPLANATION OF DESIGNATIONS – LINES

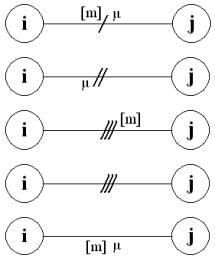
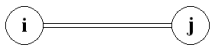
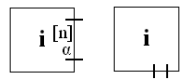
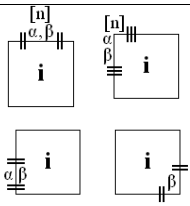
Eqn.	Line designation	Explanation
(7)-(9)		Values m and μ should be written on the scheme near to line center. Also information about branch type, in the form of diagonal lines (one for (7), two for (8) and three for (9)), should be presented there. If $\mu_{i,j}$=1, one can omit this value. If m=1 it also can be omitted. If branches in whole grid are only of first type, diagonal lines can be omitted.
-		For another line function (must be explain in text), double line should be used between nodes.

TABLE III.
EXPLANATION OF DESIGNATIONS – NODES, PART 2, FOR TRANSITION FUNCTIONS WHICH ARE BASED ALSO ON OTHER STATE VARIABLES

Eqn.	Designation of node	Explanation
(10)		n value should be written inside a square bracket. If α is omitted, it is assumed that α=1. If value n is omitted, it is assumed that n=1.
(11)-(12)		When all three values are given, one must take into account that α should be written before β (above or on the left side of β). If any parameter (α, β or n) is not presented, it is assumed that it is equal to 1; however, one should keep in mind that β can be omitted only if α is also omitted.
(13)		p_J should be written outside of the square as J's sub- or super-script (can be also from the left side). Omitted p_J means that p_J = 0.5.
-		To use another function (which must be explain in a text) double square should be used.

$$P_i^{(k)} = z_o^{(k)} = \begin{cases} \alpha \cdot \left(x_i^{(k)}\right)^m + n_o^{(k)} & m \in \mathcal{Z} \\ \alpha \cdot \left|x_i^{(k)}\right|^m + n_o^{(k)} & m \notin \mathcal{Z} \end{cases} \tag{14}$$

$$Q_i^{(k)} = z_o^{(k)} = \alpha \cdot \prod_{\substack{j=1 \\ \mu_{i,j}\neq 0}}^{N_x} x_j^{(k)} \cdot \mu_{i,j} + n_o^{(k)} \tag{15}$$

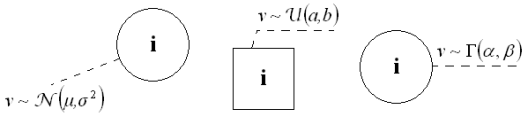


Fig. 1 Examples of internal noise designation

$$R_i^{(k)} = z_o^{(k)} = \alpha \cdot x_i^{(k)} \cdot \sum_{\substack{j=1 \\ j\neq i}}^{N_x} fl_{i,j}^{(k)} + n_o^{(k)} \tag{16}$$

$$S_i^{(k)} = z_o^{(k)} = \\ = \alpha \cdot \sum_{j=1}^{N_x} x_i^{(k)} x_j^{(k)} \mu_{i,j} \sin\left(x_i^{(k)} - x_j^{(k)} - \mu_{i,j}\right) + n_o^{(k)} \quad (17)$$

$$T_i^{(k)} = z_o^{(k)} = \alpha \cdot x_i^{(k)} \cdot \\ \cdot \sum_{\substack{j=1 \\ j \neq i}}^{N_x} x_j^{(k)} \mu_{i,j} \cos\left(\frac{x_i^{(k)} - x_j^{(k)}}{\ln\left(1.01 + \left|fl_{i,j}^{(k)}\right|\right)}\right) + n_o^{(k)} \quad (18)$$

$$P_{i,j}^{(k)} = z_o^{(k)} = \alpha \cdot \left(x_i^{(k)} - x_j^{(k)}\right) + n_o^{(k)} \quad (19)$$

$$Q_{i,j}^{(k)} = z_o^{(k)} = \alpha \cdot x_i^{(k)} \cdot x_j^{(k)} + n_o^{(k)} \quad (20)$$

$$R_{i,j}^{(k)} = z_o^{(k)} = \alpha \cdot \\ \cdot \left(x_i^{(k)} + x_j^{(k)} \cdot \sin\left(\left(x_i^{(k)} + x_j^{(k)}\right) \cdot fl_{i,j}^{(k)}\right)\right) + n_o^{(k)} \quad (21)$$

$$S_{i,j}^{(k)} = z_o^{(k)} = \alpha \cdot \\ \cdot \left(x_i^{(k)^2} - x_i^{(k)} x_j^{(k)} \mu_{i,j} \cos\left(x_i^{(k)} - x_j^{(k)} - \mu_{i,j}\right)\right) + n_o^{(k)} \quad (22)$$

$$T_{i,j}^{(k)} = z_o^{(k)} = \alpha \cdot \frac{\max\left\{x_i^{(k)} + x_j^{(k)}; x_i^{(k)} \cdot x_j^{(k)}\right\}}{0.01 + \left|\min\{A; B\}\right|} \cdot fl_{i,j}^{(k)} \\ A = \log_2\left(0.01 + \left|x_i^{(k)} x_j^{(k)}\right|\right); \; B = -0.1\left(x_i^{(k)} + x_j^{(k)}\right)^2 + 2 \quad (23)$$

The measurement noise should be marked in the same way as the internal noise – by description in text or using dashed line on the scheme. Examples of measurement noise designations are presented in Table IV.

IV. Simulation results

The simulations were performed for the networks presented in Fig. 2-4. Measurements were considered for 5 different cases – P, Q, R, S, and T. In each case all possible measurements of specific type were taken into account (10 branch measurements and 4 nodal measurements). However, one can see that differences between the second and third objects are only in line functions $fl_{i,j}$. hence only experiments for R and T cases were repeated for Ob403.

The average Root Mean Square Error was used to describe estimation quality and it can be written as

$$\text{aRMSE} = \frac{1}{N_x} \sum_{j=1}^{N_x} \text{RMSE}_j \,, \quad (24)$$

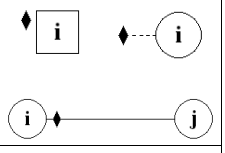
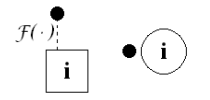
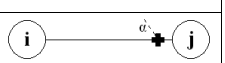
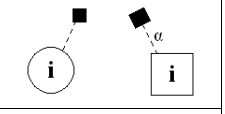
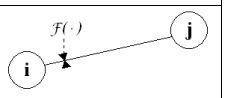
where N_x is the number of state variables, and

$$\text{RMSE}_j = \sqrt{\text{MSE}_j} = \sqrt{\frac{1}{M} \sum_{k=1}^{M} \left(x_j^{(k)} - x_j^{(k)+}\right)^2} \,, \quad (25)$$

where value with plus is the real value of state variable and M is the length of the simulation (set on the 1000 time steps in experiments).

All simulations were repeated 4000 times for Ob401, Q and T measurements, and 400 times for other cases. All results are presented in Fig. 5.

TABLE IV.
EXPLANATION OF DESIGNATIONS – MEASUREMENTS (BRANCH AND NODAL)

Eqn.	Measur. designation	Explanation
(14), (19)		Nodal measurements should be placed near the node or should be connected with node by the dashed line. The second case is needed if one want to add information about measurement noise or about measurement scaling. Branch measurements should be placed on the specific line. One should keep in mind that measurements are different at both ends of the branch, so the position of sign should be unambiguous. If one want to add information about measurement noise or its scale, also a dashed line should be used. Measurement designations can be marked in one orientation or if one wish – can be rotated; however, one should use one approach.
(15), (20)		
(16), (21)		
(17), (22)		
(18), (23)		

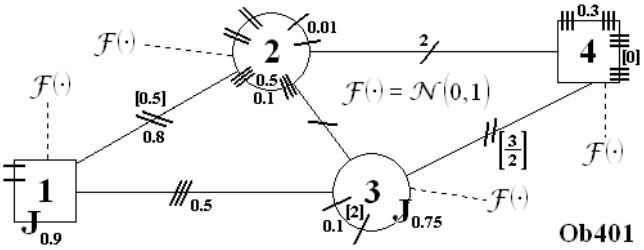


Fig. 2 Scheme of the object Ob401

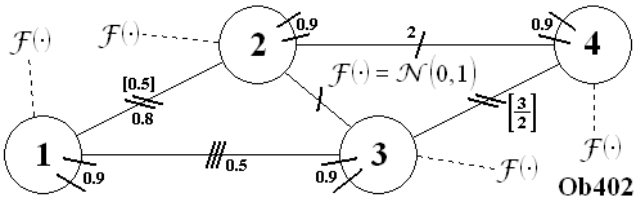


Fig. 3 Scheme of the object Ob402

V. Conclusions

Based on the obtained results one can say (by comparison results of the same cases for objects Ob401 and Ob402) that linearly changed state variables can be better tracked than ones changed nonlinearly. Moreover, the best estimation quality was obtained for linear object (linear transition model and linear measurement functions).

It is also visible (by comparison of results for R or T cases for all three objects) that complexity of measurements also impact on estimation quality (easier measurements provide better tracking).

The increase of aRMSE with higher number of particles for object Ob401 and cases Q and T is probably caused by higher chance that specific set of drawn values, which are

wrong, will match the measurements (index in one of dozens simulation had very high value). Simultaneously it is clearly visible that Q and T measurements are the hardest for estimation task. One can see also that for others measurement types even object with jump functions (Ob401) can be properly estimated.

It is also interesting that estimation quality for case Q in object Ob402 is rather weak for small number of particles *N*, but is very good for high particles number, whereas for case Q in object Ob401 the results are worst of all examined cases and objects. This is probably caused by jump functions in Ob401.

Proposed network will be widely used by the authors for further research.

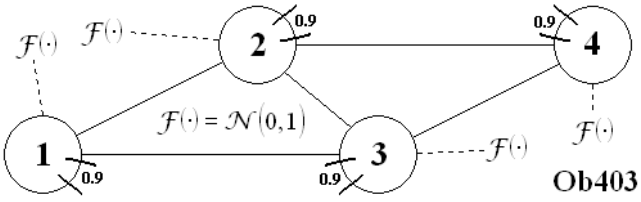


Fig. 4 Scheme of the object Ob403

REFERENCES

[1] Sutharsan S., Kirubarajan T., Lang T., McDonald M., An Optimization-Based Parallel Particle Filter for Multitarget Tracking, IEEE Transactions on Aerospace and Electronic Systems, Vol. 48, No. 2, 4/2012, pp. 1601-1618. DOI: 10.1109/TAES.2012.6178081

[2] Doucet A., Freitas N., Murphy K., Russell S., Rao-Blackwellised particle filtering for dynamic Bayesian networks, Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, 2000, pp. 176-183.

[3] Handeby G., Karlsson R., Gustafsson F., The Rao-Blackwellized Particle Filter: A Filter Bank Implementation, EURASIP Journal on Advances in Signal Processing, 2010, Article ID 724087, p. 10.

[4] Šmídl V., Hofman R., Marginalized Particle Filtering Framework for Tuning of Ensemble Filters, Monthly Weather Review, Vol. 139, No. 11, 2011, pp. 3589-3599.

[5] Djurić P. M., Lu T., Bugallo M. F., Multiple particle filtering, In 32nd IEEE ICASSP, April 2007, III pp. 1181-1184.

[6] Kozierski P., Lis M., Horla D., Level of Dispersion in Dispersed Particle Filter, in Methods and Models in Automation and Robotics (MMAR), 20th International Conference on, 2015, pp. 418-423.

[7] Kozierski P., Sadalla T., Owczarkowski A., Drgas S., Particle Filter in Multidimensional Systems, in Methods and Models in Automation and Robotics (MMAR), 21st International Conference on, 2016, pp. 806-810. DOI: 10.1109/MMAR.2016.7575240

[8] Candy J. V., Bayesian Signal Processing, WILEY, New Jersey 2009, pp. 36-44. DOI: 10.1002/9780470430583

[9] Gordon N. J., Salmond D. J., Smith A. F. M., Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proceedings-F, Vol. 140, No. 2, 1993, pp. 107-113. DOI: 10.1049/ip-f-2.1993.0015

[10] Li F., Wu L., Shi P., Lim C. C., State Estimation and Sliding Mode Control for Semi-Markovian Jump Systems with Mismatched Uncertainties, Automatica, Vol. 51, 2015, pp. 385-393.

[11] Arulampalam S., Maskell S., Gordon N., Clapp T., A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking, IEEE Transactions on Signal Processing, Vol. 50, No. 2, 2002, pp. 174-188. DOI: 10.1109/78.978374

[12] Doucet A., Johansen A.M., A Tutorial on Particle Filtering and Smoothing: Fifteen years later, handbook of Nonlinear Filtering 2009/12, pp. 656-704.

[13] Imtiaz S. A., Roy K., Huang B., Shah S. L., Jampana P., Estimation of States of Nonlinear Systems using a Particle Filter, In IEEE International Conference on Industrial Technology, ICIT 2006, December, pp. 2432-2437.

[14] Kozierski P., Lis M., Ziętkiewicz J., Resampling in Particle Filtering – Comparison, Studia z Automatyki i Informatyki, Vol. 38, 2013, pp. 35-64.

[15] Doucet A., Freitas N., Gordon N., Sequential Monte Carlo Methods in Practice, Springer-Verlag, New York 2001, pp. 225-246. DOI: 10.1007/978-1-4757-3437-9

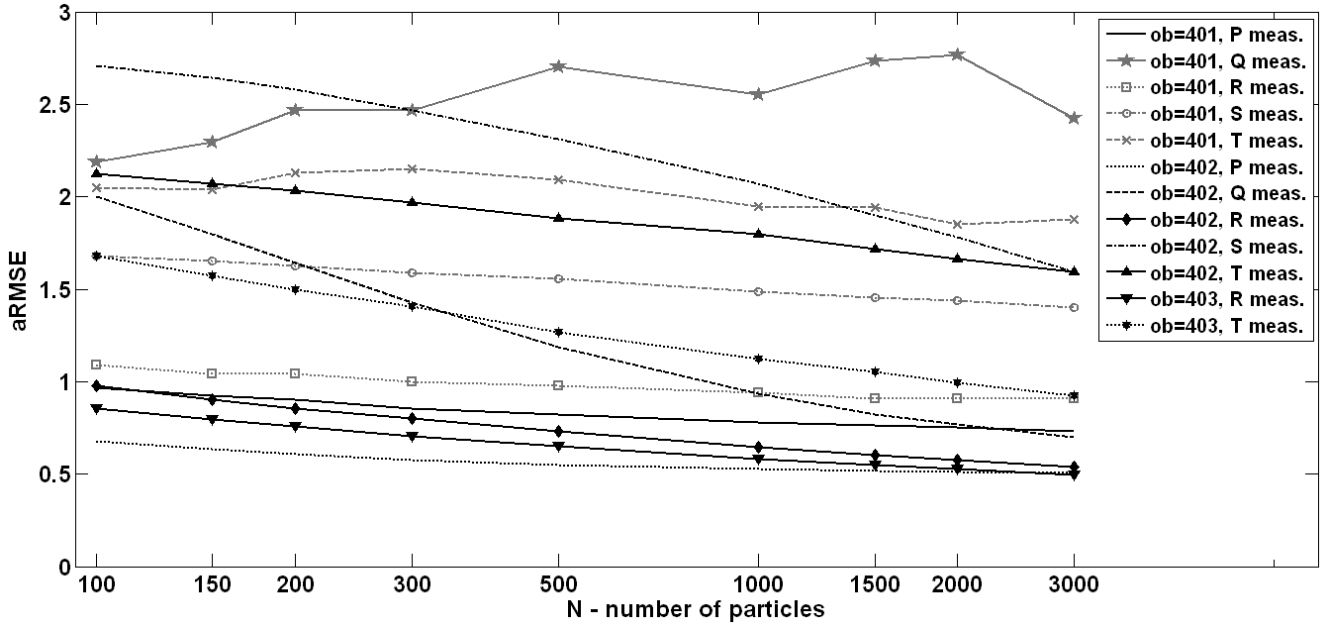


Fig. 5 Obtained results

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1079–1083
DOI: 10.15439/2018F185
ISSN 2300-5963 ACSIS, Vol. 15

UAV downwash dynamic texture features for terrain classification on autonomous navigation

J. P. Matos-Carvalho, José M. Fonseca, André Mora
Computational Intelligence Group of CTS/UNINOVA, FCT, University NOVA of Lisbon
Email: jp.carvalho@uninova.pt, jmf@uninova.pt, atm@uninova.pt

***Abstract*—The information generated by a computer vision system capable of labelling a land surface as water, vegetation, soil or other type, can be used for mapping and decision making. For example, an unmanned aerial vehicle (UAV) can use it to find a suitable landing position or to cooperate with other robots to navigate across an unknown region. Previous works on terrain classification from RGB images taken onboard of UAVs shown that only static pixel-based features were tested with a considerable classification error. This paper proposes a robust and efficient computer vision algorithm capable of classifying the terrain from RGB images with improved accuracy. The algorithm complement the static image features with dynamic texture patterns produced by UAVs rotors downwash effect (visible at lower altitudes) and machine learning methods to classify the underlying terrain. The system is validated using videos acquired onboard of a UAV.**

***Keywords*—Image processing, Texture, Machine Learning, Terrain Classification, UAV**

I. Introduction

Nowadays, due to UAVs' higher availability and capabilities, there is a research trend to explore innovative applications of UAVs useful to the society. They are having a major impact on search and rescue missions, in logistics, in precision agriculture, among other applications. Key issues are to provide a safe and reliable operation and to perceptionate the surrounding area. This latter, within this paper, will be to identify the underlying terrain. Terrain classification is a crucial functionality for a wide range of autonomous vehicles [1]: either for ground vehicles to avoid water bodies, aerial vehicles to determine suitable landing areas, or surface vehicles to detect safe passageways. As further explained in section II, several approaches have been used for terrain classification. However, there is still margin for improving accuracy by extracting more complex image features. When at lower altitudes, UAV's rotors downwash effect create singular image texture patterns depending on the type of terrain, which can be used to differentiate them.

The main goal of this paper is to propose a computer vision algorithm that using RGB images captured by a camera onboard of a UAV is capable of classifying a terrain by analysing static image features (colour and texture) and rotors downwash effect on the underlying surface. There are several issues that must be addressed in order to achieve this goal, namely: Which terrains can be more accurately classified using the downwash effect? Which are the texture and motion patterns of each terrain (water movement for example)? Which static and dynamic image features can be extracted to classify the terrain? To address these challenges, new optimization procedures and techniques will be proposed in this paper, aiming the best possible performance.

This paper is structured with six sections starting with an introductory section and followed by a presentation of related works. In the experimental setup section the system background, namely the hardware and the terrain types, are described. On the Terrain Classification Method the system architecture, the static and dynamic texture features and the machine learning classifier will be presented. The article finishes with the experimental results and drawn conclusions.

II. Related Work

UAVs (Unmanned Aerial Vehicle) play an important role on the new generation of information technology and is predicted to have a major impact in the human life in the near future [2]. One of the areas is in computer vision, where it is possible to acquire, process, analyse and understand aerial images. Many researchers have proposed terrain classification systems based on features derived from colour information [3], texture patterns [4], [5], [6] and from additional sensors, as is the case of laser scan systems [7], [8], [9]. Although many of these algorithms are for terrestrial unmanned ground vehicles, currently there is a shift towards UAVs, where the visual features have wider importance.

One of the most recent works of terrain detection and classification is presented in [10]. The authors use the concept of optical flow to detect the water texture direction in images acquired by an RGB camera onboard of a UAV. From the directions of the textural features, the algorithm determines if the terrain, where the UAV is flying over, is water or non-water. One of the problems identified is that the UAV must be stable over the target while identifying the type of terrain, which, in the best case, takes four seconds to execute. Another reason that requires the UAV to be stand still during calculations is that the computer vision algorithm does not compensate the UAV movement. Thus, when the directions of the features are calculated, the results do not represent the reality.

A classification method using colour features was proposed in [3]. The proposed method converts a RGB image into an image entitled "normal RGB", where each pixel is divided by the square root of the three colour channels. Thus, each terrain will emphasize the colour that represents it (for example, green for vegetation). The proposed method was limited due to the fact that it varies significantly with illumination.

Laser scanners have proven to be important to distinguishing between land and water as presented in [7], [8] and [9].

However, in low water depths the laser sensor produces incorrect results, due to the fact that it captures reflections from the seabed and misclassifies it as non-water terrain. Therefore, this laser scan approach, by itself, reveals to be insufficient and requires additional equipment.

III. Experimental Setup

The dynamic of different terrains when exposed to wind provoke singular texture patterns that can be used in their identification. In this paper we study the importance of static image features, such as colour and texture, when compared with the dynamic features exhibited by the downwash effect, for terrain classification.

In this work three different terrain types (water, vegetation and sand), which can benefit from the downwash effect for their identification (Figure 1) were identified. It can be seen that the downwash effect produces: on water a circular dynamic texture; on vegetation a linear spread from inside outwards; and on sand it is almost stable or it moves outwards.

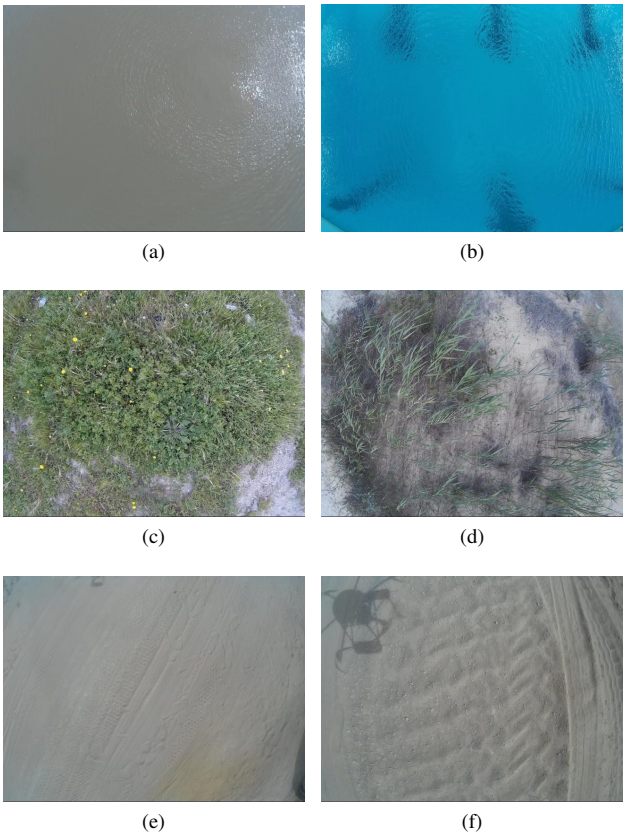


Fig. 1. Examples of terrain types: water (a)(b); vegetation (c)(d); and sand (e)(f).

IV. Terrain Classification Method

If different types of terrain behave differently when exposed to UAV rotors downwash effect, then it should be possible to obtain unique information for their identification. Based in this research hypothesis, it is possible to obtain some conclusions. When exposed to the downwash effect, water particles' movement is always greater than in vegetation and sand terrains. Also, regarding static texture, usually vegetation has a more rough texture than sand or water terrains; water only presents roughness when exposed to wind and downwash effect; and sand (fine grains) has a lower roughness. It can also be seen that sand depends on the patterns already in the terrain, showing usually a more irregular texture (figures 1.e and 1.f) when compared with water that shows a unique signature and regular texture when exposed to wind (figures 1.a and 1.b).

A. System Architecture

The proposed system architecture to classify the terrain using texture information is shown in Figure 2. As previously identified in sections I and IV, two texture features are proposed to classify the terrain, namely, static and dynamic textures. At this stage it were also assessed the features that can be computed in parallel, in order to speedup the system execution time.

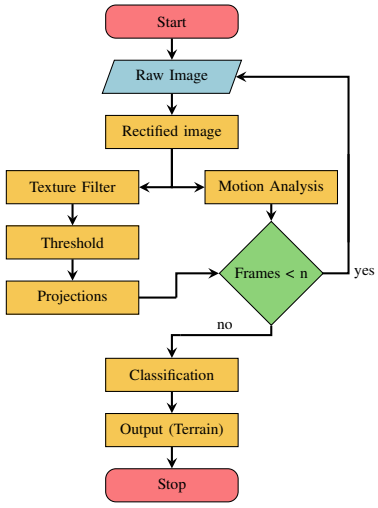


Fig. 2. Proposed system architecture.

Five main processes were identified in the architecture (figure 2), namely:

- **Rectified Image:** Performs lens geometrical corrections;
- **Texture Filter:** Extracts terrain's static textural information using Gabor filters;
- **Threshold:** A thresholding is applied to the static texture image to highlight the terrain roughness;
- **Projections:** Vertical and horizontal projections were applied to the thresholded image, extracting unique features that help differentiate the different types of terrains;
- **Motion Analysis:** Extracts information from dynamic textures. Optical flow and thresholding techniques are used to identify the moving parts;
- **Classification:** The extracted features are used as inputs of an automatic classified to identify the type of terrain. Machine learning techniques already proved to be efficient for terrain classification [11], [12], [13].

B. Static Textures

This section presents the proposed method for extracting terrain's static textures, based on the Gabor filter to be able to choose multiple texture directions. This filter is the impulse response formed by a multiplication of a sinusoidal signal with a Gaussian envelope function and can be computed using the following complex equation:

$$G(x,y,\lambda,\theta,\psi,\sigma,\gamma) = e^{\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)} e^{\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right)} \quad (1)$$

Its real and an imaginary components can be obtained by equations 2 and 3, respectively:

$$G(x,y,\lambda,\theta,\psi,\sigma,\gamma) = e^{\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)} \cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (2)$$

$$G(x,y,\lambda,\theta,\psi,\sigma,\gamma) = e^{\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)} \sin\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (3)$$

where:

$$x' = x\,\cos(\theta) + y\,\sin(\theta) \quad (4)$$

$$y' = -x\,\sin(\theta) + y\,\cos(\theta) \quad (5)$$

These equations (1, 2 and 3) require as input parameters:

- **x and y:** Filter coordinates, where x represents the columns and y the rows;
- **Lambda** (λ): Represents the sinusoid's wavelength;
- **Theta** (θ)**:** Defines the Gaussian envelope orientation;
- **Psi** (ψ)**:** Symbolizes the phase offset;
- **Sigma** (σ)**:** Describes the Gaussian envelope size;
- **Gamma** (γ)**:** Reflects the shape of the ellipse in the gabor filter space.

In this work we used only the real component of the Gabor function (equation 2). After obtaining the multiplication of a Gaussian with a sinusoidal function, i.e. the kernel of the filter, it will be convolved with the original image (equation 6). The result of the Gabor filter applied over a water surface is presented in Figure 3.

$$f[x,y] * g[x,y] = \sum_{-n1}^{n1}\sum_{-n2}^{n2} f[n1,n2]\cdot g[x-n1,y-n2] \quad (6)$$

As can be seen in figure 3, it is possible to obtain the texture of a water-type terrain when it is affected by the downwash effect of the UAV. From the binarized image, a vertical projection was made to see the singular features of this terrain type (Figure 4).

From the observed vertical projection of water type terrain (figure 4) it can be seen that it produces an undulatory effect with a local minimum in the centre of the downwash. This effect in water type terrains is due to the lower roughness

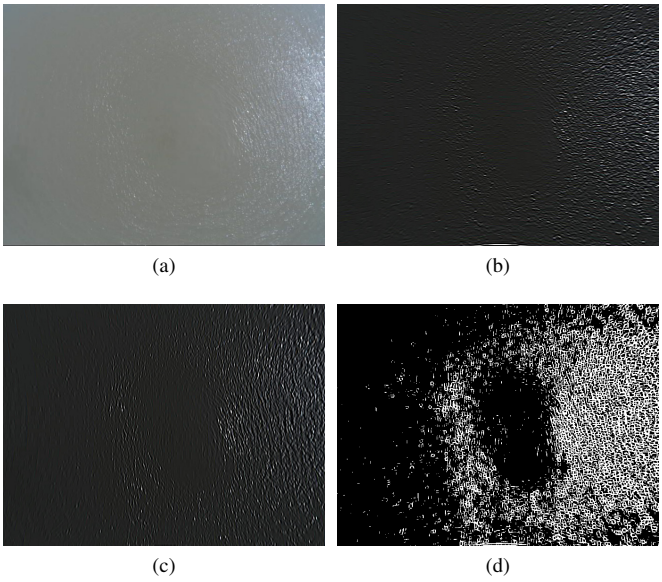


Fig. 3. Example of a static texture extraction: a) Raw image; b) c) Convolution with the Gabor filter θ-0 degrees (b) and θ-90 degrees (c); d) Sum of images b) and c) after thresholding.

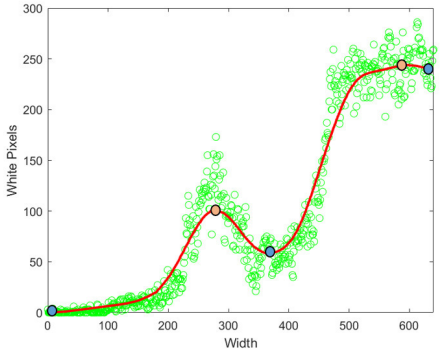


Fig. 4. Vertical projection of the example in Figure 3.d.

in the centre of the downwash. However, due to the water movement, around the centre a higher roughness is observed (white pixels in the binarized image in figure 3.d). The next step was to translate this observed feature into a computational model.

By calculating the local maxima and minima of the vertical projection in figure 3.d, it is possible to calculate a line (red line in Figure 4) that most closely approximates these points. A polynomial regression was used.

After obtaining this smoothed projection, new local minima and maxima are calculated and used to obtain two features: Area measured between the local minimum and its respective two local maxima; and Integral between the local minimum and its two respective local maxima. The first has the advantage of being relative to minima and maxima values, while the integral gives an absolute value and will vary for lower and higher roughness.

C. Dynamic Textures

This section presents the proposed method for extracting dynamic terrain textures.

As mentioned in section IV, water-type terrain only exhibit dynamic texture when exposed to the downwash effect. However, in spite having a dynamic texture, when analysing the optical flow it is never stronger than the dynamic observed for sand and vegetation. As referred in section III, the optical flow method can calculate the distance travelled by block matching features in a given frame sequence. In this paper, the Farneback algorithm [14] was used to detect the movement of these features. One of the advantages to using the Farneback algorithm is the direct flow, F_d , return of features between two frames.

With the obtained flow is then used to calculate the distance travelled (trajectory) by each feature in a sequence of frames:

$$Travel_{distance} = \sum_{i=2}^{n} \sqrt{A_x(i) + B_y(i)} \tag{7}$$

where:

$$A_x(i) = \left[x_1 - x_{i-1} + F_{dx}\right]^2 \tag{8}$$

$$B_y(i) = \left[y_1 - y_{i-1} + F_{dy}\right]^2 \tag{9}$$

and x_i and y_i are the positions in x and y in the most recent frame (n), x_1 and y_1 are the initial positions ($n = 1$) and F_{dx} and F_{dy} are the flow displacements between frames n and $n-1$. We used normalized x and y coordinates for the calculations.

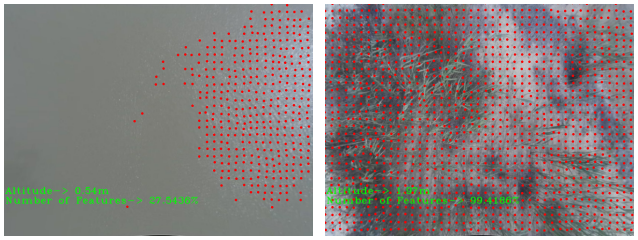
To eliminate features that did not move or were almost static in a sequence of frames, we filtered those not exceeding a pre-defined empiric threshold (1%). Then, knowing the maximum number of features, we calculate the percentage of dynamic features that appear in the image (equation 10). An example is shown in Figure 5.

$$Dynamic_{feature} = \frac{filtered\ features}{Total\ features} \cdot 100\% \tag{10}$$

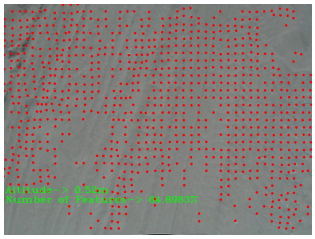
D. Classification

To increase certainty and automate the classification of the type of terrain, a machine learning technique was used, namely a feed-forward neural network (NN). The architecture of the designed neural network, was composed by two layers, a hidden layer with 10 neurons and an output layer with 3 neurons (water, vegetation and sand). A sigmoidal function was used as activation function and the final classification was derived from the output neuron with highest activation value.

The training dataset was composed by 251 samples, from which 70% were for training, 15% for testing and 15% for validation. After training the NN, it was obtained 92.9% accuracy with the training set and 93.8% with the test dataset.



(a) (b)



(c)

Fig. 5. Dynamic textures detection by Farneback algorithm and distance travelled calculation. a) water; c) vegetation and d) sand.

V. Experimental Results

To validate the proposed static and dynamic texture features for terrain classification, a total 251 frames from several types of terrains were used to validate the proposed system. From these 90 frames were for water, 88 frames for vegetation and 73 frames for sand.

Regarding the static texture feature the area and integral were calculated and displayed in Figure 6. It is possible to observe a clear separation between water, vegetation and sand, even with some outliers. In water type terrain, the three clusters that can be noticed for the integral feature, were mainly due to different water environments (lake and pool).

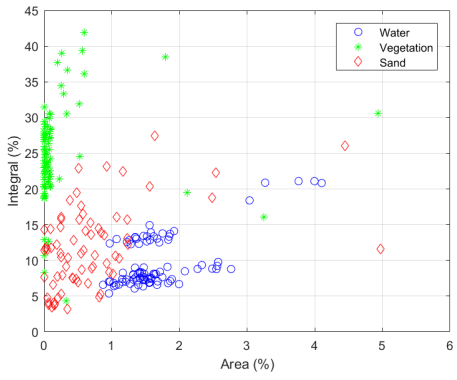


Fig. 6. Static Texture - Relation in area with respect to the integral of minimum and maxima locals.

To validate the dynamic texture feature it was calculated in a three frame period ($n = 3$) and plotted against the area feature from the static texture. This feature shows the same discriminant level to separate the different terrains. From Figure 7, it can be seen that water type terrain obtained a lower dynamic texture value ($< 45\%$), which can due to a higher

concentration of these dynamic features in the downwash centre and outside hasn't exceed the threshold. Sand and vegetation shown a more uniform pattern, obtaining a higher number of features. On average, sand presents a percentage between 55% to 90%. Finally, vegetation with a percentage of features between 90% and 100%, is the terrain with highest dynamic texture, i.e., moving features.

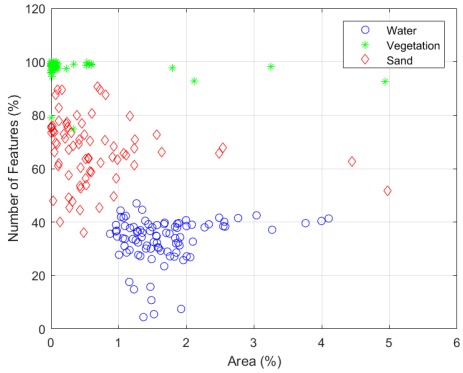


Fig. 7. Dynamic Texture - Relation in number of features with respect to the integral of minimum and maxima locals.

Finally, these features were extracted from the figures 1.a-f and shown to the neural network classifier, which outputted the automated terrain classification. The extracted features and the classification result is shown in Table I. As expected the proposed features and classification method, shown good results by classifying correctly all the six examples, reinforcing the idea that a combination of static and dynamic texture can be used to automatically extract terrain type from RGB images.

TABLE I. EXPERIMENTAL RESULTS

Figure	Static Texture		Dynamic Texture	Classification
	Area (%)	Integral (%)	Number of Features (%)	
1	1.44	8.32	32.88	water
2	1.55	7.59	36.70	water
3	0.01	27.42	98.32	vegetation
4	0.05	23.71	98.69	vegetation
5	0.24	14.67	63.74	sand
6	0.07	4.07	59.30	sand

VI. CONCLUSIONS

The main objective of this paper was to design a computer vision system capable of extracting static and dynamic image features, such as optical flow and texture features, to identify the type of terrain with improved accuracy by taking advantage of the the UAV's rotors downwash pattern effect. For this, it was necessary to conduct a research into detection methods already implemented and of interest to this work.

Texture features, such as Gabor filtering (static textures) and optical flow (dynamic textures), were studied to improve terrain classification aiming the best possible performance.

We emphasize that by implementing the static textures filter, vegetation-like terrains were found to have a higher texture than sand and water type terrains. On the other hand, water-type terrain, also presents a singular characteristic due to the downwash effect provoked by the UAV, which can be decisive to different it from other terrain types.

ACKNOWLEDGMENT

This work was partially funded by FCT Strategic Program UID/EEA/00066/203 of the Center of Technologies and System (CTS) of UNINOVA – Institute for the Development of new Technologies. In last, this work was not possible without the support and commitment of several fellow colleagues and friends, namely: Ricardo Mendonça, Francisco Marques, André Lourenço, Eduardo Pinto and José Barata.

REFERENCES

[1] Pinto, E., Marques, F., Mendonca, R., Lourenco, A., Santana, P., & Barata, J. (2014). An autonomous surface-aerial marsupial robotic team for riverine environmental monitoring: Benefiting from coordinated aerial, underwater, and surface level perception. 2014 IEEE International Conference on Robotics and Biomimetics, IEEE ROBIO 2014, 443–450. https://doi.org/10.1109/ROBIO.2014.7090371

[2] Bestaoui Sebbane, Y. (2018). Intelligent Autonomy of UAVs : Advanced Missions and Future Use. CRC Press.

[3] Ebadi, F., & Norouzi, M. (2017). Road Terrain detection and Classification algorithm based on the Color Feature extraction. In 2017 Artificial Intelligence and Robotics (IRANOPEN) (pp. 139–146). IEEE. https://doi.org/10.1109/RIOS.2017.7956457

[4] Feng, Q., Liu, J., & Gong, J. (2015). UAV Remote sensing for urban vegetation mapping using random forest and texture analysis. Remote Sensing, 7(1), 1074–1094. https://doi.org/10.3390/rs70101074

[5] Khan, Y. N., Komma, P., Bohlmann, K., & Zell, A. (2011). Grid-based visual terrain classification for outdoor robots using local features. IEEE SSCI 2011: Symposium Series on Computational Intelligence - CIVTS 2011: 2011 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems, 16–22. https://doi.org/10.1109/CIVTS.2011.5949534

[6] Pietikäinen, Matti, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. 2011. Computer Vision Using Local Binary Patterns. Vol. 40. Computational Imaging and Vision. London: Springer London. https://doi.org/10.1007/978-0-85729-748-8.

[7] Yan, W. Y., Shaker, A., & El-Ashmawy, N. (2015). Urban land cover classification using airborne LiDAR data: A review. Remote Sensing of Environment, 158, 295–310. https://doi.org/10.1016/j.rse.2014.11.001

[8] Wallace, L., Lucieer, A., Malenovsky, Z., Turner, D., & Vopěnka, P. (2016). Assessment of forest structure using two UAV techniques: A comparison of airborne laser scanning and structure from motion (SfM) point clouds. Forests, 7(3), 1–16. https://doi.org/10.3390/f7030062

[9] GruszczynSki, Wojciech, Matwij, Wojciech, Cwiakała, P. (2017). Comparison of low-altitude UAV photogrammetry with terrestrial laser scanning as data-source methods for terrain covered in low vegetation. ISPRS Journal of Photogrammetry and Remote Sensing, 126, 168–179. https://doi.org/10.1016/j.isprsjprs.2017.02.015

[10] Pombeiro, R., Mendonca, R., Rodrigues, P., Marques, F., Lourenco, A., Pinto, E., Barata, J. (2015). Water detection from downwash-induced optical flow for a multirotor UAV. In OCEANS 2015 - MTS/IEEE Washington (pp. 1–6). IEEE. https://doi.org/10.23919/OCEANS.2015.7404458

[11] Mora, A.; Santos, T.M.A.; Łukasik, S.; Silva, J.M.N.; Falcão, A.J.; Fonseca, J.M.; Ribeiro, R.A. (2017) Land Cover Classification from Multispectral Data Using Computational Intelligence Tools: A Comparative Study. Information 2017, 8, 147. https://doi.org/10.3390/info8040147

[12] Heung, Brandon, Hung Chak Ho, Jin Zhang, Anders Knudby, Chuck E. Bulmer, and Margaret G. Schmidt. 2016. "An Overview and Comparison of Machine-Learning Techniques for Classification Purposes in Digital Soil Mapping." Geoderma 265 (March): 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

[13] Giusti, Alessandro, Jerome Guzzi, Dan C. Ciresan, Fang-Lin He, Juan P. Rodriguez, Flavio Fontana, Matthias Faessler, et al. 2016. "A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots." IEEE Robotics and Automation Letters 1 (2): 661–67. https://doi.org/10.1109/LRA.2015.2509024.

[14] Farneback, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. Lecture Notes in Computer Science, 2749(1), 363–370. https://doi.org/10.1007/3-540-45103-X_50

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1085–1089
DOI: 10.15439/2018F99
ISSN 2300-5963 ACSIS, Vol. 15

MDPC decoding algorithms and their impact on the McEliece cryptosystem

Artur Janoska
Military University of Technology
Institute of Mathematics and Cryptology
ul. Urbanowicza 2, 00-908 Warsaw, Poland
Email: artur.janoska@wat.edu.pl

***Abstract*—In recent years, research has been conducted aimed at finding alternative asymmetric systems other than traditional systems such as RSA (Rivest–Shamir–Adleman algorithm) and ECC (Elliptic-curve cryptography). One of the most promising is code-based cryptosystems since their security is based on well-known NP-hard problems. Especially, the most interesting cryptosystem is system proposed by Misoczki et al. based on QC-MDPC codes which use the modified BitFlip algorithm as the decoding algorithm. This work presents a comparison of different variants of MDPC decoding algorithms and their impact on the cryptosystem. We present a complete analysis of modification of this algorithm and new results of the likelihood of correct word decoding for security systems which ensure security level 2^{128} and 2^{256}.**

I. Introduction

ASYMMETRIC cryptography is one of the most important cryptographic mechanism currently in use. It provides a number of options such as secure cryptographic keys exchanging over a public channel, digital signature and secure messaging. Most of those systems are based on number theory's problems, such as factorization of a large number (RSA) and the discrete logarithms in an elliptic curve. an important limitation of the indicated security systems is the fact that using Shor's algorithm [10] can solve these problems on the quantum computer so it means that they are not safe in the long-term security.

If a large-scale quantum computer is built for, that algorithm will become one of the most useful tools in cryptoanalysis, especially in public key cryptography. Nowadays there is a lot of interest in alternative systems that are not based on numerical problems. From that reason, NIST has initiated a process to solicit, evaluate, and standardize one or more quantum-resistant public-key cryptographic algorithms.

Promising alternative is code-based crypto algorithms those security is based on the General Decoding Problem were extensively studied for their usage in small and embedded systems. The first cryptosystems relying on coding theory was proposed by J.R. McEliece in 1978 [8]. It uses binary Goppa codes as a basis for the construction. This system presents many advantages: it is very fast for both encryption and decryption and the best-known attacks are exponential in the length of the code. Although it has proved resistance against all known attacks, it has one big flaw: the size of the public key. The public key size for the original parameters that provide security 2^{80} proposed by McEliece has 67072 bytes, against 256 bytes of a 1024-bit modulus instance of RSA.

In order to reduce key sizes, several alternative approaches for code-based cryptography were proposed. One idea to shorten key is to use codes with particular structures.Unfortunately, most of the modifications of the code structure result in the cryptosystem sensitivity to the so-called structural attacks. Such attacks aim to exploit the hidden structure, in order to recover the private key.

Another idea is to replace the binary Goppa codes with another linear code which generates matrix rolled representation and has no vulnerability to structural attacks. One of the promising algorithms was proposed by Misoczki et al. [9] They proposed to modify the McEliece cryptosystem by using a quasi-cyclic Moderate Parity Check (QC-MDPC). They achieve a very fast cryptosystem with a relatively small public key size (about 4801 b with secure level 2^{80}), lightweight implementation and still preserving the security properties of the cryptosystem. Unfortunately, in 2016 a very powerful attack was shown. The attack can recover a secret key for a system with security 2^{80} using a chosen ciphertext attack [4]. It means that this system is impractical for the encryption scheme, but still can be used as secure Key Encapsulation Mechanism.

The next section will present the motivation for the research. In the third section basic definitions and necessary statements will be presented. In the 4th section, the McEliece algorithm based on the QC-MDPC codes will be presented and in the fifth section, the decoding algorithms will be presented. Section 6 contains a description of the assumptions during the research. Section 7 contains the obtained results and their analysis. The 8th section contains a summary of completed research.

II. Motivation

The McEliece system based on QC-MDPC codes is a very interesting contribution because it has a very good performance on embedded systems and limited resources. MDPC code extends the concept a low-density parity-check (LDPC) codes [3] by using the parity check matrix with moderated sparse. Unfortunately, this leads to a significantly degraded error correction performance. However, in cryptography, We

are not interested in correcting a large number of errors, but only the number ensuring an adequate level of security.

The probabilistic decoding algorithm used in the cryptosystem is known as the modified bit flipping algorithm. During the last few years this probabilistic aspect was very intensively examined by community [1], [7]. The results of this research increased error-correcting capability for the tested systems for security 2^{80}. The quality of the solution was measured by the decoding failure rate and the number of iterations required to decrypt the message.

In the submission for the NIST standardization project, an IND-CPA[1] secure ephemeral Key Encapsulation Mechanism (KEM) based on the Quasi-Cyclic Moderate Density Parity-Check (QC-MDPC) McEliece encryption knows as QC-MDPC KEM was presented. In this work, we investigated several ways to efficiently decode erroneous MDPC codewords, especially for such codes as were proposed for the McEliece cryptosystem with security 2^{128} and 2^{256}. Additionally, we have proposed and evaluated a new way of choosing the parameter b for systems with security 2^{256}. This optimization leads to reduced decoding failure probability and fewer decoding iterations.

III. Preliminaries

In order to unify the notations and definitions, we will present a few definitions of the necessary concepts. All considerations will be conducted on finite field $\mathcal{F}_2$

The **Hamming weight** (or simply weight) of vector $x \in \mathcal{F}_2^n$ is the number of nonzero components denoted as $wt(x)$

A **binary** (n, r)**-linear code** $\mathcal{C}$ of length n and dimension r is an r-dimensional vector subspace of $\mathcal{F}_2^n$. It is spanned by the rows of a matrix $\mathbf{G} \in \mathcal{F}_2^{r\times n}$, called a generator matrix of $\mathcal{C}$. Also, it is the kernel of a matrix $\mathbf{H} \in \mathcal{F}_2^{(n-r)\times n}$ called a parity-check matrix of $\mathcal{C}$. The codeword $c \in \mathcal{C}^n$ of a vector $m \in \mathcal{F}_2^r$ is $c = mG$. The syndrome $s \in \mathcal{F}_2^{n-r}$ of a vector $\epsilon \in \mathcal{F}_2^n$ is $s = \mathbf{H}e^T$

An (n, r)-linear code is a **quasi-cyclic code** (QC) if there is some integer n_0 such that every cyclic shift of a codeword by n_0 places is again a codeword. Additionally when $n = n_0 p$ for some integer p, it is possible to have generator and parity check matrices composed by $p \times p$ circulant blocks which are completely described by their first row (or column).

An (n, r, w)-LDPC or MDPC code is a linear code of length n, dimension r which admits a parity-check matrix of constant row weight w. LDPC and MDPC codes differ in the magnitude of the row weight w. We assume for MDPC codes row weight whose scale is $O(\sqrt{n \log n})$. On the other hand, the constant row weight is usually less than 10 for LDPC.

A. MDPC and QC-MDPC code

An random (n, r, w)-MDPC code is easily generated by selecting a random parity-check matrix $\mathbf{H} \in \mathcal{F}_2^{r\times n}$ of row weight w. We only have to check that the rightmost $r \times r$ block is full rank. If not, wecan swap a few columns to get a full rank matrix. The general definition of MDPC codes can be found in [9]. For the purpose of this article, construction using $n_0 = 2$ will be discussed.

The (n, r, w)–QC-MDPC codes where $n = 2p$ and $r = p$. So then the parity check matrix has the form

$$H = [H_0|H_1]$$

where H_i is a $r \times r$ circulant block. To define the parity-check matrix $\mathbf{H}$ we pick up a random first row of weight w and the other $r - 1$ rows are obtained from $r - 1$ quasi-cyclic shift of this first row.

A generator matrix $\mathbf{G}$ in the row reduced echelon form can be easily derived from the H_i's blocks. Assuming that the block H_1 is non-singular (which particularly implies row h_i of matrix H_1 has $wt(h_i)$ odd) we construct a generator-matrix

$$G = \left[\quad I \quad \middle| \left(H_1^{-1} \cdot H_0\right)^T \right]$$

IV. QC-MDPC McEliece variant

In order to define a McEliece variant based on t-error correcting (n, r, w)–QC-MDPC code we need to fix some MDPC decoding algorithm equipped with the knowledge of $\mathbf{H}$ (denoted as Ψ_H). Encryption, decryption and key generation for the QC-MDPC McEliece variant cryptosystem are defined as follows.

Key Generation. The key Generation procedure consists of two steps. First we generate a parity-check matrix $\mathbf{H} \in \mathcal{F}_2^{r\times n}$ of a t-error-correcting(r, r, w)–QC-MDPC code by choosing the first row of the parity-check matrix. The second step is to generate the corresponding generator matrix $\mathbf{G} \in \mathcal{F}_2^{r\times n}$ in the row reduced echelon form.

The public-key of this system is the tuple $(\mathbf{G}, t)$ and the private-key is matrix $\mathbf{H}$.

Encryption. In order to encrypt message $m \in \mathcal{F}_2^r$ we need to generate random vector $\epsilon \in \mathcal{F}_2^n$ of $wt(\epsilon) \leq t$ and compute

$$x \leftarrow m\mathbf{G} + \epsilon$$

where $x \in \mathcal{F}_2^n$ is a ciphertext.

Decryption. To decrypt $x \in \mathcal{F}_2^n$ into $m \in \mathcal{F}_2^r$ we compute

$$m\mathbf{G} \leftarrow \Psi_H(x)$$

If the generated matrix $\mathbf{G}$ is in the row reduced echelon form, we extract the message m from the first r position of $m\mathbf{G}$.

A. Security and Parameter Selection

Theoretical security of the QC-MDPC McElice cryptosystem has been presented [9]. In particular, an analysis of the safety and impact a quasi-cyclic structure on the security was presented. Recently, new attacks on system using those codes have been proposed. The most important is very powerful attack using a quasi-cyclic form of the parity check matrix [4]. The attack leverages the fact that there is some probability, termed the Decoding Failure Rate (DFR), that the decoding may fail to compute the errors.

Parameters for the examined systems are based on analyzes carried out in the work [11]. The suggested parameters are presented in Table I. The tests have been carried out using these values.

[1]Indistinguishability under chosen-plaintext attack

Table I
SUGGESTED PARAMETER SETS FOR CLASSIC SECURITY AND QUANTUM SECURITY [11]

System	Classical security	Quantum Security	n	r	w	t	Public Key size
McEliece	80	58	9602	4801	90	84	4801
McEliece	128	86	19714	9857	142	134	9857
McEliece	256	154	65542	32771	274	264	32771

V. DECODING ALGORITHMS

Decoding in the McEliece cryptosystem is a more complex operation than encryption. For lightweight embedded systems, the best solution seems to be the variant of the Gallager's bit flipping algorithm [3], dedicated to the LDPC code. Positive aspects of this solution are its simplicity and lack of floating-point arithmetic. On the other hand, the disadvantage of this solution is that we find the codeword with some probability determined by threshold b, which will be later discussed.

The algorithm works as follows. At each iteration, the number of unsatisfied parity-check equations associated to each bit of the message is computed. Each bit associated with more than b unsatisfied equations is flipped and the syndrome is recomputed. This process is repeated until either the syndrome becomes zero or after a maximum number of iteration is reached. We name this algorithm Algorithm 1.

The algorithm has complexity $O(nwI)$, where I stands for the average number of iterations. The most important difference from the algorithm proposed by Gallager is how threshold b is determined. The first proposition was to precompute thresholds for each iteration i. The threshold is set as the maximum number of unsatisfied parity-check equations $b = max(\sigma_i)$. In [9] the authors suggest to use $b = max(\sigma_i) - \delta$, for some small δ (suggested in [9] is $\delta \approx 5$). In [7] the authors propose incrementing the precomputed thresholds by $\Delta = 1$ and in the case of a decoding failure increase it to $\Delta = \Delta + 1$. Decoding is restarted with the adapted Δ until reaching a predefined $\Delta_{max} = 5$. The survey of this optimalization is included in Table II

Algorithm 2 Modified Gallager's bit flipping algorithm

Input: $x \in \mathrm{F}_2^n$
Require: $H \in M_r^n, r_{max} \in \mathcal{Z}_+$
Output: $m \in C$ lub $error$
    $s \leftarrow xH^T$
2: **for** $r \in \{0, \ldots, r_{max-1}\}$ **do**
        for $i \in \{0, \ldots, n-1\}$ **do**
4: $\sigma_i \leftarrow \langle s, h_i \rangle \in Z$ $\star$
            if $\sigma_i \geq b$ **then**
6: $x_i \leftarrow x_i \oplus 1$ $\star\star$
                $s = s \oplus h_i$
8: **end if**
        end for
10: **if** $s = 0^r$ **then**
            return x
12: **end if**
    end for
14: **return** $error$

$\star$ $\langle \cdot, \cdot \rangle$ – means scalar product of two vectors, h_i – means i-th column of matrix H
$\star\star$ x_i – means i-th position in vector x

A. Effective implementation

Code-based systems allow lightweight and effective implementations in devices with limited hardware resources and all operations are much less complex when compared to the other post-quantum systems.

Encryption involved simple operations such as vector-matrix multiplication followed by an addition.

Decoding is a more complex operation, but it is possible to reduce the cost of this operation by some improvements. In [7] they propose to improve the syndrome computation. They observe that if i-the bit of ciphertext is flipped, the new syndrome is equal to the old syndrome accumulated with row h_i of the parity check matrix. The authors suggest updating the syndrome directed after flipped i-th bit. This modified algorithm is presented as algorithm 2.

Key generation is a very simple operation, as it mainly uses a pseudorandom generator. In this paper, we do not investigate selecting pseudorandom generators appropriate for the code-based systems. Some research focused on software implementations in this area is presented in [2].

VI. EXPERIMENTAL SETUP

In this work we focus on the parameter selection of decoding algorithms for the proposed McEliece systems based on QC-MDPC codes and corresponding to other security levels according to the Table II a total 10000 random decoding trials were evaluated on a computer equipped with an Intel Core i7 2670QM running at 2.20 GHz.

The research was conducted to investigate the impact of choosing the parameter value of the Decoding Failure Rate (DFR) and to examine the number of rounds needed for correct decryption of the ciphertext. The generation of plaintext and error pattern was based on a uniform distribution.

As part of the study, particular attention was paid to decoding algorithms that use the maximum value of the coefficient $sigma_i$ to calculate the b parameter. In this work the results for the following algorithms will be presented: Decoder:

Decoder A Algorithm 1, threshold b value chose using Misoczki method,

Decoder B Algorithm 2, threshold b value chose using Misoczki method,

Decoder C Algorithm 2 with method choosing threshold b proposed in this work with increment δ by two (starting from 5 to 9) every two iterations,

Table II
PROPOSED METHOD OF CHOOSING THRESHOLD b

	Proposed by	Year	Method	Comments
1	Galager [3]	1962	$b_i - const, i \in \{1, \ldots, r\}$	$b_i \in < 28, 26, 24, 22, 20 >$
2	Huffman and Pless [6]	2010	$b = \#max$	$\#max$ means the maximal value σ_i
3	Misoczki [9]	2013	$b = \#max - \delta$	Decrement δ from 5 to 0 for every incorrect decoding
4	Heyse [5]	2013	$b = b_i + \delta$	Decrement δ from 5 to 0 for every incorrect decoding
5	This paper	2018	$b = \#max - \delta$	Increment δ every two iterations by two starting from 5 to 9
6	This paper	2018	$b = \#max - \delta$	similarly to the Misoczki proposal, only the starting point is different for each level of security, starting from 5 to 7 respectively

Decoder D Algorithm 2 with method choosing threshold b proposed in this work similar to the Misoczki proposal, only the starting point is different for each level of security, starting from 5 to 7 respectively.

The method proposed by Huffman and Pless is a special case of the method proposed by Misoczki et al. and, therefore, it will not be considered in the study. [7]

VII. RESULTS

In the beginning, we focused on the analysis of the average distribution of σ_i coefficients depending on the chosen security level. The graph for these values is presented in Figure 1. The average distribution of the parameter σ_i represents the first distribution before any coding algorithm is used. As can be seen, the values of the parameter σ_i for systems with security 2^{80} and 2^{128} are similar to each other while for the 2^{256} security system, the average values σ_i are much higher and more intense.

These are the first symptoms that the decoding parameter δ for the safest system will need to be modified. An interesting relationship is that if we take the average distribution for σ_i for the files corresponding to the word correctly decoded, we can say that it is different depending on the choice of the decoding algorithm. an example of a distribution for the system with security level 2^{128} is shown in Figure 2. The given property can be used to distinguish, which decoding algorithm was used.

The Decode Failure Rates for the tested algorithms are listed in Table III for all levels of security considered in the table I.

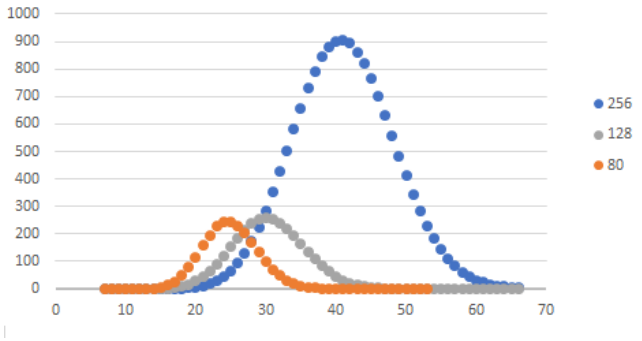


Figure 1. The average distribution of σ_i coefficients depending on the chosen security level.

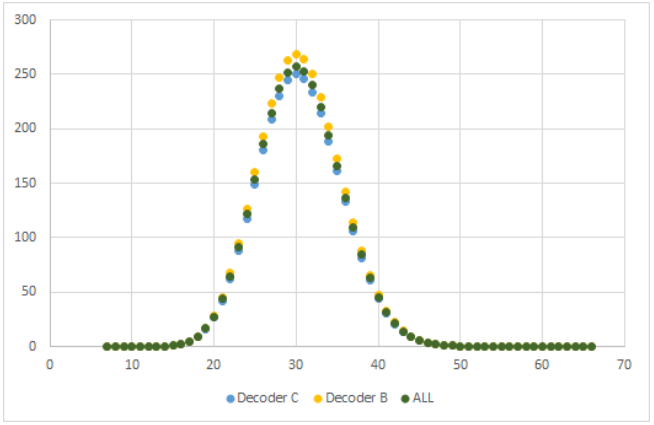


Figure 2. The average distribution of σ_i coefficients depending on the chosen decoding algorithm.

Analysing the results obtained for the Decoder A, which was proposed in the original work, it can be seen that for higher security levels it is not useful. Especially for the level 256, wherein any of the examined cases, the message was properly decoded.

Comparing the two decoders from literature (Decoder a and Decoder B), in Decoder B provides much better decoding failure rate for higher security. However, it still has too high DFR to be practically used. Our proposition to change the δ value for higher security level has a very strong impact on the decoding failure rate. In addition, our solution led to a smaller average number of decoding iterations, respectively 13% and 26% less.

Comparing decoders B, C, D we can see that the decoding failure rate and the number of decoding iterations are not strongly correlated. The improvement of the DFR level does not always result in an improvement in the number of decoding iterations. If we properly manipulate the parameters of decoding algorithm, we can get properties adapted to the specific application.

Additionally, as part of the study, the decoding failure rate was analyzed depending on the distribution of words with the desired Hamming weight. The test assumes that the codewords have the Hamming weight equal to $r/2$. It was noted that the DFR, as well as the number of rounds, increased slightly.

Table III
EVALUATION OF THE PERFORMANCE AND ERROR CORRECTION CAPABILITY OF THE TESTED ALGORITHMS. NOTE, A DFR OF 0 MEANS THAT NO DECODING ERROR OCCURRED DURING OUR EVALUATIONS BUT THE DECODERS ARE STILL PROBABILISTIC.

Name	80		128		256	
	DFR	Round Number	DFR	Round Number	DFR	Round Number
Decoder A	0.00000*	6.30000	0.21400	9.65600	1.00000	–
Decoder B	0.00000*	3.02850	0.00499	6.31087	0.16400	8.69617
Decoder C	0.01600	3.98831	0.07304	5.45662	0,00300	6,72700
Decoder D	0,00000*	3.02850	0.00440	5.54209	0,00000*	6.87500

VIII. CONCLUSIONS

In this work, we examined various variants of the decoding algorithm depending on the choice of the threshold. Additionally, we presented the method of selecting the threshold for codes used in high-security levels systems.

Additionally, as part of the work, an analysis of the possibility of improving the bit-flipping algorithm in applications to MDPC codes was presented.

An interesting fact is that if we use algorithms with a relatively high DFR coefficient, we can distinguish these algorithms based on the analysis of histograms discussed in Section VII. However, it should be noted that when using low-DFR decoding algorithms, the corollary analysis does not apply.

In the light of the achieved results, it can be concluded that the modified Bit Flipping algorithm can be successfully applied to various types of key encapsulation mechanism based on QC-MDPC codes. Especially, for the QC-MDPC-KEM algorithm reported to a process to solicit, evaluate, and standardize one or more quantum-resistant public-key cryptographic algorithms organized by NIST.

In further work I will focus on compare and analysis of other algorithms for decoding QC-MDPC codes, in particular the "One-round Bit Flipping" algorithm and their use in the KEM QC-MDPC system.

REFERENCES

[1] Julia Chaulet and Nicolas Sendrier. “Worst case QC-MDPC decoder for McEliece cryptosystem”. In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE. 2016, pp. 1366–1370.

[2] Nir Drucker and Shay Gueron. “A toolbox for software optimization of QC-MDPC code-based cryptosystems”. In: (2017).

[3] Robert Gallager. “Low-density parity-check codes”. In: *IRE Transactions on Information Theory* 8.1 (1962), pp. 21–28.

[4] Qian Guo, Thomas Johansson, and Paul Stankovski. “A key recovery attack on MDPC with CCA security using decoding errors”. In: *International Conference on the Theory and Application of Cryptology and Information Security*. Springer. 2016, pp. 789–815.

[5] Stefan Heyse, Ingo Von Maurich, and Tim Güneysu. “Smaller keys for code-based cryptography: QC-MDPC McEliece implementations on embedded devices”. In: *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer. 2013, pp. 273–292.

[6] W. Cary Huffman and Vera Pless. *Fundamentals of Error-correcting Codes*. Cambridge University Press, 2010.

[7] Ingo Von Maurich, Tobias Oder, and Tim Güneysu. “Implementing QC-MDPC McEliece Encryption”. In: *ACM Transactions on Embedded Computing Systems (TECS)* 14.3 (2015), p. 44.

[8] Robert J McEliece. “A public-key cryptosystem based on algebraic”. In: *Coding Thv* 4244 (1978), pp. 114–116.

[9] Rafael Misoczki et al. “MDPC-McEliece: New McEliece variants from moderate density parity-check codes”. In: *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE. 2013, pp. 2069–2073.

[10] Peter W Shor. “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”. In: *SIAM review* 41.2 (1999), pp. 303–332.

[11] Atsushi Yamada et al. “QC-MDPC KEM: A Key Encapsulation Mechanism Based on the QC-MDPC McEliece Encryption Scheme - First Round Submission in NIST Post-Quantum Cryptography Standardization”. In: (2017).

Author Index