Information and Data Management at PUC-Rio and UFMG

Antonio L. Furtado
Department of Informatics, PUC-Rio
Rio de Janeiro, Brazil
furtado@inf.puc-rio.br

Nivio Ziviani
Computer Science Dept., UFMG & Kunumi
Belo Horizonte, Brazil
nivio@dcc.ufmg.br

ABSTRACT

This article presents a summary of the main activities of the Database & Information Systems Research Group at Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) and the Information Management Research Group at Universidade Federal de Minas Gerais (UFMG). These two groups played a pioneering role in the development of the information and data management research area in Brazil. The survey covers about four decades of research work, aiming at theoretical and practical results, with increasing participation of other groups that they helped to initiate.

PVLDB Reference Format:
A. L. Furtado, N. Ziviani. Information and Data Management at PUC-Rio and UFMG. *PVLDB*, 11 (12): 2114-2129, 2018.
DOI: https://doi.org/10.14778/3229863.3240490

1. INTRODUCTION

This article is a brief survey of the activities of two Brazilian academic groups, the Database & Information Systems Research Group at the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), and the Information Management Research Group at the Universidade Federal de Minas Gerais (UFMG). Their work, already spanning about four decades, has exerted a major influence on the Brazilian community investing on database technology, including universities and research institutes, as well as public and private enterprises, and has enjoyed their continuing collaboration.

The first graduate program in Computer Science in Brazil was started in 1968 by the Departamento de Informática of PUC-Rio, with the initial support of the universities of Toronto and Waterloo. Graduates from this program participated in the creation of equally successful academic programs in several other Brazilian universities. Back in 1979, the first author served as program committee chair and editor of the conference proceedings, together with Howard L. Morgan, when the Fifth International Conference on Very Large Data Bases was held in Rio de Janeiro. He has also served as trustee, and is now trustee emeritus, of the VLDB Endowment.

The Information Management Research Group at UFMG was started in 1982 by the second author, Nivio Ziviani, after obtaining a PhD degree in Computer Science at the University of Waterloo in Canada. The work of the UFMG group has covered some of the key areas in modern Information Management from compression, crawling, indexing, machine learning, natural language processing to ranking. Further, its focus on addressing practical problems of relevance to society and on building prototypes to validate the proposed solutions has led to the spin-off of four successful start-up companies in Brazil, one of them acquired by Google Inc. to become its R&D center for Latin America. As a result, the group has established a solid reputation in its topics of interest and combines a large experience in technology-based enterprises with a wide network of collaborators in Brazil and abroad.

The activities of the Database & Information Systems group of PUC-Rio are reviewed in Section 2, and those of the Information Management Research Group of UFMG in Section 3. Section 4 contains closing remarks and acknowledgments.

2. DATABASE & INFORMATION SYSTEMS RESEARCH AT PUC-RIO

Research on databases at PUC-Rio dates back to the late seventies and covers a broad range of topics, from the early development of the relational model to recent interdisciplinary applications of semiotic and storytelling concepts to the design and specification of information systems. During these four decades, 92 MSc and 31 PhD students graduated from our academic program.

A number of visiting researchers had a major influence on the early development of our group, among which we may cite C.C. Gotlieb (the first author's PhD supervisor, at the University of Toronto), E.F. Codd, C.J. Date, M.M. Zloof, M.R. Stonebraker, E.J. Neuhold and R. Fagin. Larry Kerschberg, today at George Mason University, was one of our most active colleagues for several years. We are grateful to José Paulo Schiffini from IBM Brasil, who helped us organize the Fifth International Conference on Very Large Data Bases, held in Rio de Janeiro, on October 3-5, 1979.

We are pleased to recognize our long-standing collaboration with the database groups of several Federal Universities in Brazil, located respectively at the following states: Minas Gerais (UFMG), Rio de Janeiro (UFRJ, UNIRIO, UFF, UERJ), Amazonas (UFAM), Ceará (UFC), Rio Grande do

Proceedings of the VLDB Endowment, Vol. 11, No. 12
Copyright 2018 VLDB Endowment 2150-8097/18/8.
DOI: https://doi.org/10.14778/3229863.3240490

Sul (UFRGS, UFPel, UFSM), and São Paulo (UNICAMP). In special, the Database Group at the Universidade Federal do Rio Grande do Sul (UFRGS) was initiated by Clésio S. dos Santos and José M. V. de Castilho, who worked towards their PhD degree under the supervision of the first author. They would soon distinguish themselves, the former for creative leadership, and both of them for outstanding teaching and research performance. Other collaborators include research institutes such as the Instituto Nacional de Pesquisa Espacial (INPE), Instituto Brasileiro de Geografia e Estatística (IBGE) and the Laboratório Nacional de Computação Científica (LNCC). Among collaborating business corporations, we should cite Petrobras, IBM Brasil and Dell EMC. Finally, within our own department, the multimedia group has significantly participated in our projects.

Besides the research papers to be discussed in this section, the Database & Information Systems Group of PUC-Rio published a number of books, among which should be mentioned [32], [37] and [75]. As textbooks for undergraduate and introductory graduate courses, the group has started long ago with [76], describing the early hierarchic, network and relational data models, as well as file organizations and the then available database management systems.

In what follows we shall review some of the major contributions of our group, from the perspective of the first author. The contributions are organized according to the data model or to the underlying applications that they are based on: (1) the relational model; (2) the entity-relationship model; (3) several formalisms, including algebraic specification; (4) a plan generation and recognition paradigm; (5) methods to promote cooperative behavior; (6) the problem of publishing databases on the Web; (7) a semiotic approach to the conceptual specification of information systems; (8) adding story-bases to data-bases. The last four topics have received special attention.

2.1 Early Years

2.1.1 Contributions to the Relational Model

Contributions to the development of the relational model can be traced back to the 1977 SIGMOD Conference, where an algebra of quotient relations was proposed [74]. In a second early paper [62], the relational model was studied from three interdependent viewpoints. Relational databases were first modeled by directed hypergraphs, a concept derived in a straightforward way from Berge's hypergraph theory. Then, the abstract directed hypergraphs were interpreted using a linguistic model. Finally, the hypergraphs were represented with the help of relations and additional structures. Normalization was then discussed in the context of the three approaches. The design of relational databases based on functional dependencies (FDs) and inclusion dependencies (INDs) was addressed in [39]. Motivated by the above formal analysis, an investigation was undertaken on how to efficiently enforce inclusion dependencies and referential integrity [42, 43]. An analysis of the integrity constraints defined in the SQL ISO standard in the light of the entity-relationship model was also carried out [95]. Departing from the tradition of data dependencies, a database description framework [46] was introduced that accounts for both static constraints and transition constraints.

The view update and the view integration problems were addressed in several papers. The effects of a wide range of update operations on relational views were investigated [77] to identify which operations must be prohibited in order to assure harmonious interactions among database users, and which operations could be allowed, even though the structure of the view may substantially differ from the actual structure of the database. Later on, a survey on the view update problem was published [65], covering the two basic approaches proposed at that time to solve the problem. The first approach suggested treating views as abstract datatypes so that the definition of the view included all permissible view updates, together with their translations. The second approach led to general view update translators and was based either on an analysis of the conceptual schema dependencies or on the concept of view complement to disambiguate view update translations.

2.1.2 Contributions to the Entity-Relationship Model

Results about the entity-relationship model were already reported at the 1st ER Conference [126]. A datatype approach to database semantics was considered using the ER model as a framework. Later on, a method, also based on abstract datatypes, was proposed for representing a database application on a simple entity-relationship data model [78] and two constructs that capture and extend the generalization and subset abstractions were proposed [137], together with operations to maintain entity and relationship sets organized according to these constructs. As a result of the investigation on the ER model, an expert software tool, called CHRIS, was developed to help in the design and rapid prototyping of information systems containing a database component [136].

Continuing this line of research, we introduced a declarative way of specifying both the structure and the operations of an entity-relationship schema [66]. The paper proceeded to describe a plan generation algorithm and a method to introduce the time dimension, whereby the facts that hold at a certain instant can be inferred from the record of the operations executed. By combining these features, the paper showed how to extend temporal databases so as to cover past, present and future states (as determined by fixed commitments), as well as to draw plans coupled with time schedules. A second paper [44] defined a design algorithm that accepts as input an entity-relationship conceptual schema and generates an optimized relational representation for the schema (optimized in the sense that the number of dependencies of the relational schema is minimized).

The question of database redesign was retaken in [131]. A mapping strategy proposed earlier [95] was generalized in [132]. Finally, a survey, included in the Encyclopedia of Database Systems, summarized work on mapping entity-relationship schemas into relational schemas [22].

This long tradition of contributions to the ER Conferences was recognized through the first author's invited talk at the 28th ER Conference [70] and, years later, when he received the 2014 Peter P. Chen Award [68].

2.1.3 Formal Specification and Modularization

The earliest contribution to database design based on algebraic specifications was published in 1981 [125]. The paper proposed a formalism adequate for the specification of behavioral properties of data bases. Research then proceeded in three directions: complementary specifications, stepwise refinement and modular design.

A methodology was proposed for the systematic derivation of a series of complementary specifications of a database application [139]. The topic of complementary specifications was retaken in [45]. Logical, algebraic, programming language, grammatical and denotational formalisms were investigated with respect to their applicability to formal database specification. On applying each formalism for the purpose that originally motivated its proposal, the paper showed that they all have a fundamental and well-integrated role to play in different parts of the specification process.

Stepwise refinement and modularization were addressed in [129]. Modularization was discussed as another dimension in the specification process, orthogonal to stepwise refinement [41]. The modularization discipline incorporated both a strategy for enforcing integrity constraints and a tactic for organizing large sets of database structures, integrity constraints, and operations.

2.1.4 Plan Generation / Plan Recognition

We have been working with the conceptual modeling of information systems with a database component, considering their static, dynamic and behavioral aspects. The three aspects were integrated through the application of a plan-recognition / plan-generation paradigm [72]. The static aspect concerns what facts hold at some database state, conveniently described in terms of the entity-relationship model. The dynamic aspect corresponds to events that can produce state transitions. The behavioral aspect refers to the agents authorized to cause events by performing the operations.

As a further development, we have started to look at agent profiles involving three kinds of personality factors, from which a decision-making process could operate: *drives* for the emergence of goals from situations, *attitudes* for the choice of plans to achieve the preferred goal, and *emotions* to decide whether or not to commit to the execution of the chosen plan, depending on the expected emotional gain when passing from the current to the target state [17]. And, as an inducement to revise individual decisions, we included *competition* and *collaboration* interferences, as prescribed for multi-agent contexts [141].

In order to make our conceptual specifications executable, we created an environment where entity and relationship classes, operations, and goal-inference rules and agent profiles are all represented as Prolog clauses. Also written in Prolog, algorithms were provided, in increasingly extended versions [140, 49], for planning and for the simulated execution of the generated plans. Moreover, it was noted that simulation can become a useful resource to support learning or training [50].

The plan-recognition algorithm, which we adapted in part from [88], matches a few observed actions of the user against a library of previously recorded typical plans. As we explained in [73], the library of typical plans, in turn, can be constructed by inspecting the log and extracting and filtering sequences of executed operations whereby the transition indicated in some goal-inference rule has been achieved.

Treating databases as components of information systems encompassing facts, events and agents permits a shift from a purely descriptive to a narrative context [63]. Indeed, we showed how to generate template-based natural language text, by inspecting the plot-structured execution log and analyzing it against our three-level conceptual schemas [71]. It is therefore not surprising that all the discussion in this section applies in essentially the same way to literary genres [64]. We have adopted plan-based plot composition, coupled with several dramatization techniques and visual media, within an ongoing digital storytelling project [51].

The application of the plan-recognition / plan-generation paradigm to the narrative domain [64] was presented at the XIX Brazilian Symposium on Databases as an invited talk, on which occasion the author received a prize from the Brazilian Computer Society, acknowledging his contributions to database research.

2.2 A Closer Look on Selected Topics

2.2.1 Cooperative Behavior

An information system exhibits cooperative behavior to the extent that it interacts with users in ways that: contribute to the achievement of the users' goals and plans; keep the users' understanding of the system in harmony with the definition and contents of the system; conform to the established integrity and authorization constraints. In particular, when interacting with a database, a user may be tempted to infer further information from that explicitly obtained from previous queries. However, since his world model is often faulty or incomplete, a fact he infers may be false with respect to the database. Such facts are often called misconstruals.

At the 10th ER Conference, in order to achieve cooperative behavior, an algorithm was proposed that does not execute a request literally, but rather transforms the request appropriately, guided by a set of modification rules [83]. The algorithm may modify a request by invoking rules before the request is actually executed, after the request is successfully executed, or even after a failed execution. As part of project NICE, a prototype tool was implemented to run experiments with the request modification algorithm, which incorporated a plan and schedule generation algorithm, a temporal database package, a query-the-user facility and a session-monitoring feature [40].

These ideas were expanded later on into a model of question-answering [3, 130]. Taking as the starting point an input query, the system answers the query and then, in the course of a dialogue, tries to suggest new queries related to the input query. The dialogue control was based on the structure of the concepts stored in the knowledge base, on domain restrictions, and on specific constraining rules.

The problem of avoiding misconstruals was addressed in two papers that follow closely related strategies, but which differ on the formalisms used. The strategy was to create user models that capture the intuition that, whenever the user needs to derive a positive fact F in his inference, he must check whether the current log does not indicate that F must be rejected. One of the papers [85] described a model for users' inferences that directly checked if F must be rejected. The other paper adopted a more elegant solution based on Default Logic [84]. Both papers considered a cooperative interface which was responsible for all inferences from the deductive database necessary to answer the users' queries and to determine what additional information to include in the log to avoid misconstruals.

2.2.2 Publishing Data on the Web

Two basic approaches to access Deep Web data have been proposed. The first approach, called surfacing or Deep Web

Crawl, tries to automatically fill out HTML forms to query databases. Queries are executed offline and the results are translated to static Web pages, which are then indexed. The second approach, called federated search or virtual integration, suggests using domain-specific mediators to facilitate access to the databases. Hybrid strategies, which extend the previous approaches, have also been proposed.

A different approach to publish Deep Web databases was proposed in [111]. The basic strategy consists of creating a set of natural language sentences, with a simple structure, to describe Deep Web data, and publishing the sentences as static Web pages, which are then indexed as usual. The use of natural language sentences is convenient for three reasons. First, they lead to Web pages that are acceptable to Web crawlers that consider words randomly distributed in a page as an attempt to manipulate page rank. Second, they facilitate the task of more sophisticated engines that support semantic search based on natural language features. Lastly, the descriptions thus generated are minimally acceptable to human users.

A similar technique was introduced in [108] to automatically generate semantically enhanced descriptions of audio and video objects. The goal was to facilitate indexing and retrieval of the objects with the help of traditional search engines. Basically, the technique automatically generates static Web pages that describe the content of the digital audio and video objects, organized in such a way as to facilitate locating segments of the audio or video that correspond to the descriptions.

Given a spoken content, the first step is to transcribe it using an automatic speech recognition service. The set of time-aligned text excerpts thus obtained forms what we call the script of the spoken content, by analogy with the usual meaning of the word. Since the script is a textual representation of the spoken content, the spoken search problem is converted into a text search problem. The second step of the publishing technique is to translate the script to other languages, if desired, to reach a wider user population. The last step is decomposed into two substeps: (a) to transform a plain text script into a XHTML file; and (b) to annotate the content using RDFa (Resource Description Framework in attributes).

The technique was thoroughly analyzed by comparing the clickstreams to the digital content before and after the automatic generation of the descriptions. The outcomes suggest that the technique significantly improve the retrieval of items, not only in terms of visibility, but also brings down language barriers, by supporting multilingual access.

2.2.3 Semiotic Relations and Semiotic Completeness

When specifying any system, and when using it as well, some guidelines should be available. What properties are relevant to characterize an entity? What events should be observed? How do agents interact, either collaborating or competing? Is it possible to attain modularity, by setting the focus to different degrees of detail? Which integrity constraints should be enforced? Our invited paper at the ER 2014 Conference [68] shows how we approached this problem. The next paragraphs highlight some aspects covered at that occasion.

We based our proposal on studies [33, 47] asserting the completeness, as reasoning processes, of the so-called four master tropes: metonymy, metaphor, irony and synecdoche. Their universality has been repeatedly emphasized, with the indication that they may constitute "a system, indeed *the* system, by which the mind comes to grasp the world conceptually in language" [55]. Reasoning about these tropes, we identified four types of semiotic relations that can exist not only between facts, but also between events and between agents, which we denominated, respectively, syntagmatic, paradigmatic, antithetic and meronymic relations. Informally speaking, syntagmatic relations refer to connectivity, paradigmatic relations to similarity and analogy, antithetic relations to negation, and meronymic relations to hierarchy. The term *syntagmatic* has been used by linguists [128] to characterize a horizontal axis in the formation of sentences, whereas *paradigmatic* refers to drawing a vertical axis, to suggest alternative words to replace the word located at some position. *Antithetic* would recall the rules of grammar that impose bounds to the space of language (as integrity constraints do to the information space). *Meronymy*, that would correspond to a depth axis to zoom in and out across syntactic levels (sentence, word, syllable, etc.), was, curiously, treated in our very first participation in Entity-Relationship events [126], when we proposed to add semantic is-a and part-of hierarchies to the ER model. Not much later we learned about the seminal contribution of [142], where six types of part-of were distinguished.

Another paper of our group [16] can be regarded as a first attempt to deal with paradigmatic relations in the context of databases. The motivating problem was that databases, particularly when storing heterogeneous, sparse semi-structured data, tend to provide incomplete information and information which is difficult to categorize. The paper first considers how to classify entity instances as members of entity classes organized in a lattice-like generalization/specialization hierarchy. Then, it describes how the frame representation employed for instances and classes, as well as the closeness criterion involved in the classification method, favor the practical use of similarity and analogy, where similarity refers to instances within the same class, and analogy involves different classes. Finally, the paper argues that similarity and analogy facilitate querying semi-structured data. A more in-depth investigation of classification methods based on frames was the object of another work [109], which describes a tool to classify semi-structured data, represented by frames, without any previous knowledge about structured classes. The tool uses a variation of the K-Medoid algorithm and organizes a set of frames into classes, structured as a strict hierarchy.

The next step, still focusing on paradigmatic relations and the corresponding trope, metaphor, was to promote a reuse strategy, whereby new conceptual specifications might be partly derived from previous ones. We argued in [30] that analogy mappings facilitate conceptual modeling by allowing the designer to reinterpret fragments of familiar conceptual models in other contexts. This reuse strategy was further examined in [31]. This paper argued in favor of a database conceptual schema and Semantic Web ontology design discipline that explores analogy mappings to reuse the structure and integrity constraints of conceptual models, stored in a repository. A standard repository of source conceptual models previously assembled by expert designers was assumed, which less experienced designers would use to create new target conceptual models in other domains. The target models would borrow the structure and the integrity constraints from the source models by analogy. The concepts were expressed in the contexts of Description Logic, the RDF model and OWL to reinforce the basic principles and explore additional questions, such as the consistency of the target model.

Reusing a conceptual schema is of course a multi-phase process. After finding a suitable source schema, adaptations will often be needed in view of conflicts with the target schema being designed. The notion of blending [60] was exploited for this objective in [38]. To support the generation of database schemas of information systems, the paper proposed a five-step design process that explores the notions of generic and blended spaces and favors the reuse of predefined schemas. The use of generic and blended spaces is essential to achieve the passage from the source space into the target space in such a way that differences and conflicts can be detected and, whenever possible, conciliated. The convenience of working with multiple source schemas to cover distinct aspects of a target schema, as well as the possibility of creating schemas at the generic and blended spaces, was also considered. Notice that, as we would indicate more explicitly in later articles, the presence of conflicts already suggests the need to deal with antithetic relations.

In order to extend the reuse strategy to the design of dynamic schemas, we employed plots, also defined as a frame-like data structure [69]. A plot is a partially ordered set of events. Plot analysis is a relevant source of knowledge about the agents' behavior when accessing data stored in the database. It relies on logical logs, which register the actions of individual agents. The paper proposed techniques to analyze and reuse plots based on the concepts of similarity and analogy. The concept of similarity was applied to organize plots as a library and to explore the reuse of plots in the same domain. By contrast, the concept of analogy helps reuse plots across different domains. The techniques proposed in the paper find applications in areas such as digital storytelling and emergency response information systems, as well as some traditional business applications.

Frames and plots became increasingly important to our research projects. Our 28th ER Conference invited paper [70], recommends frames as an alternative to move from ER specifications to logical stage modeling, and treats frames as an abstract data type equipped with a Frame Manipulation Algebra. It is argued that frames, with a long tradition in AI applications, are able to accommodate the irregularities of semi-structured data, and that frame-sets generalize relational tables, allowing to drop the strict homogeneity requirement.

Likewise, a Plot Manipulation Algebra was proposed to handle plots in [87]. The seven basic operators, equally named in both the Frame Manipulation Algebra and in the Plot Manipulation Algebra, and working respectively on frames and plots, were introduced in view of the following four fundamental semiotic relations: syntagmatic relations (product, projection), paradigmatic relations (union, selection), antithetic relations (difference) and meronymic relations (combination, factoring).

The first three operators listed above encompass the equivalent to the five basic operators of Codd's relational algebra (product, projection, union, selection, difference). The additional two operators (combination, factoring) handle the hierarchical structures induced by the meronymic relations, being comparable to the nest / unnest operations over non-first-normal form (NF^2) tables, later admitted in the relational model (cf. our algebra of quotient relations [74]). Thus, it seems fair to claim that our algebras are *semiotically complete*, a notion that covers an ampler scope than that of Codd's relational completeness. Prototype logic-programming tools have been developed to experiment with the frame manipulation and the plot manipulation algebras.

2.2.4 From Data-bases to Story-bases

To enable practical experiments wherein information system domains could be viewed in terms of the stories emerging from their formal specification, we developed a prototype, called IDB, described in detail in a technical report [80], which allows to test our conceptually specified event-producing operations along three successive stages. The first stage deals with the logic programming (Prolog) definition of the *static*, *dynamic* and *behavioral* conceptual schemas. The operations, defined in a STRIPS-like declarative style by their pre-conditions and post-conditions (effects), can be tested by simulated execution in workspace memory, and can be handled by a backward-chaining plan-generation algorithm. At the second stage, at which Prolog communicates with Oracle via an ODBC interface, relational tables are created in correspondence to the entity-relationship static schema, and the declarative-style operations are compiled into a semi-procedural format, where predicates implementing `select` commands are generated to check the pre-conditions, and `insert`, `delete` and `update` commands to produce the effects directly on the database tables. Generated plans are treated as transactions, which are caused to backtrack if the pre-conditions of a constituent operation happen to fail. At the third stage, a second compiler converts the semi-procedural operations into independently executable Oracle storage procedures.

Keeping the orientation adopted in our early work [140], our approach consistently relies on a strict abstract datatype discipline to be maintained throughout the three stages, by requiring that database manipulation be restricted to such sets of pre-defined operations, whose pre-conditions and effects are articulated so as to enforce all integrity constraints. Hopefully, some mistakes in a specification could be detected by simulated execution or anticipated by plan-generation. Apart from error-detection, planning should help the designers to check whether each of the legitimate goals of the prospective users, specified by way of inference rules (associating a motivating situation S with a goal G) that are part of the behavioral schema, could be met, and, in contrast, whether there might exist unforeseen ways to reach inconsistent or undesirable situations.

But actual practice may still reveal previously ignored aspects. Users may develop typical plans, from whose analysis the designers would be able to devise ways to introduce corrections and improvements. For this purpose, as an important complement to plan-generation, the IDB prototype features a powerful plan-recognition facility. Whenever any of the event-producing operations is executed upon the database tables, it performs the side effect of inserting into a **Log** table a record indicating the respective transaction number, the current time stamp, and the name and parameters of the operation. The plan-recognition algorithm works on this special IDB environment that comprises the database tables, which represent entity and relationship instances, and the time-stamped **Log** records, which register the execution of operations whose pre-conditions and effects specification is available to the algorithm. Given as input a goal G that is true at the current state, it proceeds in a backward direction until reaching a motivating situation S and yields as output one or more *traces* extracted from the **Log**, consisting of those operations that contributed to achieve G, i.e., sub-sequences of the **Log** filtered to exclude operations not belonging to the same transaction or not contributing to G. Frequently occurring traces are suitable candidates for inclusion in a *library of typical plans* [73].

Users are also allowed to insert records, with the same composition as those of the **Log**, into an **Agenda** table, in order to register operations scheduled for execution at some future time. The presence of such tables, which register the past (or still non-committed future) occurrence of events, together with the plan-generation and plan-recognition algorithms, offers a temporal database environment, wherein, if periodical snapshots are taken from the database tables, it should become practically viable at a reasonable cost to recover, by a sort of interpolation method, any intermediate database state.

With the help of the IDB environment, we have been working on a *process-mining* project, which started with a preliminary investigation over the academic domain of our university [81]. The project extends our early work on what we called *plot-mining* [67]. A further extension is the combination of traces that aim at the same given goal into a *network structure*, wherein sub-sequences with similar effects are unified, and the convergence or divergence of sub-sequences is pictorially made explicit by join or fork nodes. By what may be termed *network-oriented reasoning*, one can not only realize what the traces have in common and in what they differ but can also devise new plans by traversing the network along a path composed of parts taken from different traces. It should be noted that an important related research project on process mining [1] also utilizes logs and network or workflow representations, but, contrary to our approach, is directed to legacy databases, not assuming therefore the availability of formally defined schemas.

Since traces are stories, we feel justified to regard as a story-base any repository associated with a log, and indeed we have, in our long-term digital entertainment Logtell project [51], dealt with literary stories, applying Computational Narratology notions. In a recent paper [97], we demonstrated how to construct networks by combining different variants of a folktale into a network and described a prototype to help users with no authorial background to interactively compose new variants. We have claimed [48] that this interdisciplinary approach is suitable to both literary genres and business information systems and propose to further exploit it in the continuation of our work.

3. INFORMATION MANAGEMENT RESEARCH AT UFMG

In this section we summarize Information Management research conducted at UFMG, from the perspective of the second author of this article. The work of the UFMG group has covered some of the key areas in modern Information Management from Natural Language Processing, Information Retrieval and Web Search Engines to Machine Learning.

In a period of more than 35 years, the group has engaged in a variety of projects developed at the Laboratory for Treating Information (LATIN), located at the UFMG Computer Science Department. The LATIN group has a long time tradition of cooperation with foreign institutions, such as the Instituto Superior Técnico in Lisbon (Portugal), University of Chile, University Pompeu Fabra (Spain), Virginia Tech (USA), Yahoo! Barcelona, and Yahoo! Santiago (Chile).

In Brazil, it maintains a strong collaboration with groups from several other universities. In particular, the group helped to create the Database and Information Retrieval Group (BDRI) at the Universidade Federal do Amazonas (UFAM). Four out of the five initial members of BDRI received their PhD degrees from UFMG, and two of them, Altigran S. da Silva and Edleno Silva de Moura, are today highly reputed and productive researchers in information management, having the latter worked towards his PhD degree under the supervision of the second author. Also important is the collaboration with the Universidade Federal do Rio Grande do Sul (UFRGS), Universidade de Campinas (UNICAMP) and Universidade de São Paulo (USP).

The second author formed more than sixty graduate students, many of them occupying key positions in academia and industry, and has a scientific production on Information Management spread across many of the major journals and conferences. Further, his focus on addressing practical problems of relevance to society and on building prototypes to validate the proposed solutions has led to the spin-off of four key start-up companies in Brazil. As a result, the group now combines a large experience in technology-based enterprises with a wide network of collaborators in Brazil and abroad. Next we review some of the main contributions of the group at UFMG from the perspective of the second author.

3.1 Group History

The research activities on the application of Information Retrieval (IR) techniques to natural language texts started in June 1984 when Nivio Ziviani, the group's principal investigator, visited Gaston Gonnet at the Computer Science Department of the University of Waterloo. At that time, the algorithms and data structures group of the University of Waterloo signed a 10 years contract with the Oxford English Dictionary (OED) to computerize the dictionary then comprising 21,000 pages and 600,000 word definitions. The OED project was coordinated by Gaston Gonnet and Frank Tompa. Many important results on algorithms to search natural language texts came out of that project.

Upon return to his visit to Waterloo, Nivio Ziviani started the UFMG Information Management Group still in 1984 to study efficient algorithms to retrieve information from natural language texts. The UFMG Information Management Group eventually gave birth to the Laboratory for Treating Information (LATIN) in 1994.

In 1989, the group started a fruitful cooperation with Ricardo Baeza-Yates and his group at University of Chile, with the first joint paper published in 1990 [10]. Since then Ricardo Baeza-Yates and Nivio Ziviani have published together close to 40 papers in conference proceedings and journals. In 1993, they co-founded SPIRE (International Symposium on String Processing and Information Retrieval) whose first edition was held at UFMG and the 25th one will be held on October 9th, 2018 in Lima, Peru. In 2005, they co-chaired the 2005 International ACM SIGIR Conference on Research and Development in Information Retrieval, which was held in Salvador, Brazil. They also participated in the RITOS and AMYRI projects funded by the Spanish agency CYTED in the 1990s.

In 1995, Berthier Ribeiro-Neto joined UFMG and LATIN, bringing his experience in core Information Retrieval and ranking to the group. In 1999, jointly with Ricardo Baeza-Yates, he published the book Modern Information Retrieval [12], with some chapters written in collaboration with researchers from our group (e.g., [107, 145]). This is one of the most cited publications in the history of IR, with 17,137 citations at the time of this writing, according to Google Scholar Citations.[1] A revised and greatly expanded second edition of this book was published [13] in 2011.

From 1986 to 2010, Nivio Ziviani published a series of five books on the design of algorithms and data structures[2] [143, 144, 146, 147, 148]. These five books cover algorithms on text searching, sorting, text compression, hashing, and perfect hashing, among other topics. Further, they also provide various useful algorithms and their associated programs, which can be used as a basis to build search engine prototypes.

In 1998, we launched the first Brazilian search engine called MINER, based on meta-searching, as described in [4]. In April 1998, this work led to the creation of the start-up Miner Technology Group, one of the first web technology companies in Brazil, which was sold to the group Folha de São Paulo/UOL in June 1999. This was one of the first experiences in spinning off a web start-up company from research conducted at a Brazilian university, showing that research results can be transferred to society by creating knowledge intensive start-ups.

In November 1999, in a project developed in collaboration with the UFMG Database Group headed by Alberto Laender, an important sequel happened with the launching of the TodoBR search engine [56]—a vertical search engine for the Brazilian web, owned by Akwan Information Technologies. Akwan, which became a successful start-up and a reference for web search in Brazil, was acquired by Google Inc. in July 2005—an acquisition that became worldwide news. With Akwan, Google bootstrapped its R&D Center for Latin America, which is located in Belo Horizonte.

In 2006, again in collaboration with the UFMG Database Group, we developed at LATIN the Perfil-CC Project with the objective of assessing the research and education quality of the top Brazilian Computer Science graduate programs [96, 101]. Within that project, we conducted a study of the scientific production of these programs in the 2004-2006 triennium. That study compared the scientific production of the Brazilian programs against that of reputable programs in North America and Europe and was based on data from DBLP - Digital Bibliography & Library Project.[3]

In 2009, invited by Altigran Soares da Silva and Edleno Silva de Moura from the Information Retrieval research group at Universidade Federal do Amazonas (UFAM), Ziviani co-founded Neemu Technologies—an e-commerce search and price comparison company focused on presenting product offers to web users. Neemu was acquired by Linx, the largest retail company in Brazil, in September 2015.

In 2010, Adriano Veloso joined LATIN, bringing his experience in Machine Learning and Natural Language Processing. The group became more focused on the development of new machine learning algorithms to improve recommender systems, reinforcement learning algorithms, deep neural architectures for graph data, semantic matching of images, speech emotion recognition end-to-end computationally efficient semantic parsing. Research projects have ramifications in diverse areas including health, education and arts.

Also in 2010, the Journal of Information and Data Management (JIDM) is published by UFMG. The journal, as the official publication of the Brazilian Computer Society Special Interest Group on Databases, provides a forum for academic research on databases and related topics on information retrieval, digital libraries, knowledge discovery, data mining, and geographic information systems, among others.

Following LATIN's tradition of promoting and exploring the transfer of research results and technology to society, in 2016, Ziviani co-founded Kunumi—a start-up company focused on making the bridge between businesses and machine learning. Kunumi produces Machine Learning knowledge intensive technology that is dependent on high quality research conducted in partnership with UFMG.

In 2018, the group created the Laboratory of Artificial Intelligence at UFMG, sponsored by Kunumi and coordinated by Adriano Veloso. The process of creation of the Laboratory is innovative at UFMG, specially in terms of intellectual property, which is fully transferred to Kunumi. This was possible due to a novel form of relationship between UFMG and start-up companies created by the group: UFMG's intellectual property of any know-how derived from research results is transferred to the start-up in the form of usufruct of 5% of shares. Thus UFMG is a shareholder of Kunumi, which givies it the right to participate in the social profit and receive the obtained price in case of ownership transfer.

3.2 Some of Most Relevant Research Results

This section presents some of the group's main research results covering the following topics: natural language processing, information retrieval, web search engines and machine learning. Note that most publications by Ziviani mentioned in this section has graduate students as first co-author.

3.2.1 Natural Language Processing

Text Compression. Our discussion here focuses on text compression methods that are suitable for use in an Information Retrieval environment. By suitable means to access text randomly and allow searching the compressed text directly and faster than the uncompressed one, something that was not possible to perform efficiently until the work presented in [58, 59]. We proposed a fast compression and decompression technique for natural language texts. We also have studied alternatives to inverted index compression and text compression considered simultaneously. In [105, 149], we combine index compression to block addressing and sequential search on compressed text. This illustrates a win-win case where there is no space-time trade-off.

Text Classification. Automatic text classification aims to to create models capable of associating documents with semantically meaningful categories. Automatic text classification algorithms usually employ a supervised learning strategy, where a classification model is first built using a set of pre-classified documents, i.e., a training set, which is then

[1] https://scholar.google.com
[2] http://www.dcc.ufmg.br/algoritmos/
[3] http://dblp.unitrier.de

used to classify unseen documents. In [34, 35], we focused on web directories, producing high levels of classification effectiveness (around 90%) which is good for a noisy environment such as the Web. Similar results could be obtained in other scenarios where hyperlinked information also existed such as those provided by citations among scientific documents and references among encyclopedia articles [53, 54].

3.2.2 Effective Information Retrieval

Set-Based Model. In here, we describe an algorithm that improved the Vector Space Model (VSM) for ranking web documents returned by a user query (VSM was proposed in 1975 by Gerard Salton from Cornell University), by taking into account patterns of word co-occurrence using a data mining technique called association rules, leading to significant gains in the quality of the rankings. We have developed a model that combines data mining and traditional information retrieval models [116, 117, 118]. It presents a new approach for ranking documents in the vector space model that (*i*) patterns of term co-occurrence are considered and processed efficiently; (*ii*) term weights are generated using a data mining technique called association rules, which leads to a new ranking mechanism called *set-based vector model.* The components are no longer index terms but index *termsets* (sets of index terms).

Hypergraph Model. Also related to web search, we proposed a representation of the web as a directed hypergraph, instead of a graph, where links can connect both pairs of pages and pairs of disjoint sets of pages [18]. Here, the web hypergraph is derived from the web graph by dividing the set of pages into non-overlapping sets and using the links between pages of distinct sets to create hyperarcs. Each hyperarc connects a set of pages to a single page, which provides more reliable information to link analysis methods.

Web Data Mining. Another interesting work developed by our group is WIM – Web Information Mining [110], a model for fast web mining prototyping. The underlying conceptual model of WIM provides its users with a level of abstraction appropriate for prototyping and experimentation throughout the web data mining task. The experimentation of WIM in real use cases has shown to significantly facilitate web mining prototyping. For example, in [11] we use WIM to study the evolution of textual content on the Web.

Assessing Academic Productivity in Computer Science. The identification of reputable entities is important in business, education, and many other fields. In [119], we propose to exploit the transference of reputation among entities to identify the most reputable ones. We instantiate our model in an academic search setting, by modeling research groups as reputation sources and publication venues as reputation targets. In [120], we discuss the problem of how to assess academic productivity based on publication outputs.

3.2.3 Web Search Engines

Crawling the Web. The quality of a web search engine is influenced by several factors, including coverage and the freshness of the content gathered by the web crawler. Web crawlers find, download, parse content and store pages in a repository. A large-scale web crawler has the following main components: fetcher, URL extractor, URL uniqueness verifier and scheduler. In [86], we present a new algorithm for verifying URL uniqueness in a large-scale web crawler. Focusing on freshness, one key challenge is to estimate the likelihood of a previously crawled webpage being modified. In [124], we present a genetic programming framework to generate score functions that produce accurate rankings of pages regarding their probabilities of having been modified.

Indexing. The usually large size of a search engine textual repository demands specialized indexing techniques for efficient retrieval. The two most important indexing methods are suffix arrays and inverted files. We have presented an efficient implementation of suffix arrays when the data is stored on secondary storage devices such as magnetic or optical disks [14]. We also have a series of works on parallel generation of suffix arrays [89, 106]. In a more theoretical work, we studied the problem of minimizing the expected cost of binary searching for data where the access cost is not fixed and depends on the last accessed element, such as data stored in magnetic or optical disk [15, 104]. Another theoretical work on binary search trees with costs depending on the access paths is reported in [134]. In [123], we present distributed algorithms to build global inverted files for very large text collections.

Query Processing. One of the key difficulties on query searching is that users usually submit very short and ambiguous queries. In [61], we propose a concept-based query expansion technique, which allows disambiguating queries submitted to search engines. One of our main efforts is to develop new distributed query processing strategies for search engines. In [8], we present a real distributed architecture implementation that offers concurrent query service. In [9], we study key issues related to distributed web query processing. In [6], we modeled workloads for a web search engine from a system performance point of view. The performance of parallel query processing in a cluster of index servers for modern web search systems was discussed in [7]. A model for predicting the response time of a vertical search engine was presented in [5]. A substantial fraction of web search queries contains references to entities, such as persons, organizations, and locations. In [29], we present a supervised learning approach that exploits named entities for query expansion using Wikipedia as a repository of high quality feedback documents. Content-targeted advertising, the task of automatically associating ads to a web page, constitutes a key web monetization strategy nowadays. In [90], we propose ranking functions for associating ads with web pages, which aims at learning functions that select the most appropriate ads, given the contents of a web page.

Link Analysis. Information derived from the cross-references among the documents in a hyperlinked environment is considered important since it can be used to effectively improve document retrieval. In [36, 133], we investigate how the use of local link information compares to the use of global link information.

Caching. In [127], we present an effective caching scheme that reduces the computing and I/O requirements of a web search engine without altering its ranking characteristics. The novelty is a two-level caching scheme that simultaneously combines cached query results and cached inverted lists on a real case search engine.

Detecting Replicated Web Sites. Duplicate content on the Web occurs within the same website or across multiple websites. The latter is mainly associated with the existence of website replicas sites that are perceptibly similar. In [57], we model the detection of website replicas as a pairwise classification problem with distant supervision. Finding obvious replica and non-replica cases is trivial, but learning effective classifiers requires a representative set of non-obvious labeled examples, which are hard to obtain. We employ efficient expectation-maximization algorithms in order to find non-obvious examples from obvious ones, enlarging the training-set and improving the classifiers iteratively.

Near-Optimal Space for Minimal Perfect Hashing Functions. The design of new hashing methods for *static sets of keys* is strongly related to the generation of indexes for Information Retrieval systems, since a significant portion of the time is spent in hash operations. We proposed a construction for minimal perfect hash functions that combines theoretical analysis, practical performance, expected linear constructing time and near optimal space consumption for the data structure [23, 24, 25, 26, 27, 28]. The space consumption for the data structure considering n keys and $m = n$ (m the size of the hash table) ranges from $2.62n$ to $3.3n$ bits, and for $m = \lceil 1.23n \rceil$ it ranges from $1.95n$ to $2.7n$ bits. This is within a small constant factor from the theoretical lower bounds of $1.44n$ bits for $m = n$ and $0.89n$ bits for $m = \lceil 1.23n \rceil$. An open source implementation of the algorithms is available in the C Minimal Perfect Hashing Library (CMPH)[4] under the GNU Lesser General Public License (LGPL). It was incorporated by the Debian and Ubuntu Linux distributions, which indicates how useful the results are in practice. The number of downloads of an efficient implementation on May 25th, 2018 is 35,196.

3.2.4 Machine Learning

Machine learning is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Artificial neural networks, a branch of machine learning, are able to learn complicated patterns in large amounts of data. Neural network techniques are currently the state-of-the-art for identifying objects in images and words in sounds.

Learning to Rank. The group has contributed with new algorithms that rank documents and images using machine learning. In [2], we discuss the problem of learning to rank in scenarios where labeled data is scarce. Specifically, we developed deep autoencoders that learn feature transformations using inexpensive unlabeled data and available labeled data so that it becomes easier for existing learning to rank algorithms to find better ranking models from labeled datasets that are limited in size and quality. In [138], we propose a compositional approach for fashion retrieval by arguing that the semantics of an outfit can be recognised by their constituents (i.e., clothing items and accessories). Specifically, we present a semantic compositional network in which clothing items are detected from the image and the probability of each item is used to compose a vector representation for the outfit. Finally, outfits are ranked according to the semantic distance to the query. In [102], we apply multiple diversification approaches for dynamic information retrieval. In [103], we develope systems that learn about the user's need from his or her interactive exploration.

Recommender Systems. The group became more focused on the development of new machine learning algorithms to improve recommender systems. Some of the results include solutions based on techniques as diverse as collaborative filtering [19, 20, 21], reinforcement learning algorithms [92, 93, 94], multi-objective optimization [121, 122], genetic programming [82], taxonomies [99], and association rules [100]. Specific challenges include new item recommendation and cold-start recommendation [52], balancing diversity and novelty [122], and dealing with sparse data [19]. Daily-Deals sites enable local businesses, such as restaurants and stores, to promote their products and services to increase their sales by offering customers significantly reduced prices. In [91], we propose a new algorithm for daily deals recommendation based on an explore-then-exploit strategy.

Speech Emotion Recognition. Emotion recognition from speech is one of the key steps towards emotional intelligence in advanced human-machine interaction. Identifying emotions in human speech requires learning features that are robust and discriminative across diverse domains that differ in terms of language, spontaneity of speech, recording conditions, and types of emotions. In [98], we propose an architecture that jointly exploits a convolutional network for extracting domain-shared features and a long short-term memory network for classifying emotions using domain-specific features. We use transferable features to enable model adaptation from multiple source domains, given the sparseness of speech emotion data and the fact that target domains are short of labeled data.

Deep Neural Architectures for Graph Data. More recently, we focused on the design and development of novel deep neural architectures for learning node representations on large graphs [112, 114]. Anomaly detection (a.k.a. outlier detection) aims to discover rare instances that do not conform to the patterns of majority. In [113], we present a generalized active learning approach for unsupervised anomaly detection problem.

Semantic Parsing and Question Answering. We developed efficient and effective deep learning architectures to text understanding. Application scenarios include end-to-end computationally-efficient semantic parsing [115] for mapping natural language utterances into machine interpretable meaning representations and question answering [150]. Query understanding is a challenging task primarily due to the inherent ambiguity of natural language. A common strategy for improving the understanding of natural language queries is to annotate them with semantic information mined from a knowledge base. In [79], we propose a framework for learning semantic query annotations suitable to the target intent of each individual query.

Medical Data Analysis. Recently, we applied artificial neural network algorithms to medical data analysis that lead to improved diagnoses and treatment [135]. Current endeavors are concentrated on answering big questions with direct impact in health, developing (i) new models based on convolutional and recurrent networks that are trained to perform dynamic predictions about the health condition of patients in intensive care units where predictions are interpretable, thus enabling clinical reasoning, (ii) new machine learning technologies for automatically evaluating blood panels for screening the Alzheimer's disease in a non-invasive way with the potential to affect large scale sub-populations, (iii) natural language processing models that are based on jointly processing speech and facial information for recognizing chronic pain in human adults and fetuses.

Painting Authentication. Finally, we are currently developing machine learning models to facilitate authentication and attribution of drawings and paintings of Cândido Porti-

[4] http://cmph.sf.net

nari.[5] Our models are based on quantifying his style characteristics and generate images that resemble the style of this world-famous Brazilian painter.

4. CLOSING REMARKS AND ACKNOWLEDGEMENTS

The activities of the two groups are in a broad sense complementary. They were created independently and addressed different topics, though at all times jointly participated in many initiatives towards the development of the database community in Brazil. One especially relevant research wherein the two groups have collaborated was reported in section 2.1.2 [22, 44, 95, 131, 132].

The creation and early activities of the other university groups have been reported in detail in online accessible issues of the Journal of Information and Data Management (JIDM), whose importance as a forum of database research interchange has been stressed:

- UNICAMP (https://seer.ufmg.br/index.php/jidm/article/view/229)
- UFPE (https://seer.ufmg.br/index.php/jidm/article/view/121)
- COPPE/UFRJ (https://seer.ufmg.br/index.php/jidm/article/view/119)
- UFAM (https://seer.ufmg.br/index.php/jidm/article/view/120)
- UFC (https://seer.ufmg.br/index.php/jidm/article/view/118)
- ICMC/USP (https://seer.ufmg.br/index.php/jidm/article/view/125)
- IME/USP (https://seer.ufmg.br/index.php/jidm/article/view/128)
- UFRGS (http://www.academia.edu/2957853/The_Database_Research_Group_at_UFRGS)

All these groups have fared successfully both in research and teaching. In the 2017 evaluation of the graduate programs in Computer Science carried out by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) of the Brazilian Ministry of Education, the following universities received the highest grade[6]: UFPE, COPPE/UFRJ, PUC-Rio, UFMG, ICMC/USP, UNICAMP, and UFRGS.

To report the work at PUC-Rio, the first author counted with the help of his colleague Marco A. Casanova, with whom he has been working in equal partnership along all these years. Regarding the preparation of the UFMG survey, the second author is grateful to Alberto H. F. Laender and is pleased to recognize the continuing collaboration of his Database Research Group.

The research projects of both groups are partially supported by grants from two federal government agencies: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). At the state level, the projects of the PUC-Rio group receive additional financial support from Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), and those of the UFMG group from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

5. REFERENCES

[1] W. M. Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, Berlin, Heidelberg, 2011.

[2] A. Albuquerque, T. Amador, R. Ferreira, A. Veloso, and N. Ziviani. Learning to rank with deep autoencoder features. In *Proceedings of the IEEE International Conference on Neural Networks*, 2018.

[3] J. P. Alcazar, A. S. Hemerly, M. A. Casanova, and A. L. Furtado. Cooperative interfaces for spatio-temporal databases. In *Proceedings of the International Conference on Systems Research, Informatics and Cybernetics, Focus Symposium on Database and Expert Systems, Baden-Baden, Germany*, 1994.

[4] V. Almeida, W. Meira Jr, V. Ribeiro, and N. Ziviani. Efficiency analysis of brokers in the electronic marketplace. In *Proceedings of International Conference on the World Wide Web*, pages 1–12, 1999.

[5] C. S. Badue, J. M. Almeida, V. Almeida, R. Baeza-Yates, B. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Capacity planning for vertical search engines. *CoRR*, abs/1006.5059, 2010.

[6] C. S. Badue, R. Baeza-Yates, B. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Modeling performance-driven workload characterization of web search systems. In *Proceedings of 15th ACM International Conference on Information and Knowledge Management*, pages 842–843, 2006.

[7] C. S. Badue, R. Baeza-Yates, B. Ribeiro-Neto, A. Ziviani, and N. Ziviani. Analyzing imbalance among homogeneous index servers in a web search system. *Information Processing and Management*, 43(3):592–608, 2007.

[8] C. S. Badue, R. Baeza-Yates, B. Ribeiro-Neto, and N. Ziviani. Distributed query processing using partitioned inverted files. In *Proceedings of 8th International Symposium on String Processing and Information Retrieval*, pages 10–20, 2001.

[9] C. S. Badue, R. Barbosa, P. B. Golgher, B. Ribeiro-Neto, and N. Ziviani. Basic issues on the processing of web queries. In *Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 577–578, 2005.

[10] R. Baeza-Yates, G. H. Gonnet, and N. Ziviani. Expected behaviour analysis of avl trees. In *Proceedings of 2nd Scandinavian Workshop on Algorithm Theory*, pages 143–159, 1990.

[11] R. Baeza-Yates, A. Pereira-Jr, and N. Ziviani. Genealogical trees on the web: a search engine user perspective. In *Proceedings of 18th International World Wide Web Conference*, pages 367–376, 2008.

[5] http://www.portinari.org.br/

[6] http://avaliacaoquadrienal.capes.gov.br/resultado-da-avaliacao-quadrienal-2017-2

[12] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval (First edition)*. Addison-Wesley, 1999.

[13] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval (Second edition)*. Pearson, 2011.

[14] R. A. Baeza-Yates, E. F. Barbosa, and N. Ziviani. Hierarchies of indices for text searching. *Information Systems*, 21(6):497–514, 1996.

[15] E. F. Barbosa, G. Navarro, R. A. Baeza-Yates, C. H. Perleberg, and N. Ziviani. Optimized binary search and text retrieval. In *Proceedings of 3rd European Symposium on Algorithms*, pages 311–326, 1995.

[16] S. D. J. Barbosa, K. K. Breitman, A. L. Furtado, and M. A. Casanova. Similarity and analogy over application domains. In *Anais do XXII Simpósio Brasileiro de Banco de Dados, 15-19 de Outubro, João Pessoa, Brasil, Anais*, pages 238–252, 2007.

[17] S. D. J. Barbosa, A. L. Furtado, and M. A. Casanova. A decision-making process for digital storytelling. In *Proceedings of the 2010 Brazilian Symposium on Games and Digital Entertainment*, pages 1–11, Nov 2010.

[18] K. Berlt, E. S. de Moura, A. L. da Costa Carvalho, M. Cristo, N. Ziviani, and T. Couto. Modeling the web as a hypergraph to compute page reputation. *Information Systems*, 35(5):530–543, 2010.

[19] A. Bessa, A. H. F. Laender, A. Veloso, and N. Ziviani. Alleviating the sparsity problem in recommender systems by exploring underlying user communities. In *Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management*, pages 35–47, 2012.

[20] A. Bessa, R. L. T. Santos, A. Veloso, and N. Ziviani. Exploiting item co-utility to improve collaborative filtering recommendations. *Journal of the American Society for Information Science and Technology*, 68(10):2380–2393, 2017.

[21] A. Bessa, A. Veloso, and N. Ziviani. Using mutual influence to improve recommendations. In *Proceedings of the 20th International Symposium on String Processing and Information Retrieval*, pages 17–28, 2013.

[22] A. Borgida, M. A. Casanova, and A. H. F. Laender. Logical database design: from conceptual to logical schema. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 1645–1649. Springer US, 2009.

[23] F. Botelho, D. Galinkin, W. Meira-Jr., and N. Ziviani. Distributed perfect hashing for very large key sets. In *Proceedings of 3rd International Conference on Scalable Information Systems*, 2008.

[24] F. Botelho, Y. Kohayakawa, and N. Ziviani. A practical minimal perfect hashing method. In *4th International Workshop on Efficient and Experimental Algorithms*, pages 488–500, 2005.

[25] F. Botelho, R. Pagh, and N. Ziviani. Simple and space-efficient minimal perfect hash functions. In *Proceedings of 10th Workshop on Algorithms and Data Structures*, pages 139–150, 2007.

[26] F. Botelho and N. Ziviani. External perfect hashing for very large key sets. In *Proceedings of 16th ACM International Conference on Information and Knowledge Management*, pages 653–662, 2007.

[27] F. Botelho and N. Ziviani. Practical perfect hashing algorithm in nearly optimal space. *Information Systems*, 38(1):108–131, 2013.

[28] F. C. Botelho, N. C. Wormald, and N. Ziviani. Cores of random r-partite hypergraphs. *Information Processing Letters*, 112(8-9):314–319, 2012.

[29] W. C. Brandão, R. L. T. Santos, N. Ziviani, E. S. de Moura, and A. S. da Silva. Learning to expand queries using entities. *Journal of the American Society for Information Science and Technology*, 65(9):1870–1883, 2014.

[30] K. K. Breitman, S. D. J. Barbosa, M. A. Casanova, and A. L. Furtado. Conceptual modeling by analogy and metaphor. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM'07, Lisbon, Portugal, November 6-10, 2007*, pages 865–868, 2007.

[31] K. K. Breitman, S. D. J. Barbosa, M. A. Casanova, A. L. Furtado, and M. G. Hinchey. Using analogy to promote conceptual modeling reuse. In *Proceedings of the Workshop On Leveraging Applications of Formal Methods, Verification and Validation, ISoLA 2007, Poitiers-Futuroscope, France, December 12-14, 2007*, pages 111–122, 2007.

[32] K. K. Breitman, M. A. Casanova, and W. Truszkowski. *Semantic Web: Concepts, Technologies and Applications*. NASA Monographs in Systems and Software Engineering. Springer, 2007.

[33] K. Burke. *A Grammar of Motives*. University of California Press, 1969.

[34] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of web documents. *Journal of the American Society for Information Science and Technology*, 57(2):208–221, 2006.

[35] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Gonçalves. Combining link-based and content-based methods for web document classification. In *Proceedings of 12th ACM International Conference on Information and Knowledge Management*, pages 394–401, 2003.

[36] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. S. de Moura, and I. Silva. Local versus global link information in the web. *ACM Transactions on Information Systems*, 21(1):42–63, 2003.

[37] M. A. Casanova. *Concurrency Control Problem for Database Systems*. Springer-Verlag, Berlin, Heidelberg, 1981.

[38] M. A. Casanova, S. D. J. Barbosa, K. K. Breitman, and A. L. Furtado. Generalization and blending in the generation of entity-relationship schemas by analogy. In *Proceedings of the Tenth International Conference on Enterprise Information Systems, ICEIS'08, Volume ISAS-2, Barcelona, Spain, June 12-16, 2008*, pages 43–48, 2008.

[39] M. A. Casanova, R. Fagin, and C. H. Papadimitriou. Inclusion dependencies and their interaction with functional dependencies. *J. Comput. Syst. Sci.*, 28(1):29–59, 1984.

[40] M. A. Casanova and A. L. Furtado. An information system environment based on plan generation. In *Proceedings of the Working Conference on Cooperating Knowledge based Systems*, 1990.

[41] M. A. Casanova, A. L. Furtado, and L. Tucherman. A software tool for modular database design. *ACM Trans. Database Syst.*, 16(2):209–234, 1991.

[42] M. A. Casanova, L. Tucherman, and A. L. Furtado. Enforcing inclusion dependencies and referencial integrity. In *Proceedings of the Fourteenth International Conference on Very Large Data Bases, August 29 - September 1, 1988, Los Angeles, California, USA*, pages 38–49, 1988.

[43] M. A. Casanova, L. Tucherman, A. L. Furtado, and A. P. Braga. Optimization of relational schemas containing inclusion dependencies. In *Proceedings of the Fifteenth International Conference on Very Large Data Bases, August 22-25, 1989, Amsterdam, The Netherlands*, pages 317–325, 1989.

[44] M. A. Casanova, L. Tucherman, and A. H. F. Laender. On the design and maintenance of optimized relational representations of entity-relationship schemas. *Data Knowl. Eng.*, 11(1):1–20, 1993.

[45] M. A. Casanova, P. A. S. Veloso, and A. L. Furtado. Formal data base specification - an eclectic perspective. In *Proceedings of the Third ACM SIGACT-SIGMOD Symposium on Principles of Database Systems, April 2-4, 1984, Waterloo, Ontario, Canada*, pages 110–118, 1984.

[46] J. M. V. Castilho, M. A. Casanova, and A. L. Furtado. A temporal framework for database specifications. In *Proceedings of the Eight International Conference on Very Large Data Bases, September 8-10, 1982, Mexico City, Mexico*, pages 280–291, 1982.

[47] D. Chandler. *Semiotics: the Basics*. Routledge, New York, NY, 2002.

[48] A. E. M. Ciarlini, M. A. Casanova, A. L. Furtado, and P. A. S. Veloso. Modelling interactive storytelling genres as application domains. *J. Intell. Inf. Syst.*, 35(3):347–381, 2010.

[49] A. E. M. Ciarlini and A. L. Furtado. Understanding and simulating narratives in the context of information systems. In *Proceedings of the 21st International Conference on Conceptual Modeling, ER'02, Tampere, Finland, October 7-11, 2002*, pages 291–306, 2002.

[50] A. E. M. Ciarlini and A. L. Furtado. Towards a plan-based learning environment. In *Proceedings of the I PGL Database Research Conference, PGLDB'2003, Rio de Janeiro, Brazil, April 10-11, 2003*, 2003.

[51] A. E. M. Ciarlini, C. T. Pozzer, A. L. Furtado, and B. Feijó. A logic-based tool for interactive generation and dramatization of stories. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology, ACE'05, Valencia, Spain, June 15-15, 2005*, pages 133–140, 2005.

[52] T. F. Costa, A. Lacerda, R. L. T. Santos, and N. Ziviani. Information-theoretic term selection for new item recommendation. In *Proceedings of the 21st International Symposium on String Processing and Information Retrieval*, pages 236–243, 2014.

[53] T. Couto, M. Cristo, M. A. Gonçalves, P. Calado, N. Ziviani, E. Moura, and B. Ribeiro-Neto. A comparative study of citations and links in document classification. In *Proceedings of 6th ACM/IEEE Joint Conference on Digital Libraries*, pages 75–84, 2006.

[54] T. Couto, N. Ziviani, P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, and W. C. Brandão. Classifying documents with link-based bibliometric measures. *Information Retrieval*, 13(4):315–345, 2010.

[55] J. Culler. *The Pursuit of Signs: Semiotics, Literature, Deconstruction*. Cornell University Press, Ithaca, NY, 2002.

[56] A. S. da Silva, E. A. Veloso, P. B. Golgher, B. Ribeiro-Neto, A. H. F. Laender, and N. Ziviani. Cobweb - a crawler for the brazilian web. In *Proceedings of the International Symposium on String Processing and Information Retrieval*, pages 184–191, 1999.

[57] C. R. de Carvalho, E. S. de Moura, A. Veloso, and N. Ziviani. Website replica detection with distant supervision. *Information Retrieval Journal*, 21(4):1–20, 2017.

[58] E. S. de Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Fast searching on compressed text allowing errors. In *Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–306, 1998.

[59] E. S. de Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Fast and flexible word searching on compressed text. *ACM Transactions on Information Systems*, 18(2):113–139, 2000.

[60] G. Fauconnier and M. Turner. Conceptual projection and middle spaces. Technical report, Univ. California, San Diego, 1994.

[61] B. M. Fonseca, P. B. Golgher, B. Pôssas, B. Ribeiro-Neto, and N. Ziviani. Concept-based interactive query expansion. In *Proceedings of 14th ACM International Conference on Information and Knowledge Management*, pages 696–703, 2005.

[62] A. L. Furtado. Formal aspects of the relational model. *Inf. Syst.*, 3(2):131–140, 1978.

[63] A. L. Furtado. Narratives and temporal databases: An interdisciplinary perspective. In G. Goos, J. Hartmanis, J., van Leeuwena, P. P. Chen, J. Akoka, H. Kangassalu, and B. Thalheim, editors, *Conceptual Modeling: Current Issues and Future Directions*, pages 73–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[64] A. L. Furtado. Narratives over real-life and fictional domains. In *Anais do XIX Simpósio Brasileiro de Bancos de Dados, 18-20 de Outubro, 2004, Brasília, Distrito Federal, Brasil*, pages 4–12, 2004.

[65] A. L. Furtado and M. A. Casanova. Updating relational views. In *Query Processing in Database Systems*, pages 127–142. Springer, 1985.

[66] A. L. Furtado and M. A. Casanova. Plan and schedule generation over temporal databases. In *Proceedings of the 9th International Conference on*

Entity-Relationship Approach, ER'90, 8-10 October, 1990, Lausanne, Switzerland, pages 235–248, 1990.

[67] A. L. Furtado, M. A. Casanova, S. D. Barbosa, and K. Breitman. Plot mining as an aid to characterization and planning. Technical report, Dept. of Informatics, PUC-Rio, 2007.

[68] A. L. Furtado, M. A. Casanova, and S. D. J. Barbosa. A semiotic approach to conceptual modelling. In *Proceedings of the 33rd International Conference on Conceptual Modeling, ER'14, Atlanta, GA, USA, October 27-29, 2014*, pages 1–12, 2014.

[69] A. L. Furtado, M. A. Casanova, S. D. J. Barbosa, and K. K. Breitman. Analysis and reuse of plots using similarity and analogy. In *Proceedings of the 27th International Conference on Conceptual Modeling, ER'08, Barcelona, Spain, October 20-24, 2008*, pages 355–368, 2008.

[70] A. L. Furtado, M. A. Casanova, K. K. Breitman, and S. D. J. Barbosa. A frame manipulation algebra for ER logical stage modelling. In *Proceedings of the 28th International Conference on Conceptual Modeling, ER'09, Gramado, Brazil, November 9-12, 2009*, pages 9–24, 2009.

[71] A. L. Furtado and A. E. M. Ciarlini. Generating narratives from plots using schema information. In *Proceedings of the 5th International Conference on Applications of Natural Language to Information Systems, NLDB 2000, Versailles, France, June 28-30, 2000, Revised Papers*, pages 17–29, 2000.

[72] A. L. Furtado and A. E. M. Ciarlini. The plan recognition / plan generation paradigm. In *Information engineering: state of the art and research themes*, 2000.

[73] A. L. Furtado and A. E. M. Ciarlini. Constructing libraries of typical plans. In *Proceedings of the 13th International Conference on Advanced Information Systems Engineering, CAiSE'01, Interlaken, Switzerland, June 4-8, 2001*, pages 124–139, 2001.

[74] A. L. Furtado and L. Kerschberg. An algebra of quotient relations. In *Proceedings of the 1977 ACM SIGMOD International Conference on Management of Data, Toronto, Canada, 1977*, pages 1–8, 1977.

[75] A. L. Furtado and E. J. Neuhold. *Formal Techniques for Data Base Design*. Springer-Verlag, 1986.

[76] A. L. Furtado and C. S. Santos. *Organização de Bancos de Dados*. Ed. Campus, S. Paulo, 1979.

[77] A. L. Furtado, K. C. Sevcik, and C. S. dos Santos. Permitting updates through views of data bases. *Inf. Syst.*, 4(4):269–283, 1979.

[78] A. L. Furtado, P. A. S. Veloso, and J. M. V. de Castilho. Verification and testing of S-ER representations. In *Proceedings of the Second International Conference on the Entity-Relationship Approach, ER'81, Washington, DC, USA, October 12-14, 1981*, pages 123–147, 1981.

[79] R. Glater, R. L. T. Santos, and N. Ziviani. Intent-aware semantic query annotation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 485–494, 2017.

[80] V. Gottin, M. A. Casanova, E. S. de Lima, and A. L. Furtado. A story-based approach to information systems. Technical report, Dept. of Informatics, PUC-Rio, 2015.

[81] V. Gottin, H. Jimenez, A. C. Finamore, M. A. Casanova, A. L. Furtado, and B. P. Nunes. An analysis of degree curricula through mining student records. In *Proceedings of the 17th IEEE International Conference on Advanced Learning Technologies, ICALT 2017, Timisoara, Romania, July 3-7, 2017*, pages 276–280, 2017.

[82] A. P. Guimarães, T. F. Costa, A. Lacerda, G. L. Pappa, and N. Ziviani. GUARD: A genetic unified approach for recommendation. *Journal of Information and Data Management*, 4(3):295–310, 2013.

[83] A. S. Hemerly, M. A. Casanova, and A. L. Furtado. Cooperative behavior through request modification. In *Proceedings of the 10th International Conference on Entity-Relationship Approach, ER'91, 23-25 October, 1991, San Mateo, California, USA*, pages 607–621, 1991.

[84] A. S. Hemerly, M. A. Casanova, and A. L. Furtado. Avoiding misconstruals in database systems: A default logic approach. *IEEE Trans. Knowl. Data Eng.*, 5(6):994–996, 1993.

[85] A. S. Hemerly, M. A. Casanova, and A. L. Furtado. Exploiting user models to avoid misconstruals. In R. Demolombe and T. Imielinski, editors, *Nonstandard queries and nonstandard answers: studies in logic and computation*, pages 73–97. Oxford University Press Oxford, UK, 1994.

[86] W. F. Henrique, N. Ziviani, M. Cristo, C. Carvalho, E. S. de Moura, and A. S. Silva. A new approach for verifying url uniqueness in web crawlers. In *18th International Symposium on String Processing and Information Retrieval*, 2011.

[87] B. F. Karlsson, S. D. J. Barbosa, A. L. Furtado, and M. A. Casanova. A plot-manipulation algebra to support digital storytelling. In *Proceedings of the 8th International Conference on Entertainment Computing, ICEC 2009, Paris, France, September 3-5, 2009*, pages 132–144, 2009.

[88] H. A. Kautz. A formal theory of plan recognition and its implementation. In R. J. Brachman, J. F. Allen, H. A. Kautz, R. N. Pelavin, and J. D. Tenenberg, editors, *Reasoning About Plans*, pages 69–124. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.

[89] J. P. Kitajima, B. Ribeiro-Neto, M. D. Resende, and N. Ziviani. Distributed parallel generation of indices for very large databases. In *Proceedings of 3rd IEEE Intl. Conference on Algorithms and Architectures for Parallel Processing*, pages 745–752, 1997.

[90] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto. Learning to advertise. In *Proceedings of 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 549–556, 2006.

[91] A. Lacerda, R. L. T. Santos, A. Veloso, and N. Ziviani. Improving daily deals recommendation using explore-then-exploit strategies. *Information Retrieval Journal*, 18(2):95–122, 2015.

[92] A. Lacerda, A. Veloso, R. L. T. Santos, and N. Ziviani. Context-aware deal size prediction. In *Proceedings of the 21st International Symposium on String Processing and Information Retrieval*, pages 256–267, 2014.

[93] A. Lacerda, A. Veloso, and N. Ziviani. Exploratory and interactive daily deals recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 439–442, 2013.

[94] A. Lacerda, A. Veloso, and N. Ziviani. Adding value to daily-deals recommendation: Multi-armed bandits to match customers and deals. In *Proceedings of the 2015 Brazilian Conference on Intelligent Systems*, pages 216–221, 2015.

[95] A. H. F. Laender, M. A. Casanova, A. P. de Carvalho, and L. F. G. G. M. Ridolfi. An analysis of SQL integrity constraints from an entity-relationship model perspective. *Inf. Syst.*, 19(4):331–358, 1994.

[96] A. H. F. Laender, C. J. P. de Lucena, J. C. Maldonado, E. de Souza e Silva, and N. Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bulletin*, 40(2):135–145, 2008.

[97] E. S. Lima, V. Gottin, B. Feijo, and A. L. Furtado. Network traversal as an aid to plot analysis and composition. In *Proceedings of the XVI Brazilian Symposium on Computer Games and Digital Entertainment, SBGames 2017, Curitiba, Brazil*, 2017.

[98] A. Marczewski, A. Veloso, and N. Ziviani. Learning transferable features for speech emotion recognition. In *Proceedings of ACM Multimedia, Thematic Workshops*, pages 529–536, 2017.

[99] O. Matos-Junior, N. Ziviani, F. C. Botelho, M. Cristo, A. Lacerda, and A. S. da Silva. Using taxonomies for product recommendation. *Journal of Information and Data Management*, 3(2):85–100, 2012.

[100] G. V. Menezes, J. M. Almeida, F. Belém, M. A. Gonçalves, A. Lacerda, E. S. de Moura, G. L. Pappa, A. Veloso, and N. Ziviani. Demand-driven tag recommendation. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 402–417, 2010.

[101] G. V. Menezes, N. Ziviani, A. H. F. Laender, and V. A. F. Almeida. A geographical analysis of knowledge production in computer science. In *Proceedings of International Conference on the World Wide Web*, pages 1041–1050, 2009.

[102] F. Moraes, R. L. T. Santos, and N. Ziviani. UFMG at the TREC 2016 dynamic domain track. In *Proceedings of the 25th Text Retrieval Conference*, 2016.

[103] F. Moraes, R. L. T. Santos, and N. Ziviani. On effective dynamic search in specialized domains. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 177–184, 2017.

[104] G. Navarro, R. A. Baeza-Yates, E. F. Barbosa, N. Ziviani, and W. Cunto. Binary searching with nonuniform costs and its application to text retrieval. *Algorithmica*, 27(2):145–169, 2000.

[105] G. Navarro, E. S. de Moura, M. S. Neubert, N. Ziviani, and R. Baeza-Yates. Adding compression to block addressing inverted indexes. *Information Retrieval Journal*, 3(1):49–77, 2000.

[106] G. Navarro, J. P. Kitajima, B. Ribeiro-Neto, and N. Ziviani. Distributed generation of suffix arrays. In *Proceedings of 8th Annual Symposium on Combinatorial Pattern Matching*, pages 102–115, 1997.

[107] G. Navarro and N. Ziviani. Documents: Languages and properties. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval (Second edition)*, pages 203–254. Pearson, 2011.

[108] B. P. Nunes, A. A. M. Caraballo, M. A. Casanova, and R. Kawase. Automatically generating multilingual, semantically enhanced, descriptions of digital audio and video objects on the web. In *Proceedings of the 16th Annual Conference on Advances in Knowledge-Based and Intelligent Information and Engineering Systems, San Sebastian, Spain, 10-12 September 2012*, pages 575–584, 2012.

[109] B. P. Nunes and M. A. Casanova. A frame-based system for automatic classification of semi-structured data. *Revista de Informática Teórica e Aplicada*, 16:87–92, 2010.

[110] A. Pereira-Jr, R. A. Baeza-Yates, N. Ziviani, and J. Bisbal. A model for fast web mining prototyping. In *Proceedings of 2nd ACM International Conference on Web Search and Data Mining*, pages 114–123, 2009.

[111] H. Piccinini, M. Lemos, M. A. Casanova, and A. L. Furtado. W-ray: A strategy to publish deep web geographic data. In *Advances in Conceptual Modelling - Applications and Challenges, Proceedings of the ER 2010 Workshops ACM-L, CMLSA, CMS, DE@ER, FP-UML, SeCoGIS, WISM, Vancouver, BC, Canada, November 1-4, 2010*, pages 2–11, 2010.

[112] T. Pimentei, A. Veloso, and N. Ziviani. Unsupervised and scalable algorithm for learning node representations. In *Proceedings of International Conference on Learning Representations, Workshop Track*, 2017.

[113] T. Pimentel, M. Monteiro, J. Viana, A. Veloso, and N. Ziviani. A Generalized Active Learning Approach for Unsupervised Anomaly Detection. *ArXiv e-prints*, May 2018.

[114] T. Pimentel, A. Veloso, and N. Ziviani. Fast node embeddings: Learning ego-centric representations. In *Proceedings of International Conference on Learning Representations, Workshop Track*, 2018.

[115] T. Pimentel, J. Viana, A. Veloso, and N. Ziviani. Fast and effective neural networks for translating natural language into denotations. In *Submitted*, 2018.

[116] B. Pôssas, N. Ziviani, W. Meira, and B. Ribeiro-Neto. An efficient approach for correlation-based ranking. *ACM Transactions on Information Systems*, 23(4):397–429, 2005.

[117] B. Pôssas, N. Ziviani, W. Meira Jr., and B. Ribeiro-Neto. Set-based model: A new approach

for information retrieval. In *Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 230–237, August 2002.

[118] B. Pôssas, N. Ziviani, B. Ribeiro-Neto, and W. Meira Jr. Processing conjunctive and phrase queries with the set-based model. In *Proceedings of 11th International Symposium on String Processing and Information Retrieval*, pages 171–183, 2004.

[119] S. Ribas, B. Ribeiro-Neto, R. L. T. Santos, E. de Souza e Silva, A. Ueda, and N. Ziviani. Random walks on the reputation graph. In *Proceedings of the Intl. Conference on The Theory of Information Retrieval*, pages 181–190, 2015.

[120] S. Ribas, B. A. Ribeiro-Neto, E. de Souza e Silva, A. H. Ueda, and N. Ziviani. Using reference groups to assess academic productivity in computer science. In *Proceedings of the 24th International Conference on World Wide Web - Companion Volume*, pages 603–608, 2015.

[121] M. T. Ribeiro, A. Lacerda, A. Veloso, and N. Ziviani. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the ACM Conf. on Recommender Systems*, pages 19–26, 2012.

[122] M. T. Ribeiro, N. Ziviani, E. S. de Moura, I. Hata, A. Lacerda, and A. Veloso. Multi-objective pareto-efficient approaches for recommender systems. *ACM Transactions Intelligent Systems and Technology*, 5(4):53:1–53:20, 2014.

[123] B. Ribeiro-Neto, E. S. de Moura, M. S. Neubert, and N. Ziviani. Efficient distributed algorithms to build inverted files. In *Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 105–112, 1999.

[124] A. S. R. Santos, C. R. de Carvalho, J. M. Almeida, E. S. de Moura, A. S. da Silva, and N. Ziviani. A genetic programming framework to schedule webpage updates. *Information Retrieval*, 18(1):73–94, 2015.

[125] C. S. Santos, T. S. E. Maibaum, and A. L. Furtado. Conceptual modeling of data base operations. *International Journal of Parallel Programming*, 10(5):299–314, 1981.

[126] C. S. Santos, E. J. Neuhold, and A. L. Furtado. A data type approach to the entity-relationship approach. In *Proceedings of the 1st International Conference on the Entity-Relationship Approach*, pages 103–119, 1979.

[127] P. C. Saraiva, E. S. de Moura, R. C. Fonseca, W. M. Jr., B. Ribeiro-Neto, and N. Ziviani. Rank-preserving two-level caching for scalable search engines. In *Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 51–58, 2001.

[128] F. Saussure. *Cours de Linguistique Generale*. Payot, 1995.

[129] U. Schiel, A. L. Furtado, E. J. Neuhold, and M. A. Casanova. Towards multi-level and modular conceptual schema specifications. *Inf. Syst.*, 9(1):43–57, 1984.

[130] G. J. Sena and A. L. Furtado. Towards a cooperative question-answering model. In *Proceedings of the Third International Conference on Flexible Query Answering Systems, FQAS'98, Roskilde, Denmark, May 13-15, 1998*, pages 354–365, 1998.

[131] A. S. Silva, A. H. F. Laender, and M. A. Casanova. An approach to maintaining optimized relational representations of entity-relationship schemas. In *Proceedings of the 15th International Conference on Conceptual Modeling, ER'96, Cottbus, Germany, October 7-10, 1996*, pages 292–308, 1996.

[132] A. S. Silva, A. H. F. Laender, and M. A. Casanova. On the relational representation of complex specialization structures. *Inf. Syst.*, 25(6-7):399–415, 2000.

[133] I. Silva, B. Ribeiro-Neto, P. Calado, E. S. de Moura, and N. Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–103, 2000.

[134] J. L. Szwarcfiter, G. Navarro, R. A. Baeza-Yates, J. Oliveira, W. Cunto, and N. Ziviani. Optimal binary search trees with costs depending on the access paths. *Theoretical Computer Science*, 290(3):1799–1814, 2003.

[135] L. Taniguchi, G. Salgado, J. Vieira Jr, J. Viana, and N. Ziviani. Machine learning for prediction of mortality and prolonged length of stay at intensive care unit in a brazilian cohort of critically ill patients. In *Proceedings of 37th International Symposium on Intensive Care and Emergency*, 2017.

[136] L. Tucherman, M. A. Casanova, and A. L. Furtado. The CHRIS consultant-a tool for database design and rapid prototyping. *Inf. Syst.*, 15(2):187–195, 1990.

[137] L. Tucherman, M. A. Casanova, P. M. Gualandi, and A. P. Braga. A proposal for formalizing and extending the generalization and subset abstractions in the enity-relationship model. In *Proceedings of the Eight International Conference on Entity-Relationship Approach, Toronto, Canada, 18-20 October, 1989*, pages 27–41, 1989.

[138] D. Valle, A. Veloso, and N. Ziviani. Effective fashion retrieval based on semantic compositional networks. In *Proceedings of IEEE International Conference on Neural Networks, Rio de Janeiro, Brazil, July 08-13*, 2018.

[139] P. A. S. Veloso, J. M. V. de Castilho, and A. L. Furtado. Systematic derivation of complementary specifications. In *Proceedings of the Seventh International Conference on Very Large Data Bases - Volume 7*, VLDB '81, pages 409–421. VLDB Endowment, 1981.

[140] P. A. S. Veloso and A. L. Furtado. Towards simpler and yet complete formal specifications. In *TFAIS*, pages 174–188, 1985.

[141] R. Wilensky. *Planning and Understandinga Computational Approach to Human Reasoning.* Addison-Wesley, 1983.

[142] M. E. Winston, R. Chaffin, and D. Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444, 1987.

[143] N. Ziviani. *Projeto de Algoritmos e Estruturas de Dados.* Editora Unicamp, 1986.

[144] N. Ziviani. *Projeto de Algoritmos com Implementações em Pascal e C (First edition)*. Pioneira Thomson, first edition, 1993.

[145] N. Ziviani. Text operations. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval (First edition)*, pages 163–190. Addison-Wesley, 1999.

[146] N. Ziviani. *Projeto de Algoritmos com Implementações em Pascal e C (Second edition)*. Thomson learning, 2004.

[147] N. Ziviani. *Projeto de Algoritmos com Implementações em Java e C++*. Thomson, 2007.

[148] N. Ziviani. *Projeto de Algoritmos com Implementações em Pascal e C (Third edition)*. Cengage Learning, 2010.

[149] N. Ziviani, E. S. de Moura, G. Navarro, and R. Baeza-Yates. Compression: A key for next-generation text retrieval systems. *IEEE Computer*, 33(11):37–44, 2000.

[150] G. Zuin, L. Chaimowicz, and A. Veloso. Learning transferable features for open-domain question answering. In *Proceedings of the IEEE International Conference on Neural Networks*, 2018.