Towards *Dr Inventor*: A Tool for Promoting Scientific Creativity

D.P. O'Donoghue[1], H Saggion[2], F. Dong[3], D. Hurley[1], Y. Abgaz[1], X. Zheng[3], O. Corcho[4], J.J. Zhang[5], J-M Careil[6], B. Mahdian[7], X. Zhao[8]

1 National University of Ireland Maynooth, Ireland.
2 Universitat Pompeu Fabra, Barcelona, Spain.
3 University of Bedfordshire, UK;
4 Universidad Politécnica de Madrid, Spain
5 Bournemouth University, UK
6 Intellixir, Manosque, France
7 ImageMetry, Prague, Czech Republic
8 Ansmart, Wembley, UK

Abstract

We propose an analogy-based model to promote creative scientific reasoning among its users. *Dr Inventor* aims to find *novel* and potentially *useful* creative analogies between academic documents, presenting them to users as potential research questions to be explored and investigated. These novel comparisons will thereby drive its users' creative reasoning. Dr Inventor is aimed at promoting *Big-C Creativity* and the *H-creativity* associated with true scientific creativity.

Introduction

Reasoning with analogical comparisons is highly flexible and powerful, playing a significant role in the creativity of scientific and other disciplines (Koestler, 1964; Boden, 2009). The role played by various analogies in both helping and (implicitly) hindering scientific progress is discussed by Brown (2003). Dunbar and Blanchette (2001) found that analogies were used extensively by working scientists as part of their day-to-day reasoning, playing significant roles in processes from explanation to hypothesis formation.

This paper discusses some initial work on an analogy-based model (called *Dr Inventor*), which will offer computational creativity as a web service to its users who are practising scientists. *Dr Inventor* is focused on helping research scientists by discovering creative analogical comparisons between academic documents and related sources for their consideration. So Dr Inventor will act as a creativity assistant, while its cognitively inspired architecture also offers one possible model of people thinking creatively.

The Web has become a ubiquitous source of publications, source code, data, research websites, wiki and blogs. These form the *Research Objects* (Belhajjame *et al*, 2012), used by Dr Inventor – a tool for the discovery and presentation of creative analogies between research objects. Dr Inventor is targeted on the *Big-C Creativity* (Gardner, 1993) sought by practising scientists. Indeed, the aspirations of Dr Inventor include supporting analogy driven H-creativity (Boden, 1992; 2009).

Analogies compare a *source* to a *target* problem highlight some latent similarity between them. A creative analogy uses a novel source to bring new and creative possibilities to light. Dr Inventor aims to discover novel analogies between academic resources, bringing unnoticed possibilities out of the shadows. Cognitive studies have shown that exposure to even a single analogical comparison can induce significant differences in peoples response to a given problem (Gick and Holyoak, 1980; Thibodeau and Boroditsky, 2011). This paper is focused on identifying *novelty* and *quality* (Boden, 1992) - essential qualities of creativity:

Baydin's (2012) model generated creative analogs for a given target. CrossBee (Juršič *et al,* 2012) looked for bridging concepts between documents from two given domains of interest. Kilaza (O'Donoghue and Keane, 2012) generated creative analogies but it relied on hand-coded data. Dr Inventor will offer a more complete model of creative analogising and blending (Fauconnier and Turner, 1998 Veale, O'Donoghue and Keane, 2000), addressing a broad range of the aspects of creativity.

Dr Inventor Overview

Dr Inventor will include a multi-phase model of analogy encompasing *representation, retrieval, mapping* and *validation.* It may become the first web-based system that supports the exploration of scientific creativity via a computational approach – offering creativity as a web service to its users, i.e. researchers. Dr Inventor is built upon the vision that technologies have a great potential to enhance the broader discipline of scientific creativity. It will build on technologies, such as information extraction, document summarization, semantic web and visual analytics to exploit the great potential in supplementing human ingenuity.

Dr Inventor will become researchers' personal research assistant by reporting to the researchers on a wide variety of relevant concepts through machine-powered search and visualization. It will assess an input research document through comparison with recognized research approaches and suggest new research ideas to the users in an autonomous manner. Dr Inventor will, to a degree, replicate one mode of human creativity to combine diverse information resources and generate new concepts with unexpected features. The new concepts may come from radical transformations inspired by other semantically distant but analogically similar concepts.

Dr Inventor will be based on computational models of analogical reasoning and conceptual blending. Computational models can arguably offer greater creative ability than human reasoners for at least three specific reasons. Firstly, "problem fixation" frequently acts to limit people's ability to think creatively (Lopez *et al*, 2011). Secondly, people often fail to notice analogies when they are present (Gick and Holyoak, 1980). Thirdly, people often discard useful distant analogies once they have been discovered (Lopez *et al*, 2011). People tend to rate distant analogies as less useful - even if they produce better results. People also suffer from memory limitations, selective thinking, perception limitations, biases, *etc*. A computational model may help to address some of these limitations.

The work of Dr Inventor will synergistically explore techniques for information extraction, document summarization and semantic identification to support the analysis of research objects and the generation of new ontologies for scientific creativity. Interactive visual analytics will be applied to support a user centred creative process. The outcome will be evaluated through appropriately developed evaluation metrics, baselines and benchmarks.

Dr Inventor will focus its evaluation on a specific scientific domain (i.e. computer graphics) exploring *Research Objects (RO)* from various sources. These will include: free research papers on the Web, research websites, Wikipedia, Internet forums, the home pages of many research institutes and groups, as well as individual researchers. In addition, research sites and social networks such as CiteSeer^{X}, ResearchGate and Google Scholar offer large numbers of freely accessible research papers; research source code is available from GitHub, SourceForge, etc.; and data can also be downloaded for research in computer graphics and image processing, e.g. from Flickr and from benchmarking archives. Secondly, it will use scholarly open-access journals. Finally, it will use online professional digital libraries for top-class research publications. Patents will also be considered within the scope of analysis by Dr Inventor.

The Dr Inventor Model

Research objects will be represented by *skeletons* to allow further processing. A *Research Object Skeleton (ROS)* represents the key concepts and relationships extracted from each RO. Retrieving and representing these ROS is the first task for the Dr Inventor model. The main challenges of the Dr Inventor project are now described in turn.

Information Extraction, Summarization, and RO Skeleton Generation Information Extraction (IE) and Text Summarization (TS) (Poibeau *et al.*, 2013) are two key technologies for transforming document content into concise, manageable semantic representations for use by our creativity model. Dr Inventor's IE aims to find not only general scientific concepts and relations such as: authors, institutions, research objectives, methods, citations, results, conclusions, developments, hypothesis postulation, hypothesis rejection, comparisons, etc. but also domain specific computer graphic concepts/relations such as algorithms, 3D modelling, rendering techniques, *etc*. Initial investigations have identified difficulties in extracting text from papers in PDF format. Issues include: '*ff*' and '*fi*' being represented as single characters, word-flow problems particularly in multi-column documents, representation of mathematical expressions, footnotes and page numbers appearing within the text. PDFX (Constantin *et al.*, 2013) will be used to assist in the text extraction process.

The inventory of entities to be extracted from different data-sources will be modelled in a domain ontology developed for Dr Inventor (see next Section). The most important methods to be used for IE are based on machine learning both supervised and semi-supervised. Indeed, in order for our methods to be applicable to different domains, techniques which are able to learn conceptualizations from raw text and propose new concepts are needed (Saggion, 2013), in this way IE will closely interact with ontology learning so as to expand scientific ontologies with specialized domain information. The GATE (http:///gate.ac.uk) system provides us with the basic infrastructure for developing and integrating basic and advanced IE components. Our current IE system is composed of modules for entity recognition (Ronzano et al. 2014) based on support vector machines (Li et al, 2009) and a rule-based approach for relation extraction based on dependency parsing output (Bohnet, 2010).

Summarization research in Dr Inventor is focusing on adaptation of summarization to scientific data by developing content relevance measures that take into account among other the scientific article rhetorical structure. We are producing an annotated data using an annotation schema based on work by (Liakata et al., 2010). Summaries will be used both as textual surrogates to allow scrutiny for scientist and as content briefers to identify main semantic information in the input. The work is being based on available generic summarization technology being adapted to the scientific domain (Saggion, 2008). Methods to produce these generic summaries are currently based on statistical techniques; however adaptation will be required to target the rich information present in scientific documents - eg Qazvinian *et al.*, (2010). To generate the ROS we need to extract sentence components such as the nouns and verbs, and the structure joining them. For example, from the sentence *"This paper in contrast, proposes a surface-oriented FFD"*, we extract the grammatical subject of the sentence: *paper*, the grammatical object: *FFD* and the relationship holding between them: *propose*. In addition to propositions, information regarding the structure of the article is also available (e.g., the fact that the proposition is extracted from a *purpose* rhetorical zone in the article).

Semantic Technologies & Ontology We will use existing semantic technologies to build up concepts and to identify the relationships between them. Domain ontologies will be built through the learning from a wide variety of research objects, including: documents, datasets, scripts, *etc*. Domain ontologies will also be used and connected to an upper-level ontology network, which will be developed in Dr Inventor as well, reusing existing ontologies covering scientific discourse, document structures, bibliographies and citations (e.g., DoCO, BIBO, EXPO, SPAR etc.) (Belhajjame *et al*, 2012). The extracted information related to authors, co-authors, affiliations, impact factors, h-indices, etc., will be used to facilitate the retrieval and ranking of RO’s but it will not be required in the analogy based model. We will also focus on knowledge extraction from user-defined tags associated to research objects and their aggregated objects, following on current work in ontology learning from folksonomies. In addition, extending existing work on social recommendation of research objects, we will be able to discover implicit relationships between different pieces of work that were originally not considered by the author in a basic literature exploration activity that can increment creativity in research. Such an ontology network will be designed to allow the representation of scientific discourse for scientific creativity.

With respect to ontology matching, we want to make use of existing techniques (Shvaiko and Euzenat, 2013) in the context of applying structured similarity evaluations between the aggregations of objects that are represented by research objects. In this context, knowledge extracted from documents and other artefacts should be seen as a skeleton set of information that summarizes key ideas, which allows researchers to explore the content of existing RO's in the process of their evaluation and of the generation of scientific innovation. This will contribute to the similarity measure for comparing research object skeletons for the creativity process. Finally, ontologies will also be used to provide personalized recommendations of scientific RO’s, using different sets of recommendation techniques.

Retrieval Model Retrieval will combine several techniques to identify homomorphic skeletons. A vector space model will enable quick, inexpensive comparison between skeletons, using numeric qualities representing the topology of each skeleton. This will also account for the inferences we expect to find in creative source domains.

Analogy/Blending Model Dr Inventor’s comparison model will identify and extend detailed similarities between ROS. It will typically search for a source to reinterpret a given target problem, but can also select its own targets. Dr Inventor’s final structure may be best seen as a conceptual blending (Fauconnier and Turner, 1998) model. It accepts as input two ROS, a generic space represents ontological and other commonalities while the output space represents the new creative concept (blend). (Space doesn’t permit proper treatment of the similarities and differences between analogy and conceptual blending).

Dr Inventor presents many challenges to similarity based discovery, such as; identifying a compelling source ROS, balancing structural and semantic factors in the mapping phase and performing quality assurance on the resulting inferences. Choosing the correct interpretation(s) of each domain to find an appropriate mapping will also be crucial.

The analogy-based model envisages the re-description of any given target using a pre-stored collection of sources with which to re-interpret that problem. This requires a rich memory of background knowledge to seek creative interpretations of the targeted problem through an extensive analogical comparison to a wide range of objects. In this context, Dr Inventor aims at exploring the potential of web resource to promote scientific creativity. From the previous example sentence in the IE section, we have the graph `[paper]→(propose)→[ffd]` where [] is a concept node and () denotes a relation connecting concept nodes.

Visual Analytics In Dr Inventor, visual analytics will serve to visualize the analogical reasoning and conceptual blending processes. Graphs visualization is a natural choice for the visualization of the ROS, which can also be supported by other means of visualizations. This could involve a large number of skeletons with a considerable level of uncertainty originated from similarity measures between the ROSs. Also, to allow effective handling of large scale visualization, we will investigate aggregation techniques such as binning, abstraction, hierarchical clustering to create effective aggregation of data at different levels of details.

User interaction with a creative system is an interesting research issue. The interaction techniques are categorized as select, explore, reconfigure, encode, abstract/elaborate, filter, and connect. An important task of user interaction is to help user navigation of the data. To this end, the interaction will follow the recommendation of *“overviews first and details-on-demands”* by working together with the data aggregation. Also, techniques that support zoom in within local areas, focus+context and coordinate views will help users to interactively explore comparisons without losing the perception towards the overall data structure.

Web-Based Creativity Service Dr Inventor will present a web-based system for exploring scientific creativity. It will offer a front-end web interface and a back-end mechanism addressing data transfer, access and federation, resource management, *etc*. The backend crawler will constantly gather research objects from the web extending the ROS repository. Information extraction and subsequent activities will be applied, as previously discussed.

At the front end, a web-based interface will be built to provide interface to allow interactive browse, search and visualization of the analogies of ROSs from the repository that contains analogically matched skeletons to inspire user creativity; to provide interface to assess an input RO; to provide interface for a creativity inspiration engine that allows scientific creativity promotion in highly interactive ways. The system is expected to be linked to a social network service (e.g. LinkedIn, Facebook or Twitter) to enhance the interaction and to explore common interest between the researchers. Finally, APIs will be developed to support further development.

Evaluation Among the remaining significant challenges will be evaluation of Dr Inventor, assessing its impact on the creativity of its user groups. This will rely heavily on access to a group of domain (computer graphics) experts for assessment and evaluation. Just as important to Dr Inventor is the development of a set of benchmarks and metrics for evaluating progress of this project.

Conclusion

Models of analogical reasoning are presenting new horizons for intelligently processing information, unearthing creative possibilities in new and surprising ways. Using analogy-based models upon academic resources is a broad and open-ended challenge, requiring advances in areas like document analysis, representation, ontology, analogy & blending, visualization *etc*. Dr Inventor aspires to Big-C Creativity (Gardner, 1993) hoping to support the transformational creativity (Boden, 1992) associated with significant scientific progress. Boden (1998) identifies two major "bottlenecks" for transformational creativity. Firstly, the domain expertise required for mapping the conceptual spaces to be transformed and secondly, valuing the results produced by a transformationally creative system. We believe that both challenges will be addressed by the combined efforts of the different activities in Dr Inventor, leading to a powerful tool that will invigorate the research communities opening up new and exciting possibilities.

A number of high-level issues arise related to Dr Inventor. Firstly, is documented information sufficiently complete to allow fruitful comparisons to be drawn between research papers, collections of papers or other sources? Can Dr Inventor adequately identify creative analogies from such sources? Will users be sufficiently receptive to accept creative inspiration from Dr Inventor? How can we maximize the impact from each component of Dr Inventor to produce comparisons with the greatest effect on its users? These and many other challenges await.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement 611383.

References

Belhajjame, K.; Corcho, O.; Garijo, D.; Zhao, J.; Missier, P.; Newman, D.R.; Palma, R.; Bechhofer, S.; Garcia-Cuesta, E.; Gómez-Pérez, J.M.; Klyne, G.; Page, K.; *et al* 2012 *Goble CA: Workflow-Centric Research Objects*. Proc. ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Greece.

Boden, M.A. 1992. *The Creative Mind*. London: Abacus.

Boden, M.A. 1998. *Creativity and artificial intelligence*, Artificial Intelligence, 103, 347-356.

Boden, M.A. 2009. *Computer Models of Creativity*, AI Magazine, 23-34.

Bohnet, B. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. Proc. COLING, pp 89-97.

Brown, T.L. 2003. *Making Truth: Metaphor in Science*, University of Illinois Press.

Constantin, A.; Pettifer, S.; Voronkov, A. 2013. *PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature*. ACM Symp. Document Engineering, 177-180.

Fauconnier, G. and Turner, M. *Conceptual Integration Networks*, Cognitive Science 22(2), 133–187, 1998.

Gardner, H. 1993. *Creating Minds*, Basic Books, NY.

Gick, L. M. and Holyoak, J. M., 1980. *Analogical problem solving,* Cognitive Psychology, 12, pp 306-355.

Juršič, M.; Cestnik, B.; Urbančič, T.; and Lavrač, N. 2012. *Finding Bridging Concepts with CrossBee*, International Conference on Computational Creativity, Ireland, 33-40.

Koestler, A. 1964. *The Act of Creation*, Penguin, NY.

Lopez, R.; Lindsey, J.S.; and Smith, S.M. 2011. *Characterizing the effect of domain distance in design-by-analogy*, ASME IDETC-Design Theory and Methodology Conf.

Li, Y., Bontcheva, K. and Cunningham, H. 2009: Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, Vol. 15, pp. 241-271, Cambridge University Press (2009)

Liakata M.; Teufel S.; Siddharthan A.; Batchelor C. 2010. Corpora for conceptualisation and zoning of scientific papers. 7th Conf. Intl. Language Resources and Evaluation.

O'Donoghue, D.P. and Keane, M.T. 2012. A Creative Analogy Machine: Results and Challenges, *4th International Conference on Computational Creativity,* Ireland, 17-24.

Poibeau, T.; Saggion, H.; Piskorski, J.; and Yangarber, R. 2013. *Multi-source, Multilingual Information Extraction and Summarization++,* Theory and Applications of Natural Language Processing, Springer.

Thibodeau, P.H. and Lera Boroditsky L. 2011. *Metaphors We Think With*, PlosOne, 6(2) 1-11.

Qazvinian, V.; Radev, D.R.; and Özgür, A. 2010. *Citation Summarization Through Keyphrase Extraction,* COLING 2010, 895-903.

Ronzano, F.; Casamayor, G.; and Saggion, H. 2014 Semantify CEUR-WS Proc.: towards the automatic generation of highly descriptive scholarly publishing Linked Datasets. Proc. ESWC-14 Challenge on Semantic Publishing.

Saggion, H. 2008. SUMMA: A Robust and Adaptable Summarization Tool. Traitement Automatique des Langues 49(2). pp103-125.

Saggion, H. 2013. Unsupervised Learning Summarization Templates from Concise Summaries. Proc. North American Chapter of Assoc. Computational Linguistics, 270-279.

Veale, T.; O'Donoghue, D.; Keane, M. 2000. *Computation and Blending*, Cognitive Linguistics, 11, 3/4, 253-281.

Shvaiko, P.; Euzenat, J. 2013 Ontology Matching: State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176.