GERINDO: Managing and Retrieving Information in Large Document Collections*

Nivio Ziviani[1] Alberto H. F. Laender[1] Edleno S. de Moura[2]
Altigran S. da Silva[2] Carlos A. Heuser[3]
Wagner Meira Jr.[1]

[1] Depatamento de Ciência da Computação
Universidade Federal de Minas Gerais
{nivio,laender,meira}@dcc.ufmg.br

[2] Depatamento de Ciência da Computação
Universidade Federal do Amazonas
{edleno,alti}@dcc.fua.br

[3] Instituto de Informática
Universidade Federal do Rio Grande do Sul
{heuser}@inf.ufrgs.br

Abstract

We present in this report a summary of the main results produced in the five years of the GERINDO research project. The aim of this project is to address the increasing demand for software tools capable of dealing with information available in large document collections, such as the World Wide Web. It involves efforts of researchers from three Brazilian universities to develop core technologies for a number of document management applications demanded by today's information society. These efforts are concentrated in six main research topics: document categorization, semistructured data management, agents and focused crawlers, information retrieval models and searching techniques, efficiency issues, and data mining. Besides specific contributions in these five research topics, the project has stimulated the interaction among the researchers of the three universities who have worked together to solve challenging problems using a combination of different approaches. Moreover, the project has promoted other collaborations with research groups from North America and Europa.

*GERINDO means *managing* in Portuguese and is an acronym for "GErência e Recuperação de INformação em DOcumentos" (Managing and Retrieving Information in Documents).

1 Introduction

This report describes the main results produced in the five years of GERINDO (`http://www.dcc.ufmg.br/gerindo`), a research project that is supported by the Brazilian National Council for Scientific and Technological Development (CNPq/CT-INFO/Grant 55.2087/02-5) and has been carried out by a group of researchers from three Brazilian universities: the Federal University of Minas Gerais (UFMG), the Federal University of Amazonas (UFAM), and the Federal University of Rio Grande do Sul (UFRGS).

The aim of this project is to address the increasing demand for software tools capable of dealing with information available in large document collections, such as the World Wide Web. New advances in computer and communication technologies have driven our society to deal with an amount of electronic information never experienced before. Besides the virtually limitless amount of documents available on the Web, it is common to find nowadays institutions (e.g., companies and governmental departments) where most of the documents produced are stored in electronic media available in their intranets. In this scenario, it is not surprising that there is today an increasing demand for new technologies capable of efficiently managing and retrieving information available in electronic documents.

The GERINDO project involves the efforts of more than 30 researchers (professors and research students) from the three universities to develop core technologies for a number of document management applications demanded by today's information society. These efforts are concentrated on six main research topics: document categorization, semistructured data management, agents and focused crawlers, information retrieval models and searching techniques, efficiency issues, and data mining. Since the project addresses these topics from the same perspective, very often results produced in one specific topic impact the work carried out in another one. Thus, this has stimulated the three groups to cooperate very intensively to develop integrated solutions. In addition, the project has fomented the collaboration with Akwan Information Technologies (`http://www.akwan.com.br`), a Brazilian search technology company acquired by Google Inc. in July 2005, as well as with research groups from several universities from North America and Europe, such as Virginia Tech, the University of Rochester, the University of Utah and the University of Pennsylvania, in the USA, the University of Alberta, in Canada, the University Pompeu Fabra, in Spain, and the Instituto Superior de Tecnologia, in Portugal.

This report is organized as follows. Section 2 presents basic methodological aspects of the project. Section 3 summarizes the main results produced so far in each research topic addressed by the project. Finally, Section 4 presents conclusions and directions for future research.

2 Infrastructure and Methodology

This project adopts as its main methodological strategy the use of a unified repository to store the work produced by students and researchers. The repository uses the Savannah environment (`http://www.dcc.ufmg.br/repositorio`), a GNU Public License (GPL) software that provides facilities for project management, such as version control and concurrent access. The idea of using a centralized software repository is to provide support for reuse of code and easy access to previous research work, make easier the transfer of technology to society, and support collaborative work among researchers of the three universities involved. Using Savannah and communication tools (e.g., voice over ip software and messengers), we have been able to work in a collaborative environment, involving people from different institutions, and to conduct regular remote technical meetings whenever necessary.

In addition, the project researchers have visited each other regularly. These technical visits have provided opportunities for defining new research directions and for conducting collaborative work involving researchers from the three universities. We have also organized workshops to discuss new results, to evaluate partial results, and to plan future research directions.

The project has also played an important role in making reference collections available for its research groups. Reference collections are essential for evaluating new algorithms and information retrieval models, and therefore we have not only acquired a number of such collections but also developed new ones.

3 Research Results

In this section, we sumarize below the main results produced so far in each research topic addressed by the project.

3.1 Document Categorization

The aim of this research topic is to develop algorithms capable of automatically identifying important features of documents and then apply such features to determine whether the documents belong or not to a specific category.

One of the main focus of our research in this topic is the *automatic categorization of web documents* (Calado et al., 2006, 2003; Couto et al., 2006; Cristo et al., 2003; Zhang et al., 2004). We have investigated how link information can be accurate in predicting document categories. Our approach uses a Bayesian network framework (Calado et al., 2004) for proposing and evaluating different alternative models for categorizing web pages. As a result, we have obtained a method that has improved the accuracy from an average micro F1 value of 41 to roughly 87. Further, we have found that the best categorization results can be obtained by using only the title of the web pages, combined with anchor text information and with link information. This means that full-text might be discarded during the categorization process, which significantly reduces the computational efforts to determine the class of each page.

In another work (Couto et al., 2006), we made a study to verify if the same techniques that are used on the Web can be applied to collections of documents containing citations between scientific papers. In this work, we present a comparative study of digital library citations and web links, in the context of automatic text classification. We show that there are in fact differences between citations and links in this context. For the comparison, we run a series of experiments using a digital library of computer science papers and a web directory. In our reference collections, measures based on co-citation tend to perform better for pages in the web directory, with gains up to 37% over text based classifiers, while measures based on bibliographic coupling perform better in a digital library.

Another research direction related to categorization aims to determine a mechanism for detecting and retrieving documents from the Web with a similarity relation to a suspicious document. The categorization problem in this case is to *determine whether a document is a plagiarism or not* (Pereira Jr and Ziviani, 2003, 2004). The algorithm we have developed has important practical applications, such as the verification of the originality of exams and homework at schools and of articles submitted to conferences. It might also help authors find related work when writing a paper. So far, we have proposed and studied several strategies for solving this problem. The most successful one comprises the following steps: (a) generation of a "fingerprint" of the suspicious document, (b) gathering candidate documents from the Web and (c) comparing each candidate document to the suspicious document. In the first step, the fingerprint of the suspicious document is used as its identification. The fingerprint is composed of representative sentences of the document. In the second step, the sentences composing the fingerprint are submitted as queries to a search engine. The documents identified by the URLs returned by the search engine are collected to form a set of similar candidate documents. In the third stage, the candidate documents are compared to the suspicious document. The process of comparing the documents uses two different methods: Shingles and Patricia trees. Preliminary results with these methods can be find in (Pereira Jr and Ziviani, 2003).

Another practical categorization problem we have studied *is how to determine the best advertisement to be shown for each Web page presented to a user in a Web portal.* In this problem, an advertisement company has a set of clients that are willing to pay every time a user clicks on their advertisements. Thus, advertisements should be presented to the users trying to maximize the chance of reaching people interested in their subjects. The final goal is to maximize the gain with the advertisements shown. The task of automatically associating ads to a Web page based on its content is known as content-targeted advertising. It constitutes a key Web monetization strategy nowadays and introduces new challenging technical problems and raises interesting questions. For instance, how to design ranking functions able to satisfy conflicting goals such as selecting advertisements (ads) that are relevant to the users and suitable and profitable to the publishers and advertisers? In (Lacerda et al., 2006) we propose a new method for associating ads with web pages based on Genetic Programming (GP). The GP method aims at learning functions that select the most appropriate ads, given the contents of a web page. These ranking functions are designed to optimize overall precision and minimize the number of misplacements. By using a real ad collection and web pages from a newspaper, it was obtained a gain over a state-of-the-art baseline method of 61.7% in average precision. Further, by evolving individuals to provide good ranking estimations, GP was able to discover ranking functions that are very effective in placing ads in web pages while avoiding irrelevant ones.

3.2 Semistructured Data Management

The main aim of the research in this topic is to develop methods and tools for dealing with data available on the Web and in other non-structured sources (e.g., XML documents), thus providing facilities similar to those available in traditional database systems for managing such data. Specific problems addressed in this topic include or are related to data integration (Borges et al., 2003; Carvalho and da Silva, 2003; de Carvalho et al., 2006), data extraction (Oliveira et al., 2006; Reis et al., 2004; Vieira et al., 2006), query processing (Calado et al., 2004; da Silva et al., 2003; da Silva et al., 2007; Goncalves et al., 2004; Mesquisa et al., 2007; Stasiu et al., 2005) and XML views (Braganholo et al., 2004, 2006).

Data Integration

An interesting data integration problem we have addressed is that of *integrating web data and geographical knowledge into spatial databases* (Borges et al., 2003; Laender et al., 2004). In this work, we see the Web as a rich data source that stores daily facts that often involve textual geographic descriptions. These descriptions can be perceived as indirectly georeferenced data - e.g., addresses, telephone numbers, zip codes and place names. Under this perspective, the Web becomes a large geospatial database, often providing up-to-date local or regional information, and can be used as an important source of urban geographic data for enhancing existing Geographic Information Systems (GIS). We have designed an environment that allows the extraction of geospatial data from web pages, converts them to XML, and uploads the converted data into spatial databases for later use in urban GIS. The effectiveness of our approach has been demonstrated by a real urban GIS application that uses street addresses as the basis for integrating data from different web sources, combining these data with high-resolution imagery (Borges et al., 2003).

We have also worked on the problem of *integrating data from multiple web sources* (Carvalho and da Silva, 2003). We consider web sources with objects that can have different formats and structures, which makes it difficult to identify those that can be matched together. Thus, in this work, we have introduced a model for representing data in this kind of web source, and have studied and proposed strategies for identifying and finding similar identities among objects from distinct sources. In our approach, object matching works like the relational join where a similarity function, based on the vector space model, takes the place of the equality condition. Our approach differs from others in the literature since it can be used to identify and match objects more complexly structured (e.g., XML documents) and not only objects with a flat structure such as relations. The effectiveness of our approach has been demonstrated by means of experiments with real web data sources from different domains, whose results have reached precision levels above 75% (Carvalho and da Silva, 2003).

As a step forward on this research topic, we have also developed a new method for *identifying record replicas in digital libraries and other types of digital repositories* (de Carvalho et al., 2006). This problem is crucial to improve the quality of content and services provided by such systems as well as to yield eventual sharing efforts. Several deduplication strategies are available, but most of them rely on manually chosen settings to combine evidence used to identify records as being replicas. Our method deploys a novel strategy based on Genetic Programming (GP), which is able to automatically generate similarity functions to identify record replicas in a given repository. The generated similarity functions properly combine and weight the best evidence available among the record fields in order to tell when two distinct records represent the same real-world entity. Experimental results have shown that our method outperforms the well-known method by Fellegi and Sunter by more than 12% when identifying replicas in a data set containing researcher's personal data, and by more than 7%, in a data set with article citation data.

Data Extraction

Extracting data from the Web has been a challenging problem over the past years. Although several techniques and tools have been developed to address this problem (Laender et al., 2002), their use is still not spread mostly because of the need for high human intervention and the low quality of the extraction results. Thus, in some practical situations the best solution for extracting data from the Web is to develop domain-oriented methods. In this project, we have proposed a domain oriented approach to *automatically extracting news from web sites*. Our approach is based on a highly efficient tree structure analysis that produces very effective results. We have tested it with several important Brazilian on-line news sites and achieved very precise results, correctly extracting 87.71% of the news in a set of 4088 pages distributed among 35 different sites (Reis et al., 2004).

In many cases, the data items to be extracted for a specific application occur in semi-structured texts without explicit delimiters and embedded within an implicit structure. Examples of such a kind of texts are classified ads, postal addresses, bibliographic references, commercial lists, and curriculum vitae, etc. For dealing with situations such as this, we have developed a novel approach to *extracting data from semi-structured texts* that is based on Hidden Markov Models (HMMs) (Oliveira et al., 2006). Distinctly from previous proposals in the literature that also use HMMs, our approach emphasizes the extraction of metadata in addition to the extraction of data items themselves. Our approach consists of a nested structure of HMMs, in which a main HMM identifies implicit attributes in the text and a set of internal HMMs, one for each attribute, identifies data and metadata. The HMMs are generated by training using a fraction of the set of the texts from which data is to be extracted. Our experiments with classified ads taken from the Web demonstrate that the extraction approach reaches quality levels superior to 0.97, considering the F-measure, even if the text fraction used for training is small.

In many other situations, a distinct but relevant problem is that of *extracting noisy information from web pages*. This is the case of the so called templates, i.e, pieces of HTML code presenting common information that are automatically inserted into web pages of a given site. The widespread use of templates on the Web is considered harmful not only because they compromise the relevance judgment of many information retrieval and mining methods, such as clustering and categorization, but because the negatively impact the performance and resource usage any tool that processes web pages. For dealing with templates, we have developed a new method that efficiently and accurately extracts templates found in web pages collections (Vieira et al., 2006). Our method works in two steps. First, the costly process of template detection is performed over a small set of sample pages. Then, the detected template is extracted from the remaining pages in the collection. This leads to substantial performance gains when compared to previous approaches that combine template detection and extraction. As shown by our experimental evaluation, our approach to identifying and extracting templates is quite effective, achieving F-measure values around 0.9.

Query Processing

Regarding query processing, we have addressed a number of distinct, but related problems. The first one deals with *automatic structuring of keyword-based queries when searching web databases* (Calado et al., 2004; da Silva et al., 2003). We have proposed an approach that allows the use of keywords (as in a Web search engine) for querying databases over the Web. The approach is based on a Bayesian network model and provides a suitable alternative to the use of interfaces based on multiple forms with several fields. Two major steps are involved when querying a Web database using this approach. First, structured (database-like) queries are derived from a query composed only of the keywords specified by the user. Next, the structured queries are submitted to a Web database, and the retrieved results are presented to the user as ranked answers. To demonstrate the feasibility of this approach, a simple prototype Web search system was developed and carefully tested. Experimental results obtained with this system indicate that our approach allows for accurately structuring the user queries and retrieving appropriate answers with minimum intervention from the user (Calado et al., 2004). Moreover, considering that structured or fielded metadata is the basis for many digital library services, including searching and browsing, we have successfully applied this approach to automatically structuring queries for such services (Goncalves et al., 2004).

This approach was further extended to deal with valuable information stored in relational databases. For many purposes, there is an ever increasing demand for having these databases published on the Web, so that users can query the data available in them. An important requirement for this to happen is that query interfaces must be as simple and intuitive as possible. For addressing such a requirement, we have developed LABRADOR, a system for efficiently publishing relational databases on the Web by using a simple text box query interface (Mesquisa et al., 2007). This system operates by taking an unstructured keyword-based query posed by a user and automatically deriving an equivalent SQL query that fits this user's information needs, as expressed by the original query. The SQL query is then sent to a relational DBMS and its results are processed by LABRADOR to create a relevance-based ranking of the answers. Experiments we have carried out show that LABRADOR can automatically find the most suitable SQL query in more than 75% of the cases, and that the overhead introduced by the system in the overall query processing time is almost insignificant. Furthermore, the system operates in a non-intrusive way, since it requires no modifications on the target database schema.

Another query processing problem we have addressed deals with *vagueness when processing queries over XML documents*. The classical approaches for accessing data, query languages and keyword search, cannot be directly applied to applications accessing data whose content the user is unaware of its representation. The same problem also happens in a database in which the instances result from data extracted from the Web or when the user query conditions may have misspelling errors (Dorneles et al., 2004). This generates a scenario where queries having equality operators can led to empty results. A solution would be the use of similarity metrics for comparing data. In this research project, we propose and study methods for querying XML documents that rely on *textual* similarity metrics. Further, as in XML documents we usually handle nested structures – i.e., collections of values – these metrics must also support this kind of structure. Thus, our main results include a similarity search strategy to deal with vagueness and a set of metrics for comparing elements of different types in XML documents.

Also an important problem related to vague or imprecise queries, more specifically to *range-queries*, is that of determining the threshold to be applied when processing such queries in a database. This type of query is relevant in many applications and appears in situations known as *entity resolution* (finding data instances that represent the same real world object) or *approximate join* (joining two data-sets by considering attributes with similar values). Thus, regarding this problem, we have developed a process for the estimation of precision/recall values for several thresholds for a database attribute and a specific similarity function (Stasiu et al., 2005). In this process, the user does not need to label relevant and irrelevant values (as usual in the recall/precision computation process) but has just to inform the number of distinct real world objects that appear in a sample of values. We have performed experiments on a data-set with 10,180 strings containing 387 different city names and five similarity functions. These experiments have shown that recall/precision values estimated from a sample of 160 city names are close (mean square deviation less then 0.07) to recall/precision values obtained when querying the complete data-set.

As a step forward on this problem, in (da Silva et al., 2007) we introduce the concept of *discernibility*, a measure to evaluate similarity functions for range queries. We propose two different approaches, one algorithmic and the other one statistical. As a by-product, our approach also generates an optimum threshold to be used in range queries (here, optimum means that this is the threshold that minimizes false positives and false negatives). Experiments with ten different similarity functions over a set of 150 paper titles have shown that the algorithmic and the statistical approaches are equivalent, i.e., both compute similar values of discernability.

At last, we have addressed the problem of *querying over compressed XML documents*. Although XML has become a de facto standard for data exchange over the Internet, efficiently storing and querying XML data is still an open problem. Thus, deploying new techniques to make it possible to query over compressed XML documents has become a key research issue. Based on a method originally proposed to deal with textual documents, we have developed a very efficient query-aware compressor for XML documents (Lage et al., 2006). As shown by our experimental results, this compressor is able to process XPath queries faster other query-aware compressors presented in the literature, and also achieves better compression ratios and decompressing performance. Moreover, its XPath processor includes a powerful pattern matching engine, based on text compressing techniques proposed by (de Moura et al., 2000), which allows to efficiently deal with query patterns that cannot be handled by other query-aware compressors.

XML Views

Finally, as a very important issue related to semistructured data management, we have addressed the problem of *updating relational databases through XML views* (Braganholo et al., 2004, 2006). Using query trees to capture the notions of selection, projection, nesting, grouping, and heterogeneous sets found in most XML query languages, we have studied how XML views expressed using query trees can be mapped to a set of corresponding relational views. We then have studied how updates on the XML views are mapped to updates on the corresponding relational views. Existing work on updating relational views can then be leveraged to determine whether or not the relational views can be updated with respect to the relational updates, and if so, to translate the updates to the underlying relational database.

3.3 Agents and Focused Crawlers

Another research topic we have addressed within the GERINDO project is the *automatic generation of agents or crawlers for collecting web pages.* As the Web grows, more and more data has become available under dynamic forms of publication, such as legacy databases accessed by an HTML form (the so called hidden Web). In situations such as this, integration of this data relies more and more on the fast generation of agents that can automatically fetch pages for further processing. As a result, there is an increasing need for tools that can help users generate such agents.

As one of the results in this topic, we have created a method for automatically generating agents to collect hidden web pages for data extraction (Lage et al., 2004). This method uses a pre-existing data repository for identifying the contents of these pages and takes the advantage of some usual patterns that are found in many web sites to identify the navigation paths to follow. To demonstrate the effectiveness of our method, we have carried out experiments with sites from different domains. The results of these experiments have shown that our method has been able to succesfully generate a complete agent for 80considered (Lage et al., 2004).

Agents are also required by many web applications (e.g., web directories and digital libraries) to build collections of similar pages they need to accomplish their talks. For some of these applications, the criteria to determine weather a page belongs to a collection are related to the page content. However, there are important situations in which the inner structure of the pages provides a better criteria to guide the crawling process than their content. With this in mind, we have develop a new structure-driven approach for generating web agents that requires a minimum effort from the users (Vidal et al., 2006) to construct them. The idea is to take as input a sample page and an entry point to a web site, and then to generate a structure-driven agent or crawler based on *navigation patterns*, i.e., sequences of patterns for the links that should be followed to reach the pages structurally similar to the sample page. In the experiments we have conducted, structure-driven crawlers generated according to our approach have been able to collect all pages that match the samples given, including those pages added after the crawlers have been generated.

Focused crawlers are also important tools when we are interested in specific document collections. They have as their main goal to crawl web pages that are relevant to a certain topic or user interest, playing an important role for a great variety of applications. In general, they work by trying to find and crawl all kinds of pages deemed as related to an implicitly declared topic. However, users are often not simply interested in any document about a topic, but instead many times they want only documents of a given type or genre on that topic to be retrieved. In this project, we have propose a new *approach to focused crawling that exploits genre and content-related information present in Web pages to guide the crawling process* (Assis et al., 2007). The effectiveness, efficiency and scalability of the proposed approach have been demonstrated by a set of experiments involving the crawling of pages related to syllabi (genre) of computer science courses (subject) and job offers (genre) in the computer science field (subject). The results of these experiments have shown that focused crawlers constructed according to our approach achieve levels of F1 superior to 88%, requiring the analysis of only 60% of the visited pages in order to find almost all relevant pages.

3.4 Information Retrieval Models and Searching Techniques

A major aim of this project is to develop novel techniques for producing a new generation of information retrieval systems. This includes the *development of new information retrieval models, query refinement techniques, and noise removal methods for improving the quality of the results provided by such systems.*

Models are at the core of the information retrieval technology. They determine the accuracy in providing relevant answers to the users, and are also the technological basis of the main component of any information retrieval system, the query processor. Therefore, in this project we have concentrated significant research effort in developing new information retrieval models (Ahnizeret et al., 2004; Pôssas et al., 2005; Pôssas et al., 2005, 2004; Ribeiro-Neto et al., 2001; Vale et al., 2003).

A major result in this topic is a model that combines data mining techniques with traditional information retrieval models. This has resulted a new technique for computing term weights for index terms, which leads to a new ranking mechanism, referred to as the *set-based model* (Pôssas et al., 2005, 2004). The components of this new model are no longer terms, but *termsets*. The novelty is that we compute term weights using a data mining technique, association rules, which is time efficient and yet yields important improvements in retrieval effectiveness. The set-based model function for computing the similarity between a document and a query considers the termset frequency in the document and its scarcity in the document collection. Experimental results show that our model improves the average precision of the answer set for all three collections evaluated. For the TReC-3 collection, which is almost a standard for comparing information retrieval systems, our set-based model led to a gain, relative to the standard vector space model, of 37% in average precision curves and of 57% in average precision for the top 10 documents. Like the vector space model, the set-based model has linear time complexity in the number of documents in the collection (Pôssas et al., 2005, 2004).

We have also worked on the development of models that use taxonomies for categorizing and retrieving information. We have firstly tested this idea in the medical domain (Freitas-Junior et al., 2006; Ribeiro-Neto et al., 2001; Vale et al., 2003), using the International Code of Diseases (ICD) as the taxonomy to categorize and retrieve information available in medical document collections. In this work, the ICD codes are represented as a directed acyclic graph, and supplemented with acronym and synonym dictionaries (Ribeiro-Neto et al., 2001). For each section of each document the acronyms and synonyms are converted to code strings and root node codes are identified. A window of document terms around each root node term is created and the longest path from the graph including these terms is extracted. These codes are assigned to the document in a ranked order by relative path length for that root. As a result, we have a model that allows the development of high quality information retrieval systems and high quality categorization systems that deal with medical documents. Moreover, this model has provided a very effective framework for cross-language information retrieval in the medical domain (Freitas-Junior et al., 2006). Based on these results, we are now working on a generalization of this strategy in order to apply it to other applications, such as categorization of Web news, processing of juridical information (Silveira and Ribeiro-Neto, 2004), and categorization of office and clerical documents found in many company intranets.

Another work related to information retrieval models is the design of models to improve the quality of web intra-site search systems. The idea in this case is to modify the web site design in order to improve the effectiveness of the search system developed for the site modeled. In web site design, a principle accepted by many authors is the separation between information content, navigation structure and visualization. This idea facilitates maintenance tasks since each of those components can be managed separately (Cavalcanti and Robertson, 2003; Vasconcelos and Cavalcanti, 2004). Our proposed approach to web site development is based on these ideas but innovates by modeling information retrieval aspects of the application (Ahnizeret et al., 2004). According to our assumption, by modeling specific aspect of the information content of a web site it is possible to develop search engines that reach a significative improvement in the overall ranking quality. In our experiments, our approach has provided a 48% of improvement in the average precision when compared with traditional implementations of intra-site search engines. Our approach merges an IR-aware methodology and a model-aware intra-site for search engine development (Ahnizeret et al., 2004). We are now working on evolving this idea to automatically discover the structure of a Web site in order to apply our structured information retrieval model using this structure.

Query refinement techniques are an important issue for improving the quality of query results provided by information retrieval systems. In this project, we have proposed a method that automatically generates suggestions of related queries to queries submitted to a search engine (Fonseca et al., 2004, 2003). The method uses information on previously submitted queries extracted from the search engine's log by using algorithms for mining association rules. Experimental results obtained with a commercial search engine indicate that our method generates valid related queries for 90.5% of the top 5 suggestions for common queries extracted from its log. Further, the related queries can also be used as information for a query expansion strategy, resulting in an improvement in the final quality of the answers provided by the search engine.

Following the previous results, we proposed a novel concept-based query expansion technique that allows disambiguating queries submitted to search engines (Fonseca et al., 2005). The concepts are extracted by analyzing and locating cycles in a special type of query relations graph. The concepts related to the current query are then shown to the user who selects the one concept that he interprets is most related to his query. This concept is used to expand the original query and the expanded query is processed instead. Using a web test collection, we show that our approach leads to gains in average precision figures of roughly 32%. Further, if the user also provides information on the *type* of relation between his query and the selected concept, the gains in average precision go up to 52%.

We have also worked on the problem of detecting and removing noisy information from web document collections (Carvalho et al., 2007). One of our results is a method to detect replicated sites. Identifying replicated sites is an important issue for search engines since it can reduce data storage costs, improve query processing time, and remove noise that might affect the quality of the final answers given to the users. Our method uses the structure of the web sites and the content of their pages to identify possible replicas of sites in web document collections. Previous solutions for finding replicated web sites do not take into account the content of the pages, with the claim that the content turns the process unacceptably expensive. We have proposed an alternative method that identifies replica candidate pairs by using page content combined with structural information, while it does not increase the processing times. As we show through experiments, such a combination improves the precision and reduces the overall costs related to the replica detection task. In our experiments, the proposed method achieves a quality improvement of 47.23% when compared to previously proposed approaches.

Another related result is a novel method for removing noisy links from search engine web document collections (da Costa Carvalho et al., 2006). Unlike prior works, we have proposed a method to detect and remove noisy link structures residing at the site level, instead of at the page level. We thus proposed site level versions of previously proposed noise detection techniques, such as site level mutual reinforcement relationships, abnormal support coming from one site towards another, as well as complex link alliances between web sites. Our experiments with a collection crawled from the Brazilian web have shown a substantial increase in the quality of the output rankings after applying our noisy link removal method. Our experiments show an improvement of 26.98% in Mean Reciprocal Rank for popular bookmark queries, 20.92% for randomly selected bookmark queries, and up to 59.16% in Mean Average Precision for topic queries. Furthermore, our algorithms identified up to 16.7% of the links from our collection as noisy, thus demonstrating that searching for noisy links in search engine document collections is more important than searching only for spam.

Finally, as the Web grows at a fast pace and little is known about how new content is generated, another major objective of the GERINDO project is to *study the dynamics of content evolution on the Web*. The work in (Baeza-Yates et al., 2006) addresses this problem aiming at giving answers to questions like: How much new content has evolved from the Web old content? How much of the Web content is biased by ranking algorithms of search engines? We used four snapshots of the Chilean Web containing documents of all the Chilean primary domains, crawled in four distinct periods of time. If a page in a newer snapshot has content of a page in an older snapshot, we say that the source is a parent of the new page. Our hypothesis is that when pages have parents, in a portion of pages there was a query that related the parents and made possible the creation of the new page. Thus, part of the Web content is biased by the ranking function of search engines. We also define a genealogical tree for the Web, where many pages are new and do not have parents and others have one or more parents. We present the Chilean Web genealogical tree and study its components. To the best of our knowledge this is the first paper that studies how old content is used to create new content, relating a search engine ranking algorithm with the creation of new pages.

Following the same direction, in (Pereira-Jr et al., 2006) we study duplicates on the Web, using collections containing documents of all sites under the .cl domain that represent accurate and representative subsets of the Web. We identify duplicate and near-duplicate documents in our collections, studying the distribution of documents in clusters of duplicates. We also study the occurrence of duplicates in both parts of our Web graphs – connected and disconnected component – aiming to identify where duplicates occur more frequently. We originally show that the number of duplicates in the Web is expressively greater than the number of duplicates in the connected component of the Web graph. Works that previously estimated the number of duplicates on the Web used collections of connected components of the Web. In those cases the sample of the Web was biased.

3.5 Efficiency Issues

Information retrieval systems need to be not only highly effective but also extremely efficient, since query throughput is a central problem in these systems.

One of our main efforts in this topic is the development of new *distributed query processing strategies for search engines* (Badue et al., 2007a, 2006, 2007b, 2005a; Badue, 2003; Badue et al., 2001, 2005b). The novelty of our work in this topic is a real distributed architecture implementation that offers concurrent query service. The distributed system we are proposing adopts a network of workstations model and the client-server paradigm. The document collection is indexed with an inverted file.

Web search engines are expensive to maintain, expensive to operate, and hard to design. Predicting their performance is usually done empirically through experimentation, requiring a costly setup. Thus, modeling is of natural appeal in this context. In (Badue et al., 2007a) we introduce a capacity planning model for Web search engines. Our model, which is based on a queueing network, is simple and yet accurate. We discuss how we tune it up and how we use it to predict, for instance, the impact on the query response time when parameters such as disks and collection size are altered. This allows the manager of the search engine to determine a priori whether a new configuration of the system will keep the query response under specified constraints. Our approach is distinct and, we believe, useful to predict the performance of real Web search engines.

The performance of parallel query processing in a cluster of index servers is crucial for modern web search systems. In such a scenario, the response time basically depends on the execution time of the slowest server to generate a partial ranked answer. Previous approaches investigate performance issues in this context using simulation, analytical modeling, experimentation, or a combination of them. Nevertheless, these approaches simply assume balanced execution times among homogeneous servers (by uniformly distributing the document collection among them, for instance)—a scenario that we did not observe in our experimentation in (Badue et al., 2007b). On the contrary, we found that even with a balanced distribution of the document collection among index servers, correlations between the frequency of a term in the query log and the size of its corresponding inverted list lead to imbalances in query execution times at these same servers, because these correlations affect disk caching behavior. Further, the relative sizes of the main memory at each server (with regard to disk space usage) and the number of servers participating in the parallel query processing also affect imbalance of local query execution times. These are relevant findings that have not been reported before and that, we understand, are of interest to the research community.

Another research direction in this topic is the development of new *pruning methods for search engines* (de Moura et al., 2005). One way to address query processing efficiency without losing effectiveness is to reduce the amount of data to be processed at query time. We adopt a new pruning strategy capable of greatly reducing the size of search engine indices. Experiments show that our technique can reduce the indices storage costs by up to 60% over traditional lossless compression methods, while keeping the loss in retrieval precision to a minimum.

We have also been working on the use of *data compression algorithms to reduce the size of text and indexes* (Ziviani, 2004, 2007; Ziviani and Moura, 2003) and, at the same time, to improve the efficiency of IR systems. Particularly, we are now addressing the problem of compressing XML documents. Altough XML has become a de facto standard for data exchange over the Internet, efficiently storing and querying XML data is still an open problem. Thus, several recent efforts have been made to deploy techniques to directly query over compressed XML data. We have worked on the development of a system that efficiently compress XML documents that allows querying over the compressed document without requiring decompression. Preliminary experimental results indicate that this system is faster than other systems presented in the literature. It also achieves better compression ratios and decompressing performance, and deals efficiently with query patterns that cannot be used in other systems (Lage, 2004).

As a final work in this topic, we are developing *new hashing methods for static sets of keys* (Botelho, 2004; Botelho et al., 2005, 2007). This problem is strongly related to the generation of indexes for IR systems, since a significant portion of the time spent to generate IR indexes is spent in hash operations. If the set of keys is *static*, then it is possible to compute a function $h(x)$ to find any key in the table in one probe. This function is called a *perfect hash function*. A perfect hash function that stores a set of records in a table of the size equal to the number of keys times the key size is called a *minimal perfect hash function*. A minimal perfect hash function totally avoids the common problem of wasted space and time. Minimal perfect hash functions are used for memory efficient and fast retrieval of items from static sets, such as words in natural languages, reserved words in programming languages or interactive systems, universal resource locations in web search engines, or itemsets in data mining techniques. Therefore, there are applications for minimal perfect hash functions in information retrieval systems, database systems, hypertext, hypermedia, language translation systems, electronic commerce systems, compilers, and operating systems.

In (Botelho et al., 2005) we propose a novel algorithm based on random graphs to construct minimal perfect hash functions h. For a set of n keys, our algorithm outputs h in expected time $O(n)$. The evaluation of $h(x)$ requires two memory accesses for any key x and the description of h takes up $1.15n$ words. This improves the space requirement to 55% of a previous minimal perfect hashing scheme due to Czech, Havas and Majewski. A simple heuristic further reduces the space requirement to $0.93n$ words, at the expense of a slightly worse constant in the time complexity. Large scale experimental results are presented. For a collection of 100 million keys, each key 63 bytes long on average, our algorithm finds a minimal perfect hash function in 811 seconds on average.

Perfect hash functions can potentially be used to compress data in connection with a variety of database management tasks. Though there has been considerable work on how to construct good perfect hash functions, there is a gap between theory and practice among all previous methods on minimal perfect hashing. On one side, there are good theoretical results without experimentally proven practicality for large key sets. On the other side, there are the theoretically analyzed time and space usage algorithms that assume that truly random hash functions are available for free, which is an unrealistic assumption. In (Botelho et al., 2007) we attempt to bridge this gap between theory and practice, using a number of techniques from the literature to obtain a novel scheme that is theoretically well-understood and at the same time achieves an order-of-magnitude increase in performance compared to previous "practical" methods. This improvement comes from a combination of a novel, theoretically optimal perfect hashing scheme that greatly simplifies previous methods, and the fact that our algorithm is designed to make good use of the memory hierarchy. We demonstrate the scalability of our algorithm by considering a set of over one billion URLs from the World Wide Web of average length 64, for which we construct a minimal perfect hash function on a commodity PC in a little more than 1 hour. Our scheme produces minimal perfect hash functions using slightly more than 3 bits per key. For perfect hash functions in the range $\{0, \ldots, 2n-1\}$ the space usage drops to just over 2 bits per key (i.e., one bit more than optimal for representing the key). This is significantly below of what has been achieved previously for very large values of n.

3.6 Data Mining

Mining distributed document databases (Otey et al., 2004; Veloso et al., 2003) is emerging as a fundamental computational problem. A common approach for mining distributed databases is to move all of the data from each database to a central site where a single model is built. This approach is accurate, but too expensive in terms of processing time. For this reason, several approaches have been developed to efficiently mine distributed databases, but they still ignore a key issue – privacy. Privacy is the right of individuals or organizations to keep their own information secret. Privacy concerns can prevent data movement – data may be distributed among several custodians, none of which is allowed to transfer its data to another site.

In this research topic we have proposed an efficient approach to mining frequent itemsets in distributed databases. Our approach is accurate and uses a privacy-preserving communication mechanism. The proposed approach is also efficient in terms of message passing overhead, requiring only one round of communication during the mining operation. Our privacy-preserving distributed approach has superior performance when compared to the application of a well-known mining algorithm in distributed databases (Otey et al., 2004; Veloso et al., 2003).

Another data mining research effort pursued in the project is towards a novel and efficient classification technique named “Lazy Associative Classification” Veloso et al. (2006a); Veloso and Meira Jr. (2005a,b, 2006); Veloso et al. (2006b), which showed to be more efficient in terms of computational costs, since it focus on the sample to be classified. Further, by focusing on the sample-relevant information we are able to improve the precision of the classifier, in particular for datasets where the popularity of classes vary significantly. The technique has been applied to various types of data, from proteomics to spam detection, going through digital libraries. In all cases our technique performed better than state-of-the-art approaches, such as SVM both in terms of computational costs and classification precision.

4 Conclusions and Future Directions

Developing core technologies for managing and processing information on electronic documents has been the focus of the GERINDO project. Several algorithms and techniques proposed represent the state-of-the-art in document management and information retrieval solutions. This has called the attention of the international research community to our group and gives to Brazil an excellent opportunity to be seen in the future as a leading country in software development in this area.

Besides these important research contributions, which can be assessed by the quality of the publications produced by our group, the project has also made other significant achievements. First, it has provided a stimulating environment for collaboration in six distinct research topics which produced solutions for a number of problems using a combination of different approaches. Second, the project has been an important source of new and challeging problems which served as research topics for several MSc and PhD students. Third, its results have been applied to practical problems and helped to improve existing tools and applications such as search engines, digital libraries, and geographical information systems. Finally, the project has opened a number of opportunities for collaboration with other research groups, being worth stressing the connections of the group with the Brazilian company Akwan Information Technology, which is now part of Google Brasil. Therefore, we expect the project will keep its course of action and produce even stronger results in the future.

References

Ahnizeret, K., Cavalcanti, J. M. B., Oliveira, D., de Moura, E. S., and da Silva, A. S. (2004). Information retrieval aware web site modelling and generation. In *Proceedings of 23rd International Conference on Conceptual Modeling*, pages 402–419, Shangai.

Assis, G. T., Laender, A. H. F., Gonçalves, M. A., and da Silva, A. S. (2007). A genre-aware approach to focussed crawling. Submitted.

Badue, C., Almeida, J., Almeida, V., Baeza-Yates, R., Ribeiro-Neto, B., Ziviani, A., and Ziviani, N. (2007a). A capacity planning model for web search engines. Submitted.

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

Badue, C., Baeza-Yates, R., Ribeiro-Neto, B., Ziviani, A., and Ziviani, N. (2007b). Analyzing imbalance among homogeneous index servers in a web search system. *Information Processing and Management*. To appear.

Badue, C., Barbosa, R., Golgher, P., Ribeiro-Neto, B., and Ziviani, N. (2005a). Trade-offs in the processing of web queries. ACM SIGIR Workshop on Heterogeneous and Distributed Information Retrieval.

Badue, C. S. (2003). Distributed query processing using partitioned inverted files. Thesis proposal presented to the Graduate Course in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Badue, C. S., Baeza-Yates, R. A., Ribeiro-Neto, B. A., and Ziviani, N. (2001). Distributed query processing using partitioned inverted files. In *Proceedings of the 11th International Symposium on String Processing and Information Retrieval*, pages 10–20.

Badue, C. S., Barbosa, R. A., Golgher, P. B., Ribeiro-Neto, B. A., and Ziviani, N. (2005b). Basic issues on the processing of web queries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, pages 577–578.

Baeza-Yates, R., Pereira-Jr, A., and Ziviani, N. (2006). The evolution of web content and search engines. In *Proceedings of the 8th ACM Workshop on Web Mining and Web Usage Analysis*, pages 68–73, Philadelphia, PA, USA.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., da Silva, A. S., and Davis, C. B. (2003). The web as a data source for spatial databases. In *Anais do V Brazilian Symposium on Geoinformatics*, Campos do Jordão, Brasil.

Botelho, F. C. (2004). Estudo comparativo do uso de hashing perfeito mínimo. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte. Supervisor Nivio Ziviani.

Botelho, F. C., Kohayakawa, Y., and Ziviani, N. (2005). A practical minimal perfect hashing method. In *Proceedings of the Workshop on Experimental Algorithms*, volume 3505 of *Lecture Notes in Computer Sciency*, pages 488–500. Springer Verlag.

Botelho, F. C., Pagh, R., and Ziviani, N. (2007). Perfect hashing for relational data compression. Submitted.

Braganholo, V. P., Davidson, S., and Heuser, C. A. (2004). From XML view updates to relational view updates: Old solutions to a new problem. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pages 276–287, Toronto, Canada.

Braganholo, V. P., Davidson, S. B., and Heuser, C. A. (2006). Pataxó: A framework to allow updates through xml views. *ACM Trans. on Database Syst.*, 31(3):839–886.

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

Calado, P., Cristo, M., Moura, E., Ziviani, N., Ribeiro-Neto, B., and Gonçalves, M. (2003). Combining link-based and content-based methods for web document classification. In *Proceedings of the Twelveth ACM International Conference on Information and Knowledge Management*, pages 394–401, New Orleans, Louisiana, USA.

Calado, P., da Silva, A. S., Vieira, R. C., Laender, A. H. F., and Ribeiro-Neto, B. (2004). A bayesian network approach to searching databases through keyword-based queries. *Information Processing and Management*, 40(5):773–790.

Carvalho, A., de Moura, E. S., da Silva, A. S., Berlt, K., and Bezerra, A. (2007). A cost-effective method for detecting web site replicas on search engine databases. *Data and Knowledge Engineering*. To appear.

Carvalho, J. C. P. and da Silva, A. S. (2003). Finding similar identities among objects from multiple web sources. In *Proceedings of the Fifth International Workshop on Web Information and Data Management*, pages 90–93, New Orleans, Louisiana, USA.

Cavalcanti, J. M. B. and Robertson, D. (2003). Web site synthesis based on computational logic. *Knowledge and Information Systems*, 5(3):263–287.

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

Cristo, M. A. P., Calado, P. P., Moura, E., Ziviani, N., and Ribeiro-Neto, B. (2003). Link information as a similarity measure in web classification. In *Proceedings of the 10th Symposium On String Processing and Information Retrieval*, Lecture Notes in Computer Sciency, pages 66–71, Manaus, Brazil. Springer Verlag.

da Costa Carvalho, A. L., Chirita, P.-A., de Moura, E. S., Calado, P., and Nejdl, W. (2006). Site level noise removal for search engines. In *Proceedings of the 15th International World Wide Web Conference*, pages 73–82.

da Silva, A. S., Calado, P., Vieira, R. C., Laender, A. H. F., , and Ribeiro-Neto, B. (2003). *Effective Databases for Text and Document Management*, chapter Keyword-Based Queries over Web Databases, pages 74–92. Idea Group, Inc., Hershey, USA.

da Silva, R., Stasiu, R., Orengo, V. M., and Heuser, C. A. (2007). Measuring quality of similarity functions in approximate data matching. *Journal of Informetrics*. (To appear).

de Carvalho, M. G., Gonçalves, M. A., Laender, A. H. F., and da Silva, A. S. (2006). Learning to deduplicate. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 41–50.

de Moura, E. S., dos Santos, C. F., Fernandes, D. R., da Silva, A. S., Calado, P., and Nascimento, M. A. (2005). Improving web search efficiency via a locality based static pruning method. In *Proceedings of the 14th International World Wide Web Conference*, pages 235–244, Chiba, Japan.

de Moura, E. S., Navarro, G., Ziviani, N., and Baeza-Yates, R. A. (2000). Fast and flexible word searching on compressed text. *ACM Trans. on Inf. Syst.*, 18(2):113–139.

Dorneles, C. F., Heuser, C. A., Lima, A. E. N., da Silva, A. S., and de Moura, E. S. (2004). Measuring similarity between collection of values. In *Proceedings of the 6th International Workshop on Web Information and Data Management*, pages 56–63, Washington, DC, USA.

Fonseca, B., Golgher, P., Moura, E. S., Pôssas, B., and Ziviani, N. (2004). Discovering search engine related queries using association rules. *Journal of Web Engineering*, 4(2):215–227.

Fonseca, B., Golgher, P., Moura, E. S., and Ziviani, N. (2003). Using association rules to discover related queries on search engines. In *Proceedings of the First Latin American Web Conference*, pages 66–71, Santiago, Chile.

Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., and Ziviani, N. (2005). Concept-based interactive query expansion. In *Proceedings of the 14th ACM International Conference on Information and knowledge Management*, pages 696–703, New York, NY, USA. ACM Press.

Freitas-Junior, H. R., Ribeiro-Neto, B. A., de Freitas Vale, R., Laender, A. H. F., and de Lima, L. R. S. (2006). Categorization-driven cross-language retrieval of medical information. *Journal of the American Society for Information Science and Tecnology*, 57(4):501–510.

Goncalves, M. A., Fox, E. A., Krowne, A., Calado, P., Laender, A. H. F., da Silva, A. S., and Ribeiro-Neto, B. (2004). The effectiveness of automatically structured queries in digital libraries. In *Proceedings of the 4th IEEE/ACM Joint Conference on Digital Libraries*, pages 98–107, Tucson, Arizona, USA.

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

Laender, A. H. F., Gonçalves, M. A., and Roberto, P. (2004). Bdbcomp: Building a digital library for the brazilian computer science community. In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries*, pages 23–24, Tucson, Arizona, USA.

Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. (2002). Brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93.

Lage, J. P. (2004). Consulta a documentos XML comprimidos. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte. Supervisor Alberto H. F. Laender.

Lage, J. P., da Silva, A. S., Golgher, P. B., and Laender, A. H. F. (2004). Automatic generation of agents for collecting hidden web pages for data extraction. *Data and Knowledge Engineering*, 49(2):177–196.

Lage, J. P., Laender, A. H. F., and de Moura, E. S. (2006). YAQCX: A word-based query-aware compressor for XML data. In *Proceedings of the 21st Brazilian Symposium on Databases*, pages 251–264, Florianópolis, Brazil.

Mesquisa, F., da Silva, A. S., de Moura, E. S., Calado, P., and Laender, A. H. F. (2007). Labrador: Efficiently publishing relational databases on the web by using keyword-based query interfaces. *Information Processing & Management*. To appear.

Oliveira, R., Mesquita, F., da Silva, A. S., and Cortez, E. (2006). Extração de dados e metadados em textos semi-estruturados usando HMMs. In *Anais do XXI Simpósio Brasileiro de Bancos de Dados*, pages 89–94.

Otey, M., Veloso, A., Wang, C., Parthasarathy, S., and Meira Jr, W. (2004). Parallel and distributed methods for incremental frequent itemset mining. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 34(6):2439–2450.

Pereira-Jr, A. R., Baeza-Yates, R., and Ziviani, N. (2006). Where and how duplicates occur in the web. In *Proceedings of the Fourth Latin American Web Congress*, pages 127–134, Cholula, Mexico.

Pereira Jr, A. R. and Ziviani, N. (2003). Syntactic similarity of web documents. In *Proceedings of the First Latin American Web Congress*, pages 194–200, Santiago, Chile.

Pereira Jr, A. R. and Ziviani, N. (2004). Retrieving similar documents from the web. *Journal of Web Engineering*, 4(2):247–261.

Pôssas, B., Ziviani, N., Ribeiro-Neto, B. A., and Meira Jr., W. (2005). Maximal termsets as a query structuring mechanism. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 287–288.

Pôssas, B., Ziviani, N., Wagner Meira, J., and Ribeiro-Neto, B. (2005). Set-based vector model: An efficient approach for correlation-based ranking. *ACM Trans. on Inf. Syst.*, 23(4):397–429.

Pôssas, B., Ziviani, N., Ribeiro-Neto, B., and Meira, W. (2004). Processing conjunctive and phrase queries with the set-based model. In *Proceedings of the 11th International Symposium on String Processing and Information Retrieval*, volume 3246 of *Lecture Notes in Computer Sciency*, pages 171–182, Padova, Itália. Springer Verlag.

Reis, D. C., Golgher, P. B., da Silva, A. S., and Laender, A. H. F. (2004). Automatic web news extraction using tree edit distance. In *Proceedings of the 13th International World Wide Web Conference*, pages 502–511, New York, NY, USA.

Ribeiro-Neto, B. A., Laender, A. H. F., and de Lima, L. R. S. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Tecnology*, 52(5):391–401.

Silveira, M. L. and Ribeiro-Neto, B. (2004). Concept-based ranking: A case study in the juridical domain. *Information Processing and Management*, 40(5):791–805.

Stasiu, R. K., Heuser, C. A., and da Silva, R. (2005). Estimating recall and precision for vague queries in databases. In *Proceedings of the 17th International Conference on Advanced Information Systems Engineering*, volume 3520 of *Lecture Notes in Computer Science*, pages 187–200. Springer.

Vale, R. F., Ribeiro-Neto, B., Lima, L. R. S., Laender, A. H. F., and Freitas Jr., H. R. (2003). Improving text retrieval in medical collections through automatic categorization. In *Proceedings of the 10th International Symposium on String Processing and Information Retrieval*, pages 197–210, Manaus, Brazil.

Vasconcelos, W. and Cavalcanti, J. M. B. (2004). An agent-based approach to web site maintenance. In *Proceedings of the International Conference on Web Engineering*, volume 3140 of *Lecture Notes in Computer Sciency*, pages 271–286, Munique. Springer Verlag.

Veloso, A., Cristo, M., Meira Jr., W., Goncalves, M., and Zaki, M. (2006a). Multi-evidence, multi-criteria, lazy associative classification. In *Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management*, pages 218–227, Arlington, VA, EUA.

Veloso, A. and Meira Jr., W. (2005a). Eager, lazy and hybrid algorithms for multi-criteria associative classification. In *Proceedings of the Data Mining Algorithms Workshop*, pages 17–25, Uberlândia,MG.

Veloso, A. and Meira Jr., W. (2005b). Rule generation and rule selection techniques for cost-sensitive associative classification. In *Proceedings of the 20th Brazilian Symposium on Databases*, pages 295–309, Uberlândia,MG.

Veloso, A. and Meira Jr., W. (2006). Lazy associative classification for content-based spam detection. In *Proceedings of the Fourth Latin American Web Congress*, Cholula, Mexico.

Veloso, A., Meira Jr., W., and Zaki, M. (2006b). Lazy associative classification. In *Proceedings of the 2006 IEEE International Conference on Data Mining*, Singapore.

Veloso, A. A., Meira Jr, W., Parthasarathy, S., and Carvalho, M. B. (2003). Efficient, accurate and privacy-preserving data mining for frequent itemsets in distributed databases. In *Proceedings of the 18th Brazilian Symposium on Databases*, pages 281–292, Manaus, Brazil.

Vidal, M. L. A., da Silva, A. S., de Moura, E. S., and Cavalcanti, J. M. B. (2006). Structure-driven crawler generation by example. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 292–299.

Vieira, K. M., da Silva, A. S., de Moura, E. S., Cavalcanti, J. M. B., Pinto, N., and Freire, J. (2006). A fast and robust method for template detection and removal. In *Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management*, pages 258–267.

Zhang, B., Goncalves, M. A., Fan, W., Chen, Y., Fox, E., Calado, P., and Cristo, M. (2004). Intelligent fusion of structural and citation-based evidence for text classification. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pages 162–163, Washington DC.

Ziviani, N. (2004). *Projeto de Algoritmos com Implementações em Pascal e C*. Thomson, second edition.

Ziviani, N. (2007). *Projeto de Algoritmos com Implementações em Java e C++*. Thomson.

Ziviani, N. and Moura, E. S. (2003). *Advances in Computers: Information Repositories*, volume 57, chapter Adding Compression to Next-Generation Text Retrieval Systems, pages 171–204. Academic Press.