Research on Complex Data Management and Analysis at UFSC

Ronaldo dos Santos Mello, Renato Fileto, Carina F. Dorneles, Vania Bogorny

Universidade Federal de Santa Catarina, Florianópolis, Brazil
{ronaldo,fileto,dorneles,vania}@inf.ufsc.br

Abstract. Since the end of the 20th century data have been gathered in an unprecedented and growing scale, by using a variety of devices and information systems, ranging from sensors to the social Web. The resulting heterogeneous collections of unconventional data, particularly complex data (XML, Web data, maps, time series, etc.), cannot be constrained just to relational tuples, and require new methods for data handling and information extraction. This article introduces the research efforts on unconventional complex data management and analysis being developed by the database group of the Universidade Federal de Santa Catarina. First, it gives a brief description of the group. Second, it presents some of the main research efforts and directions of the group, in the topics of data modeling, data matching, information retrieval, and spatial and spatio-temporal data analysis. Third, it discusses the impact of these research efforts, as well as the interactions of the group with other institutions. Finally, it points out some research perspectives and collaborations with academia and the industry.

Categories and Subject Descriptors: H.2.m [**Database Management**]: Miscellaneous; H.2.8 [**Database Management**]: Database Applications; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; E.m [**Data**]: Miscellaneous

Keywords: complex data, data modeling, approximate data matching, information retrieval, spatio and spatio-temporal data analysis.

1. INTRODUCTION

The Database Group (*GBD*) of the Universidade Federal de Santa Catarina (*UFSC*)[1] is dedicated to the research and development of solutions in complex unconventional data management systems. It is a young group that started its activities in 2006 and was expanded in 2009, with the addition of two faculty members (all of them with PhD in Computer Science). Nowadays, two members have CNPq Productivity Fellowship and the group has several graduate and undergraduate students, as well as collaborators, working either partial or full-time in two labs. The group works at the Informatics and Statistics Department (INE) of UFSC, located at Florianópolis, in Southern Brazil. Florianópolis is today a technological center of the IT industry, that in 2008 had already more than 600 PhDs[2]. INE/UFSC offers two bachelor degrees, Computer Science and Information Systems, with an average of 100 new graduates per year, which is not enough to supply the local industry needs. It also hosts a graduation program in Computer Science (PPGCC).

[1] http://www.gbd.inf.ufsc.br
[2] http://veja.abril.com.br/081008/p_158.shtml

Our work has been supported by CNPq (research grants 550845/2005-4, 481392/2007-6, 307588/2008-4, 152029/2010-9, 307992/2010-1, 476857/2010-4, 481569/2010-3), FAPESC (research grants 12.552/2007-0, PRONEX 93/2008, 005/2009-PPP), CAPES, FEESC, FAPEU, the UnA-SUS program of the Brazilian Ministry of Health, and the MODAP project financed by the European Community (EU FET-OPEN 2009-2012).

The research of our group is funded by CNPq and CAPES, which are the main federal research agencies in Brazil, as well as FAPESC, our state research foundation. UFSC also provides an important support for our research by means of the Dean of Research and Extension (PRPG), the Foundation of Research and Extension (FAPEU), and the Santa Catarina Foundation of Engineering Teaching (FEESC). The first one financially helps research and extension projects. The latter are organizations linked to UFSC that manage resources from projects in cooperation with the industry and government. Furthermore, we have received support from industry and government institutions.

The current efforts of the group are synchronized with the grand research challenges in Computer Science in Brazil, as manifested by the Brazilian Computer Society (SBC) in 2006 [Medeiros 2008]. From the five proposed challenges, we try to contribute to the first one, which focuses on *Management of information over massive volumes of distributed multimedia data*. Specifically, we focus on research related to the treatment of large collections of complex unconventional data. *Complex data* can be roughly defined as any kind of data that cannot be represented just by simple data types (strings, numbers, etc.). Examples of complex data include documents, data on the Web, images, temporal series, spatial and spatio-temporal data, among others. Much of these data are unconventional, in the sense that they are not constrained to conventional data models, like the relational model. Our current activities focus in four main research subareas: data modeling, approximate data matching, information retrieval, and spatial and spatio-temporal data analysis.

Data modeling sub-area aims at solving problems related to the suitable representation of non-conventional data, which is becoming more and more available in large data sources. Here we are not only interested in data (static) modeling, but also in related research topics, like dynamic modeling (e.g., data constraints), indexing, design methodologies and model mappings. Solutions in this area are useful as a basis to solve problems related to other areas, such as information retrieval, data integration, and data mining problems. For indexing purposes, for example, we have to model relevant object properties in order to design appropriate data structures to index them. We are working on it for Web form data, specifically on fields in form interfaces that provide search conditions over hidden databases. Another example is data integration task, in which people usually define a canonic model that represents a set of local and heterogeneous data schemata, and specify mappings between these models. We have been working on it for XML data. In this sub-area, we have contributions or ongoing research related to XML data and constraints modeling [Mello and Heuser 2005; Rodrigues and Mello 2007; Schroeder and Mello 2009], Web forms' modeling and indexing of data and constraints [Mello et al. 2010; Gonçalves et al. 2011], logical modeling of cloud data stores, conceptual modeling of georreferenced trajectories [Bogorny et al. 2010], and ontology modeling applied to data warehouses for geographic data [Deggau et al. 2010; Filho et al. 2010]. The research on this sub-area has CAPES and CNPq financial support, as well as collaborations with prof. Juliana Freire (New York University, USA) and prof. Carmem Hara (UFPR, Brazil).

Approximate data matching focuses on defining and developing techniques and algorithms to manage data that represent the same real world object but originate from different data sources. Data matching processes can be applied to several data management research areas, such as query/search over data sources (databases or Web data sources), integration of heterogeneous data sources, and so on. In all of these scenarios, data may contain inconsistencies like different conventions, missing fields, typographic, typing and grammatical errors, which must be solved before being stored or processed. Such a problem is very common on the Web, specially because data are generated by different people or software. Some contributions [Dorneles et al. 2007; Dorneles et al. 2009; Dorneles et al. 2011] in this context have been done, including those in collaboration with of Prof. Renata Galante, from UFRGS [Moro et al. 2009; Manica et al. 2010], and Angelo Frozza, from IFC-Camboriu [Frozza and Mello 2007], both from Brazil. Research in this sub-area is part of a project partially funded by a CNPq Research Grant and also includes an initial external cooperation with Prof. Marco A. Winckler, from IRIT, France and Prof. Carla Freitas, from UFRGS, Brazil. We have contributed in terms of defining similarity metrics for complex data[Dorneles et al. 2004], providing interoperability of geographic data, executing data

Table I. Summary of the *GBD/UFSC*'s Research Topics in each Sub-area

Data Modeling	Approximate Data Matching	Information Retrieval	Spatial and Spatio-Temporal Data Analysis
XML data	Metrics for atomic and complex data	Metadata-based IR	Trajectories
Web data	Similarity-based search for Web forms	Content-based IR	Spatial, temporal, and semantic extensions for data warehouses
Cloud data stores	Temporal data search	Semantic annotation	Semantic enriching of trajectories
Analytic data	Approximate structured queries over the Web	Semantic search	Trajectories data mining
	Matching OWL and GML schemas	Context for IR	Spatial networks
	Data cleaning for XML and Web form matching	Hybrid IR (via metadata and contents)	
	Approximate Web data search	Hybrid indexing	

cleaning in XML instances for matching purposes [Gonçalves and Mello 2007], performing integration of XML instances and implementing similarity search for Web forms [Gonçalves et al. 2011]. Other initial results of this research will be applied to the context of the problems solved in the InCod[3], a National Institute of Science and Technology, which aims at, among other things, searching metadata extracted from heterogeneous medical images.

Information retrieval from collections of complex data can be done via metadata and contents similarity. However, neither of these approaches alone is suitable for many applications. On the one hand, annotating large collections of complex data with metadata is a labor intensive task, and some complex data just cannot be properly and completely described by metadata. On the other hand, similarity search requires the automatic extraction of information from contents, and it is hampered by the semantic gap between what current systems can automatically handle, and what is relevant for the user. Our research on information retrieval has exploited the semantics expressed in artifacts like ontologies and annotations [D'Agostini and Fileto 2009; da Rocha et al. 2009; Rigo et al. 2010; Rigo et al. 2011], as well as the users' context related to these artifacts [D'Agostini et al. 2008; Fasolin et al. 2009]. Relevance feedback [Ruthven and Lalmas 2003] enables keeping track of the user's semantic context, i.e., her context related to the ontology used to annotate complex data and process semantic searches. Spreading activation [Crestani 1997] on the user's semantic context and the underlying ontology help to automatically solve ambiguities and semantically expand queries, in order to retrieve results with higher levels of recall and precision than traditional lexical and syntactical approaches. This research has been supported by CNPq and the Brazilian Ministry of Health, with most projects involving collaborations with other computer scientists and domain experts, from institutions contributing to InCoD or just having interest in applications of this research.

The subarea of *spatial and spatial-temporal data analysis* focuses on data warehouses and data mining algorithms. We have proposed spatial extensions [Moreno et al. 2009a; Filho et al. 2010], temporal extensions [Moreno et al. 2009b; Moreno et al. 2010], and semantic extensions [Deggau et al. 2010; Filho et al. 2010] to support information analysis in data warehouses. The research in this area has been done in collaboration with Dr. Francisco Moreno, from the Universidad Nacional

[3]http://www.incod.ufsc.br/?lang=en

de Colombia at Medellín. Part of this research has also been done in collaboration with the industry, in sectors like agriculture, energy, and natural disasters prevention and management, with financial support from companies in the respective sectors and CEPED (Brazilian Agency of Studies and Research in Disasters). In spatio-temporal data analysis the focus has been on trajectories of moving objects, developing new methods for adding semantics to trajectories, and new spatio-temporal data mining algorithms. Because of the price reduction of mobile devices as GPS and cell phones, an explosion of data generated by these devices has become available. Mobile devices can capture the position of an object in time, therefore generating a trajectory of the moving object. As such data have a lot of noise, and no explicit meaning associated to it, intelligent and novel techniques are needed to process, semantically enrich, and extract information and knowledge from these data, which can be of interest for different application domains such as human mobility analysis, animal tracking, traffic analysis, and so on. Our research in this area has been supported by CNPq, FAPESC, and the European Union. Research in this area has collaboration with Dr. Chiara Renso from the University of Pisa, Prof. Valéria Cesário Times from the Universidade Federal de Pernambuco (UFPE), Prof. Fernando José Braz from Univille and Prof. Luis Otavio Alvares at UFSC. Several results came out of these collaborations, and will be detailed in Section 5.

Table I summarizes our research topics in each subarea. In the remaining of this article, we go deeper into these specific research areas. Sections 2, 3, 4 and 5 present the research in data modeling, approximate data matching, information retrieval, and spatio and spatio-temporal data analysis, respectively. Section 6 is dedicated to the conclusion and future perspectives for *GBD/UFSC* research.

2. DATA MODELING

Challenges in complex data modeling are mainly related to the definition of effective conceptual and logical structures for data that do not have a simple or regular structure, like XML and Web data. By effective structuring we mean well designed data schemata that could capture their semantics, including dynamic properties like constraints and relationship types. Examples of complex data modeling are element hierarchies for XML data, attribute dependencies for Web forms, as well as mappings between data schemata and contents. The same reasoning holds for indexing structures to access such data in terms of performance. In the following, we detail our research interests and results in this subarea, including past and current projects.

XML data management is the first internal *GBD/UFSC* project with financial aid of PPGCC and Technological Center (CTC) of UFSC. This project is motivated by the increasing availability of XML sources and the need to manage them by several applications [Moro et al. 2009]. A system for XML schema integration called *BInXS* is the first research effort of this project [Mello and Heuser 2005]. *BInXS* aims at generating a global conceptual schema from a set of local XML schemata. The research on this system had generated contributions related to a reverse engineering process of an XML schema to a conceptual schema, a conceptual-to-XML schema mapping strategy, and an integration method suitable to conceptual representations of semistructured data. A prototype for *BInXS* is in final stage of development. This project also focuses on a schema design methodology for XML sources that extends traditional relational database modeling methodologies based on the Entity-Relationship model [Schroeder and Mello 2009]. The main contribution of this work is a *workload-based* design process. It takes into account information about transactions' workload of an application with the purpose of generating XML structures that provide fast access to the data manipulated by the transactions in the XML sources. A prototype tool that supports this methodology was implemented, being awarded in the *Regional Database School (ERBD)* event that occurs yearly in Southern Brazil [Lima et al. 2008].

The *Digitex* project, in collaboration with UFRGS and UCPel, Brazil, and funded by CNPq, was an initiative to allow indexing and further accessing to digital library documents that could be encoded in XML. In that context, we have developed an approach called *DIInCX*, that complements schema

information of these documents with integrity constraints' properties [Rodrigues and Mello 2007]. *DIInCX* applies data mining techniques over a sample of XML documents to extract constraints that are stored in *SWRL* language for further access. Experiments with a tool that performs the *DIInCX* activities have revealed promising results even for a small sample of XML documents.

Another research effort regards Web data. With partial CAPES support, we have a partnership with Prof. Juliana Freire, from New York University, related to Web forms' modeling and querying. Web forms are query interfaces for hidden databases in the Web (the *Deep Web*) which provide useful services in several domains, like car dealers and flight booking. We are working on the modeling of Web form general properties that are relevant to be queried, as well as on the design of index structures based on these properties. Preliminary experiments on a relational data sample with more than 3,000 forms in eight domains had produced good results in terms of performance, if compared to traditional relational indexes [Mello et al. 2010].

We also just started working, in another partnership with Prof. Carmem Hara (UFPR, Brazil), on modeling cloud data storage. The focus is on fragmentation design of these data stores by considering XML data stored on them. We propose an extension of the schema design methodology for XML sources previously presented, considering now a distributed architecture and workload information available for each node. This work is currently supported by a PhD scholarship from CAPES.

3. APPROXIMATE DATA MATCHING

An approximate matching process aims at defining whether two data represent the same real world object or not. This is a complex problem when the data come from different sources, because they differ from each other in the way they are represented. Heterogeneity can happen in data structure as well as in data value, so a data matching process must be able to analyze both structure and data value. The problem of approximate data matching has motivated research on various areas and contexts, and has been extensively investigated and surveyed in the literature [Dorneles et al. 2011]. Some application areas are data integration, data cleaning, duplicate data reduction in digital libraries, similarity-based queries processing (or approximate queries), including similarity joins. These data may contain inconsistencies, like different conventions, missing fields, typographic, typing and grammatical errors, which must be solved before being stored. Such situations are very common in the Web, specially in cases in which data is generated by various people or software. Because there is no way to assure that two data instances refer to the same object in the real world, a process for matching data must use a similarity function to compare their values.

For XML data sets [Moro et al. 2009], this issue should be appropriately treated since in such databases the data structure is organized in multiple levels, and can involve collections of values. These features make XML data sets underlying details even harder to grasp for external users. On accessing such an XML data set, a query processor must be able to properly treat multi-level data and collection of values. However, while related work [Motro 1988; Bilenko and Mooney 2003; Chaudhuri et al. 2007] has combined metrics for flat structures, we have studied the combination of atomic similarity metrics for nested structures, particularly to collections of values, typically found in XML documents. We have proposed a new approach [Dorneles et al. 2004] for the use of similarity metrics as a step forward when querying XML databases. Considering that XML elements may be atomic or complex, we have defined two types of similarity metrics: MAV (metrics for atomic values) and MCV (metrics for complex values). The MAVs are used with atomic XML elements, depending on the application domain of their values. The MCVs are used with complex elements. This work was developed in cooperation with Prof. Altigran da Silva and Edleno S. de Moura (UFAM, Brazil) and Carlos A. Heuser (UFRGS, Brazil).

Two objects represent the same real world object if the similarity score returned by a function is greater than a predefined threshold value, whose choice is not trivial. The scores returned by a

function depend on the internal details of the algorithm that implements it. They usually have no meaning to the user. As a consequence, the user chooses a value that is hardly satisfactory. Moreover, as the score values returned by different functions have different value distributions, the quality of the result may vary from one function to another when a specific threshold is considered. This means that a threshold value that has been predefined for a specific function may not be adequate to another. We have proposed a strategy for allowing meaningful and comparable scores [Dorneles et al. 2007]. The main idea is that instead of defining the threshold in terms of the scores returned by a similarity function, the user specifies the precision (*adjusted score*) expected from a matching process. Precision is a well known quality measure and has a clear interpretation from the user's viewpoint. This work was nominated for the best interdisciplinary paper award at ACM CIKM in 2007. We have extensively run experiments in order to show that existing methods for combining scores for computing the similarity between records may be enhanced if adjusted scores are used [Dorneles et al. 2009]. This research was also developed in collaboration with the professors from UFAM and UFRGS.

Similarity search, in the context of Web forms, is an important approach since a lot of hidden databases' interfaces in a same domain have forms with similar structures and filtering capabilities. On supporting this kind of search, a user that is not satisfied with a service provided by a form interface is able to look for other similar forms. We have been developing an environment for similarity searching over Web forms. The strong points of our approach is a similarity metric suitable to form data, and a similarity-based index based on a clustering algorithm. For combining values in a data set, we use a MCV metric, the *SubSetSim* metric, as defined in [Dorneles et al. 2004]. The method has the advantage of reducing the number of index entries because the index is designed to provide one entry per cluster of similar Web form properties. We had just submitted preliminary promising results in terms of precision/recall quality over a sample of 1,000 Web forms to a relevant conference. This research is part of a CNPq project called *WF-Sim* and is developed in collaboration with Rodrigo Gonçalves and Caio D'Agostini, both from industry.

Still on the idea of searching data coming from different sources, we are also working on specifying and constructing techniques and algorithms for accessing them in a structured fashion without considering building semantic evidence *a priori*. In this sense, we have been developing a tool, called *FindMe*, which aims at allowing more expressive queries over HTML documents, using a SQL-like approach. The idea is to enable the user to pose meaningful queries, allowing the construction of predicates (as done in a WHERE clause in SQL language) with approximated search criterion and the specification of objects in which the predicates will be applied. In order to do so, we have defined an index structure for objects such that their attributes and relationships are easily detected. This work was published in *Regional Database School (ERBD)*, in the 2011 edition, and won the best paper award at that event. This research is recent and it has been developed in the context of an ongoing CNPq Project, called *Aglomerante*. The project focuses on manipulating heterogeneous data sources using the dataspace-aware search engines idea, in which there is no semantic integration a priori (schema matching, for instance) in order to pose a structured query. This project is executed in collaboration with Prof. Renata Galante, UFRGS, and Prof. Marco A. Winckler, IRIT/França.

As already mentioned, when data differ from each other in the way they are represented, they may contain inconsistencies, like different conventions. The ability to correctly access such data in an environment like the Web is still a challenge, and a good example is the search for temporal information. In the context of Web, a simple date can be stored in several formats, like ‘‘`2009 - October 12th`’’, “`2009-10-12`”, “`12/10/2009`”. A temporal constraint in a keyword search can be expressed in different ways that are not treated correctly by a conventional search engine. We developed a proposal that, besides identifying temporal information presented in the query, it also recognizes the temporal information contained in the documents [Manica et al. 2010]. Our proposal is based on identifying temporal constraints in a keyword query and intercepting the query processing, executed by a conventional XML search engine, in order to evaluate those constraints. Our approach

allows users to find the temporal information that they are interested in without having to learn a complex query language or needing prior knowledge of the structure of the underlying data. This work has been developed in collaboration with Prof. Renata Galante from UFRGS, and Prof. Edimar Manica, from IFC - Canoas.

In the context of geographical data, we proposed a method for allowing interoperability of GML schemata. GML is an XML-based language for geographic data definition. In this context, we had developed a method for matching different GML schemata based on semantic information about the domain provided by an OWL ontology. GML is an XML-based language for geographic data definition. The contribution here is a prototype tool that provides semi-automatic definition of similarities between OWL and GML elements. These OWL-to-GML schema mappings allow the inference of semantic correspondences between heterogeneous GML elements if they point to the same ontology concept [Frozza and Mello 2007]. This work has been developed in the context of the *XML data management* project, described in Section 2.

We also have contributions related to the integration of XML data. One of them is a set of preprocessing procedures applied to XML instances with the purpose of improving the quality of the scores produced by similarity metrics for XML data [Gonçalves and Mello 2007]. These procedures accomplish several activities, like removal of stop words, stemming, as well as naming standardization based on an ontology for the domain. Experiments on a couple of metrics had raised a promising percentage of quality improvement, and we further extend these approach with a method for unification of similar XML instances [Jr and Mello 2008]. This method is based on a set of integration operators for XML data that homogenizes content and hierarchical structures to simplify the generation of the unified instance. Ontology support is also considered here to rule the transformations on the XML hierarchical structures. The idea is that the XML (logical) hierarchical relationships must be in accordance to the semantic relationships in the Ontology. This research is part of the *Digitex* project.

4. INFORMATION RETRIEVAL

Information retrieval (IR) [Baeza-Yates and Ribeiro-Neto 1999] appears is some of the most popular applications of Computer Science (search engines, digital libraries, etc.), but has challenging open problems yet. Nowadays, the data handled by IR systems can vary from structured tuples, to barely structured documents and multimedia. Huge amounts of information objects, with associated metadata or not, are available in repositories and on the Web. Searchers may be interested in whole information objects, specific chunks of information within some of them, or metadata about these objects. The widespread and growing use of sensor equipment to collect complex data, such as images and temporal series, is contributing to the growing number and the variety of available collections of complex data. It poses new challenges with respect to IR scalability, objects' nature and complexity, and the myriad of domains and applications that must be tackled.

An IR process begins when a user submits an query, which can be specified as: (i) formal and precise statements of information needs, written in some query language (e.g., SQL, XQUERY) or posed in some visual interface; (ii) a set of simple or composite keywords, connected by explicit or implicit operators (AND, OR, NOT, etc.); (iii) data objects serving as examples for retrieving similar ones. Several objects from a collection may match a query, perhaps with different degrees of relevance to the user. Thus, many IR systems compute a score, which estimates how well each returned object from the database matches the query, the needs of who submitted it and/or a particular context. These scores are used to rank the objects in the query answer, presenting top ranked objects first. The user can mark results that she considers relevant. Then, the process may continue in an interactive way. The user can refine queries according to her wishes, inserting or removing keywords, before asking for execution. The system can gradually improve the ranked results, by using what it is able to learn about the users' preferences, as she poses queries and marks relevant results.

Our research in the area of information retrieval follows some ideas of our previous works, that exploit the formal description of the semantics of metadata and information contents with ontologies and semantic annotations in order to efficiently solve semantic search [Fileto et al. 2003; Fileto et al. 2005]. We also incorporate mechanisms based on relevance feedback [Ruthven and Lalmas 2003] to capture information about the user's context [Mani and Sundaram 2007]. It helps to improve query refinement, ranking, and the users' experience with IR systems, as illustrated by the description of some of our works in the following subsection.

4.1 Keyword-based Search exploiting the User's Semantic Context

The information contents that may satisfy a user depend on her personal knowledge, preferences, and the specific semantics intended by her when submitting a query [Challam et al. 2007]. However, the evolving individual context related to shared knowledge bases or ontologies has not been exploited by typical semantic search systems [Mangold 2007]. Thus, we have pursued this theme in a project sponsored by CNPq's universal program, whose main results are described below.

In [D'Agostini et al. 2008] we propose a contextual semantic search system for processing keyword-based searches according to evolving user's interests. This system gradually learns contextual information related to the ontology used to annotate the contents. We represent this user's semantic context as graph [D'Agostini and Fileto 2009]. Each node of this graph is mapped to a specific concept or instance of an underlying knowledge base. Each edge represents a connection between topics, established by the user when including in the same query keywords referring to the pair of topics linked by the edge. This user-specific semantic context layer, on top of the shared knowledge layer, enables the system to provide results better aligned to each user's interests.

Our algorithms for processing keyword-based searches [D'Agostini and Fileto 2009] exploit the user's semantic context to disambiguate and semantically extend queries. These algorithms, based on ant colony optimization (ACO) and spreading activation, are integrated to a wider process that includes the user's interactions with the system for posing queries and analyzing the results. The user can explicitly disambiguate keywords and check results in semantically classified ranked lists, as necessary or desired. Whenever a user poses a query, disambiguates a keyword, or checks results, her semantic context is updated accordingly, by applying the relevance feedback. As the system collects more semantic context information about a user, it can produce more relevant results for that user.

We have developed a prototype of this system, called *Praestro*[4], in order to test our approach [Fasolin et al. 2009]. In the experiments done to evaluate our proposal, users were asked to search for Wikipedia documents, by interacting with Praestro, which was configured to use DBPedia as the shared ontology. The results of these experiments showed slight gains in the alignment of the top ranked results provided by Praestro with the user's interests. Praestro's was also able to retrieve results that were semantically related but not lexically related to any of the keywords provided in the users' searches. In the future, we plan to test Praesto in some specific domains, with annotations being created by using the systems described in the next subsection.

4.2 Semantic Annotation and Semantic Search in Digital Libraries

Support for efficient semantic annotation can contribute to contents description and support sophisticated ways to understand and retrieve complex objects in digital libraries. We have developed some projects in this area in close cooperation with domain experts, aiming to generate useful applications and tools. DLNotes [da Rocha et al. 2009] is a tool that we have been developing to support semantic annotation of documents, in cooperation with Prof. Roberto Willrich, an expert in digital libraries. This tool is part of the goals of an ongoing thematic project jointly financed by CNPq and FAPESC,

[4]http://www.lisa.ufsc.br/projects/praestro

through the PRONEX program, and coordinated by Prof. Alckmar Luiz dos Santos, from the Center for Research in Computer Science, Literature and Linguistics (NUPILL/UFSC).

DLNotes adopts an extension of the Annotea Schema, which allows readers of documents to make both free-text and semantic annotations of the documents contents. A semantic annotation relates a portion or a string (with one or more occurrences) selected in the text (e.g., all occurrences of *Romeo*) with concepts described in an ontology (e.g., *Man*), or a particular instance of a concept (e.g., *Juliet* an instance of *Woman*). DLNotes users can insert instances in the same knowledge base where the semantic annotations are kept, as they identify these instances in the document, during the annotation process. The users can also establish relationships between instances (e.g., *Juliet hasCousin Tybalt*). Any user can see and extend the collection of instances, instances connections, and annotations inserted in the knowledge base and validated by curators. Properties of classes define the possible types of relationships that can be inserted between pairs of their respective instances (e.g., family relations can be allowed between instances of *Person*, and the relation *bornIn* from an instance of *Person* to an instance of *Place*).

DLNotes can be easily embedded in different digital libraries, and used with different domain ontologies. It has been integrated with the Digital Library of Brazilian Literature[5], in order to demonstrate and evaluate its functionalities. Literature students and teachers have been able to cooperatively use DLNotes, with a literature ontology, to build intricate semantic networks about the contents of plays and novels of this digital library.

In another ongoing project within UnA-SUS[6], a research and development program sponsored by the Brazilian Ministry of Health, we have been developing techniques and tools based on domain knowledge to support semantic annotation and semantic retrieval on large complex data collections [Rigo et al. 2010; Rigo et al. 2011; Fileto et al. 2011] . We have employed visualization techniques to support navigation on large ontologies, and developed an efficient component for automatically completing keywords, as the user types characters, with alternative terms from the knowledge base that are lexically or semantically related to what has been typed. Tests with real users annotating multimedia objects, intended for e-learning activities in the health area and stored in a Web repository[7], showed average gains of around 51% in annotation time, by using the proposed semantic support [Rigo et al. 2011]. Other experiments showed that proper configuration values for spreading activation parameters enable its efficient execution with vast knowledge bases and data collections in practical situations [Fileto et al. 2011].

5. SPATIAL AND SPATIO-TEMPORAL DATA ANALYSIS

5.1 Data Warehouses with Spatial, Temporal, and Semantic Extensions

Data warehouses can be extended to allow spatial objects as members of dimensions, and as measures of the fact table [E. Malinowski 2008]. The exploitation of these spatial extensions requires geographical operators and special aggregation functions. Spatial operators can filter spatial measures based on topological, metric, and position relations between these measures and spatial members (e.g., geographical points of accidents which are inside a city). Specific spatial aggregation functions are necessary to handle filtered spatial measures in some applications. Temporal extensions, by their turn, can be useful to exploit temporal associations between members, different kinds of time, and measures that change as time goes on (e.g., the position of moving objects in a data warehouse about their trajectories). Semantics formalized in spatial, temporal, and domain ontologies, can be employed to describe the structure, contents, operators, and aggregate functions available in a data warehouse,

[5]http://www.literaturabrasileira.ufsc.br
[6]http://unasus.ufsc.br
[7]http://repositorio.unasus.ufsc.br

helping users to solve their information needs by posing queries in a higher abstraction level. In the following we describe some of our results related to these themes.

In cooperation with a visitor researcher sponsored by the Colombian government, we developed an extension of the map cube operator for supporting other aggregation functions besides geographical union [Moreno et al. 2009a]. It allows the construction of new geometries as the result of spatial OLAP operations (e.g., center of mass, convex hull, or Voronoi diagram).

The same cooperation also produced a dimensional model with extensions for keeping track of reclassifications, i.e., members of dimension levels changing parents (e.g., players changing their teams, teams changing divisions, a hurricane moving from one region to another one) [Moreno et al. 2009b]. Then, using this model, [Moreno et al. 2010] introduced the notion of season (a continuous period of time in which a member from a dimension level is associated to a member of a higher level of the same dimension), proposed a formal OLAP operator that enables season queries in a concise and simple way, and showed how it can be embedded in a multidimensional query language. Using the season operator one can pose queries such as "give the number of goals scored by player p in each one of his seasons with team t".

A project in cooperation with the Agricultural Research and Rural Extension Institute of Santa Catarina (EPAGRI), supported by FAPESC, allowed us to explore the use of ontologies to help users find data marts related to domain specific keywords, explore their contents, and compose spatial operators and aggregation functions to build spatial OLAP queries, by using a knowledge-based graphical user interface (GUI) [Deggau et al. 2010]. The results of usability tests of the proposed GUI suggested that it needs some improvements, but meets users' needs. Through that GUI, users from the agriculture domain, without training in OLAP or geographical data processing, were able to build spatial OLAP queries using one topological operator to filter measures. However they had difficulties with the composition of spatial operators (e.g., buffer and intersects).

Finally, in a cooperation with CELESC, the Eletricity Distribution Company of Santa Catarina State, we developed an extension of the Malinowski multidimensional model [E. Malinowski 2008] to keep dynamic measures associated to spatial elements which are connected in a network. It provides the basis for analyzing temporal series of data about the state of interconnected elements, using spatial OLAP operators, and the visualization of the evolving state of portions of the network on maps [Filho et al. 2010]. Tests in a case study using a portion of the CELESC's electricity distribution network allowed the identification of some tendencies and recurring patterns in the demand and load on specific equipments and portions of this spatial network.

5.2 Spatial and Spatio-Temporal Data Mining

Research on spatial data mining addressed the idea of integrating *database modeling* with *database mining*. While on the one hand database design has the objective of modeling data and their previously known relationships, data mining on the other hand has the objective to discover patterns (implicit relationships among data) that are previously unknown. However, by grouping a set of relations in a database for data mining, large amounts of previously known relationships come together, generating well known, obvious, and uninteresting patterns. These well known patterns extracted from relationships, apart from not being interesting, are mixed among large amounts of patterns, making their analysis very difficult from the user's point of view. In order to reduce this number of uninteresting patterns, using the knowledge stored in database schemas to eliminate well known relationships in spatial data mining has become obvious, and has been used for reducing spatial association rules [Bogorny et al. 2008] and closed frequent geographic itemsets [Bogorny et al. 2010]. In general words, the idea is to consider semantics in data mining. This work gave to a member of the group the prize of best Brazilian PhD Thesis in 2007, awarded by the Brazilian Computer Society.

The research on spatio-temporal data analysis, specifically on trajectories of moving objects, has

focused on three main directions: (i) trajectory data and pattern modeling, (ii) trajectory semantic enrichment, and (iii) trajectory data mining. These topics are being explored in five research projects (AATOM - funded by FAPESC, ATACT - funded by CNPq, MMT - funded by CNPQ, MODAP - funded by the European Union, and SEEK - funded by the European Union (FP7 Program)). In these topics there has been a collaboration between UFSC and II/UFRGS, KDD-Lab Pisa, and UFPE.

Knowledge discovery in databases has become very popular in the last years, and it is well known that data preprocessing is the most effort and time consuming step in the discovery process. In part, it is because database designers do not think about data mining during the conceptual design of a database, therefore data are not prepared for mining. This problem increases for spatio-temporal data generated by mobile devices, which involve both space and time. To overcome this problem, one solution was the proposal of the first data mining query language for trajectories [Bogorny et al. 2009]. This language, apart from considering different data mining tasks and the semantics of trajectories, supports automatic data transformations to multiple levels of granularity, which is an important issue in data mining. Another solution to reduce the gap between databases and data mining in the domain of trajectories is proposed with a conceptual data model (meta-model) for modeling trajectories with the focus on mining [Bogorny et al. 2010]. In other words, when designing a spatio-temporal database the user is already modeling the data for mining. This work had a collaboration with Prof. Carlos A. Heuser, from UFRGS.

The second research topic has a collaboration with Dr. Chiara Renso at KDD-Lab Pisa, Prof. Valeria Cesario Times, from UFPE, and Prof. Luis Otavio Alvares from UFSC. The main idea in this topic is to develop new data preprocessing and mining methods to enrich trajectories with semantic and context information. Trajectory data generated by mobile devices are simple points located in space and time, with very little or no semantic information, making their analysis and extraction of interesting knowledge very complex from the user's point of view. The first work considering semantics in trajectories was developed in 2007 [Alvares et al. 2007], in order to facilitate SQL queries on trajectories and to reduce their cost. The idea was to add to trajectories, geographic places in which the moving object has stayed longer than a user defined minimal amount of time (this work had in 2011 more than 60 citations in *Google scholar*, being very innovative in trajectory queries). In 2008, a spatio-temporal clustering algorithm was proposed to identify low speed regions in trajectories [Palma et al. 2008]. This solution is interesting in applications like traffic management. Later, in 2010, in cooperation with UFPE, a new spatio-temporal clustering algorithm was developed to find interesting places in trajectories considering direction variation as the main measure [Rocha et al. 2010]. This work is useful for applications like fishing activity control. Since the previous works may violate the privacy of the individuals, a semantic-based privacy-preserving data mining method was developed together with KDDLab Pisa. This work was awarded at a workshop held together with the 2010 ACM GIS conference, and an extended version of this work appeared in [Monreale et al. 2011].

The third research topic is being addressed in the context of the projects ATACT and MODAP, with KDD-Lab Pisa. Our first work, together with II/UFRGS and KDDLab/Pisa has been awarded as one of the three best papers at GeoInfo 2010, and a full version appeared in [Alvares et al. 2011]. The set of works developed in spatio-temporal data preprocessing and mining has resulted in the first prototype for trajectory data analysis, named Weka-STPM [Bogorny et al. 2011]. This is the first tool that, extracting data from spatial databases, supports semantic trajectory pattern discovery.

In 2010, a tutorial on trajectory data mining was presented by a member of the group in the second largest data mining conference (IEEE - ICDM), in Sydney - Australia. A workshop on the idea of considering semantics in data mining has happened since 2008 together with the conference IEEE ICDM. This workshop is a collaboration between Chiara Renso (KDDLAB/Pisa) and professor Hui Xiong, of Reuters University - USA. In 2010, a special issue on this research topic has also been organized for the journal KAIS (Knowledge and Information Systems).

The most recent research topics in this context are being developed in a European Project FP7,

between Brazil, Italy and Greece, being coordinated by Chiara Renso. The project named SEEK (Semantic Enriched Knowledge Discovery) has its focus on semantic trajectories and behavior analysis. The works [Alvares et al. 2011] and [Siqueira and Bogorny 2011] specifically focus on two novel types of behaviors: avoidance and chasing.

In cooperation with CEPED (the Brazilian agency in Florianópolis responsible for studies and prevention of natural disasters), we are developing a national database to register all natural disasters occurred in Brazil in the last 20 years.

6. CONCLUSION AND FUTURE ACTIVITIES

The *GBD/UFSC* group is engaged in active research in the database area, mainly related to the management of non-conventional data, like XML and Web data, metadata and semantic annotations, warehouse and spatio-temporal data. This current focus is in consonance with the grand challenges in Computer Science in Brazil, as stated by SBC. As detailed in Sections 2 to 5, our group has contributions for some natures of non-conventional data, in several aspects of data management: *modeling*, *matching*, *retrieval*, and *analysis*. Our future activities include other topics of research, as described below.

While current research in spatio-temporal data analysis focuses on trajectory modeling and mining, addressing the use of semantic information, future research will focus on modeling and mining behavior patterns of moving objects, based on their trajectories. We also intend to contribute, in cooperation with IFC-Camboriu, in the development of a methodology for conceptual and logical modeling of geographic databases that considers information for analytical purposes, i.e., information that helps people in decision-making tasks. For cloud data, in a joint-effort with UFPR, we plan to submit a project of a system architecture that provides relational views of large data stores based on DHT model, which is a common cloud data model. The idea is to consider the mature and well-known relational technology for accessing data in the cloud. Challenges here are related to relational-to-cloud mapping and query performance.

In the context of approximate data matching, we have the following three perspectives. First, the Aglomerante is a recent approved project that, as mentioned early, manipulates heterogeneous data originated from different data sources. Beyond the stated objectives in Section 3, we are working on user feedback, which must be requested over the resulting ranking (implicit by the detection of actions taken with the mouse, for instance, or explicit by answering questionnaires). In this sense, a more elaborated project, called VIDAS (Visualization and Interaction in Data Aggregated by Similarity), has been submitted to the CAPES/COFECUB program (support projects between Brazil and France). The project addresses the investigation of visualization techniques that aid users on interacting with clusters generated from similarity comparison. This project is a collaboration among UFRGS and UFSC, from Brazil, and IRIT, from France. Second, amongst the objectives defined to the project WF-Sim, we are working on defining a query language and operators for searching Web Forms. Furthermore, we intend to focus on data cleaning techniques to be applied over extracted forms from the Web in order to improve their automatic classification in application domains and, as a consequence, to improve the quality of the forms' matching in a same domain. We plan, in a cooperation with the University of Utah, to test such research results on form similarity search in the *DeepPeep*[8] system, a search engine for Web forms. Finally, in the context of InCOD, an Institute that aims at disseminating technology in the applications areas of health, communication and weather, we intend to collaborate in two issues: (i) developing algorithms and methods for feature extraction and indexing of images metadata in order to support decision making; and (ii) building methods for searching interactive content. The first one is related to the health applications and we are planning to define a similarity metric that can properly be applied to metadata previously extracted from medical

[8]http://www.deeppeep.org/

images and then generate a ranking with the lowest number of false positives as possible. The second goal is related to communication, more specifically, Digital TV. In this scenario, we are studying how to develop a similarity search approach for retrieving interactive content from digital TV.

In the area of information retrieval we are investigating data models, query processing techniques, optimization strategies, and access methods for efficiently supporting queries referring, at the same time, to conditions on metadata and content similarity of complex data. This has been done in a postdoc supported by CNPq, in collaboration with Prof. Caetano Traina Junior and Prof. Agma Juci Machado Traina, both from ICMC/USP. This collaboration has the following goals: (i) develop data models based on complex tuples (i.e., tuples with complex data and conventional attributes related to their contents), that include extractors of content descriptors, similarity metrics, and context as first class citizens; (ii) devise ways to optimize queries with conventional and similarity clauses in these models; (iii) develop access methods exploiting lexical indexing, semantic relationships among data (which can be described in ontologies, for example), and similarity metrics among complex tuples; (iv) organize customized methods for extraction of characteristics, indexing, and retrieval of large collections of heterogenous complex tuples.

In spatio-temporal data mining we are focusing on the behavior of moving objects, taking into account properties of trajectories, like speed and acceleration, and context information that is relevant for the application domain. As in the classical data mining field, future research will focus on semantic aspects in data mining, considering not only objective measures for pattern discovery and evaluation, but subjective measures that may help to extract more semantic and understandable patterns.

Finally, it is relevant to note that *GBD/UFSC* also works on general solutions for data management issues, including tools that help traditional Database activities, as well as data management tasks for specific applications. One example is the *brModelo* tool[9], which gives support to relational database design from ER conceptual modeling to SQL/DDL script generation. This tool has been used in Database courses by several Universities in Brazil for teaching of relational database modeling. Another examples are tools that help the application of normal forms over relational tables, and provide the extraction and publication of georeferencing information for applications that use *Google Earth*. This last work was awarded as the best Undergraduate paper in *Regional Database School (ERBD)*, in the 2007 edition [Vasel and Mello 2007].

REFERENCES

Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B., and Vaisman, A. A model for enriching trajectories with semantic geographical information. In *Proceedings of the ACM International Symposium on Geographic Information Systems*. Seattle, USA, pp. 162–169, 2007.

Alvares, L. O., Loy, A. M., Renso, C., and Bogorny, V. An algorithm to identify avoidance behavior in moving object trajectories. *Journal of the Brazilian Computer Society* 11 (3), 2011.

Baeza-Yates, R. A. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

Bilenko, M. and Mooney, R. J. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*. Washington, D.C., pp. 39–48, 2003.

Bogorny, V., Avancini, H., de Paula, B. C., Kuplich, C. R., and Alvares, L. O. Weka-stpm: a software architecture and prototype for semantic trajectory data mining and visualization. *Transactions in GIS* 15 (2): 227–248, 2011.

Bogorny, V., Heuser, C. A., and Alvares, L. O. A conceptual data model for trajectory data mining. In *Proceedings of the International Conference on Geographic Information Science*. Zurich, Switzerland, pp. 1–15, 2010.

Bogorny, V., Kuijpers, B., and Alvares, L. O. Reducing uninteresting spatial association rules in geographic databases using background knowledge: a summary of results. *International Journal of Geographical Information Science* 22 (4): 361–386, 2008.

[9]http://www.sis4.com/brModelo/

Bogorny, V., Kuijpers, B., and Alvares, L. O. St-dmql: a semantic trajectory data mining query language. *International Journal of Geographical Information Science* 23 (10): 1245–1276, 2009.

Bogorny, V., Valiati, J. F., and Alvares, L. O. Semantic-based pruning of redundant and uninteresting frequent geographic patterns. *GeoInformatica* 14 (2): 201–220, 2010.

Challam, V., Gauch, S., and Chandramouli, A. Contextual search using ontology-based user profiles. In *Proceedings of the Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*. Pittsburgh, Pennsylvania, USA, pp. 612–617, 2007.

Chaudhuri, S., Chen, B.-C., Ganti, V., and Kaushik, R. Example-driven design of efficient record matching queries. In *Proceedings of the International Conference on Very Large Data Bases*. Vienna, Austria, pp. 327–338, 2007.

Crestani, F. Application of spreading activation techniques in informationretrieval. *Artificial Intelligence Review* 11 (6): 453–482, 1997.

da Rocha, T. R., Willrich, R., Fileto, R., and Tazi, S. Supporting collaborative learning activities with a digital library and annotations. In *Proceedings of the World Conference on Computers in Education*. Bento Gonçalves, RS, pp. 349–358, 2009.

D'Agostini, C. S. and Fileto, R. Capturing users' preferences and intentions in a semantic search system. In *Proceedings of the International Conference on Software Engineering & Knowledge Engineering*. Boston, Massachusetts, USA, pp. 587–591, 2009.

D'Agostini, C. S., Fileto, R., Dantas, M. A. R., and Gauthier, F. A. O. Contextual semantic search - capturing and using the user's context to direct semantic search. In *Proceedings of the International Conference on Enterprise Information Systems*. Barcelona, Spain, pp. 154–159, 2008.

Deggau, R., Fileto, R., Pereira, D., and Merino, E. Interacting with spatial data warehouses through semantic descriptions. In *Proceedings of the Brazilian Symposium on Geoinformatics*. Campos do Jordao, SP, pp. 122–133, 2010.

Dorneles, C. F., Gonçalves, R., and dos Santos Mello, R. Approximate data instance matching: a survey. *Knowledge and Information Systems* 27 (1): 1–21, 2011.

Dorneles, C. F., Heuser, C. A., Lima, A. E. N., da Silva, A. S., and de Moura, E. S. Measuring similarity between collection of values. In *Proceedings of the ACM International Workshop on Web Information and Data Management*. Washington DC, USA, pp. 56–63, 2004.

Dorneles, C. F., Heuser, C. A., Orengo, V. M., da Silva, A. S., and de Moura, E. S. A strategy for allowing meaningful and comparable scores in approximate matching. In *Proceedings of the ACM Conference on Information and Knowledge Management*. Lisbon, Portugal, pp. 303–312, 2007.

Dorneles, C. F., Nunes, M. F., Heuser, C. A., Moreira, V. P., da Silva, A. S., and de Moura, E. S. A strategy for allowing meaningful and comparable scores in approximate matching. *Information Systems* 34 (8): 740–756, 2009.

E. Malinowski, E. Z. *Advanced data warehouse design: from conventional to spatial and temporal applications*. Springer, 2008.

Fasolin, K., D'Agostini, C. S., Fileto, R., and Besen, R. Praesto - a system for contextual semantic search. In *Anais da Seção de Demos do Simpósio Brasileiro de Banco de Dados*. Fortaleza, CE, 2009.

Fileto, R., Liu, L., Pu, C., Assad, E. D., and Medeiros, C. B. POESIA: An ontological workflow approach for composing web services in agriculture. *VLDB Journal* 12 (4): 352–367, 2003.

Fileto, R., Medeiros, C. B., Pu, C., Liu, L., and Assad, E. D. Building a semantic web system for scientific applications: An engineering approach. In *Proceedings of the International Conference on Web Information Systems Engineering*, A. H. H. Ngu, M. Kitsuregawa, E. J. Neuhold, J.-Y. Chung, and Q. Z. Sheng (Eds.). New York, NY, USA, pp. 633–642, 2005.

Fileto, R., Rigo, W., Junior, V. C. P., Willrich, R., and Oliveira, V. A. Performance evaluation and tunning of spreading activation for associative information retrieval. In *Proceedings of the IADIS International Conference on WWW/Internet (ICWI). Accepted for publication*. Rio de Janeiro, RJ, 2011.

Filho, S. I. V., Fileto, R., Furtado, A. S., and Guembarovski, R. H. Towards intelligent analysis of commplex networks in spatial data warehouses. In *Proceedings of the Brazilian Symposium on Geoinformatics*. Campos do Jordao, SP, pp. 134–145, 2010.

Frozza, A. A. and Mello, R. d. S. A Method for Defining Semantic Similarities between GML Schemas. In C. Davis (Ed.), *Advances In GeoInformatics*. Springer-Verlag, 2007.

Gonçalves, R., D'Agostini, C. S., Silva, F. R., Dorneles, C. F., and Mello, R. d. S. A similarity search method for web forms. In *IADIS International Conference WWW/Internet. Accepted for publication*. Rio de Janeiro, Brazil, 2011.

Gonçalves, R. and Mello, R. d. S. Improving XML instances comparison with preprocessing algorithms. In *Proceedings of the International Conference on Database and Expert Systems Applications*. Regensburg, Germany, pp. 13–22, 2007.

Jr, C. A. S. and Mello, R. d. S. An ontology-driven process for unification of XML instances. In *Anais do Simpósio Brasileiro de Sistemas Multimídia e Web*. Vila Velha, ES, Brazil, 2008.

Lima, C., Schroeder, R., and Mello, R. d. S. Uma ferramenta para conversão de esquemas conceituais eer para esquemas lógicos xml (in portuguese). In *Escola Regional de Banco de Dados*. Florianópolis, SC, Brazil, 2008.

Mangold, C. A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontology* 2 (1): 23–24, 2007.

Mani, A. and Sundaram, H. Modeling user context with applications to media retrieval. *Multimedia Systems* 12 (4): 339–353, 2007.

Manica, E., Dorneles, C. F., and de Matos Galante, R. Supporting temporal queries on XML keyword search engines. *Journal of Information and Data Management* 1 (3): 471–486, 2010.

Medeiros, C. B. Grand research challenges in computer science in brazil. *IEEE Computer* 41 (6): 59–65, 2008.

Mello, R. d. S. and Heuser, C. A. Binxs: A process for integration of XML schemata. In *Proceedings of the International Conference on Advanced Information Systems Engineering*. Porto, Portugal, pp. 151–166, 2005.

Mello, R. d. S., Pinnamaneni, R., and Freire, J. Indexing web form constraints. *Journal of Information and Data Management* 1 (3): 343–358, 2010.

Monreale, A., Trasarti, R., Pedreschi, D., Renso, C., and Bogorny, V. C-safety: a framework for the anonymization of semantic trajectories. *Transactions on Data Privacy* 4 (2): 73–101, 2011.

Moreno, F., Arango, F., and Fileto, R. Extending the map cube operator with multiple spatial aggregate functions and map overlay. In *Proceedings of the International Conference on Geoinformatics*. FairFax, VA, pp. 1–7, 2009a.

Moreno, F., Arango, F., and Fileto, R. A multigranular temporal multidimensional model. In *Proceedings of the Conference on Business Intelligence Systems*. Opatija, Croatia, pp. 206–210, 2009b.

Moreno, F., Fileto, R., and Arango, F. Season queries on a temporal multidimensional model for OLAP. *Mathematical and Computer Modelling* 52 (7-8): 1103–1109, 2010.

Moro, M. M., Braganholo, V. P., Dorneles, C. F., Duarte, D., de Matos Galante, R., and Mello, R. d. S. XML: some papers in a haystack. *SIGMOD Record* 38 (2): 29–34, 2009.

Motro, A. Vague: A user interface to relational datbases that permits vague queries. *Transactions on Information Systems* 6 (3): 187–214, 1988.

Palma, A. T., Bogorny, V., Kuijpers, B., and Alvares, L. O. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the ACM Symposium on Applied Computing*. Fortaleza, Ceara, Brazil, pp. 863–868, 2008.

Rigo, W., Fileto, R., Junior, V. C. P., Oliveira, V. A., Junior, D. I. R., and Silveira, R. A. Web knowledge-based interfaces for resource annotation and repository management. In *Proceedings of the Simpósio Brasileiro de Informática na Educação*. João Pessoa, PB, 2010.

Rigo, W., Fileto, R., Willrich, R., Junior, V. C. P., von Wangenheim, C. G., Oliveira, V. A., and Brasil, L. S. B. Multimedia content annotation in repositories with knowledge-based human-computer interfaces in the web. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. Florianópolis, SC, 2011.

Rocha, J. A. M. R., Times, V. C., Oliveira, G., Alvares, L. O., and Bogorny, V. Db-smot: A direction-based spatio-temporal clustering method. In *Proceedings of the IEEE Conference of Intelligent Systems*. London, UK, pp. 114–119, 2010.

Rodrigues, K. R. and Mello, R. d. S. Diincx: An approach to discovery of implicit integrity constraints from XML data. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*. Las Vegas, Nevada, USA, pp. 606–611, 2007.

Ruthven, I. and Lalmas, M. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review* 18 (2): 95–145, 2003.

Schroeder, R. and Mello, R. d. S. Designing XML documents from conceptual achemas and workload information. *Multimedia Tools and Applications* 43 (3): 303–326, 2009.

Siqueira, F. L. and Bogorny, V. Discovering chasing behavior in moving object trajectories. *Transactions in GIS* 15 (5), 2011.

Vasel, R. and Mello, R. d. S. Um sistema de extração e publicação de informações georreferenciadas em um domínio turístico (in portuguese). In *Escola Regional de Banco de Dados*. Passo Fundo, Brazil, 2007.