Gerindo: New Technologies for Managing and Processing Information in Documents

Nivio Ziviani[1] Edleno S. de Moura[2] Alberto H. F. Laender[1]
Altigran S. da Silva[2] Berthier A. Ribeiro-Neto[1]
Carlos A. Heuser[3] João M. B. Cavalcanti[2] Mara Abel[3]
Renato A. Ferreira[1] Wagner Meira Jr.[1]

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

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

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

Processo CNPq - 55.2087/2002-5

Abstract

We present in this report a summary of the main results produced in the first two years of the Gerindo research project. The aim of this project is to address the increasing demand for software 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 five main research topics: document categorization, semistructured data management, information retrieval models, efficiency issues in information retrieval, 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. As a consequence, we expect the project will produce even stronger results in the next two years.

1 Introduction

This report describes the main results produced in the first two years of the Gerindo research project (`http://www.dcc.ufmg.br/gerindo`), a project for managing and retrieving information available in electronic documents. The project 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 the Federal University of Minas Gerais (UFMG), 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 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, governmental departments) where most of the documents produced are stored in electronic media available in their intranets. In this scenario, it is not a surprise 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 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 on five main research topics: document categorization, semistructured data management, information retrieval models, efficiency issues in information retrieval, 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. This has stimulated the three groups to cooperate very intensively to develop integrated solutions.

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: document classification, semistructured data management, information retrieval models, efficiency issues in information retrieval, and data mining. Finally, Section 4 presents conclusions and future directions.

2 Infrastructure and Methodology

This project adopts as its main methodologycal 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 uni-

versities 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 aquired a number of such collections but also developed new ones.

3 Research Results

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., 2004a, 2003; 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., 2004b) for proposing and evaluating different alternative models for classifying 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.

Following a different research direction, we are currently studying methods for *determining the geographical scope of Web pages* (Moraes, 2004). We propose a geographical classification procedure for Web pages that takes advantage of different sources of information to enable the classification of individual Web pages, instead of whole sites. Since Web pages usually contain very little data, we propose the use of text from other

previously classified Web pages as a complementary source of information. Again, we show that classification accuracy can be increased by combining this information with link-based measures, but now using links to estimate how related a page is to a location. Preliminary experiments indicate that, as in the case of general Web page categorization, link measures can significantly improve the accuracy of the final categorization. This work provides, therefore, an effective approach to the problem of classifying Web pages according to their geographical scope, by making use of sources of information commonly available on the Web, such as links and text.

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 classification problem in this case is to *determine whether a document is a plagiarism or not* (Pereira Jr and Ziviani, 2003, 2004a,b). The algorithm we have developed has important practical applications, such as the verification of the originality of exams and homeworks 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 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 indicates we are getting close to a very efficient solution to this problem (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. To solve this problem, we are currently studying approaches based on k-nearest neighbors strategy and a model that represents the problem as a Markovian system. This work is in a very preliminary stage, but the first experimental and theoretical results obtained indicate that we will probably produce a very competitive solution, are very encouraging when compared to available commercial products.

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, 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 Silva, 2003; Laender et al., 2004), data extraction (Lage et al., 2004; Reis et al., 2004), query processing (Camillo et al., 2003; Goncalves et al., 2004; Silva et al., 2003) and XML views Braganholo et al. (2004).

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. We have focused on using the Web as an important source of urban geographic information and propose to enhance urban Geographic Information Systems (GIS) using indirectly georeferenced data extracted from it. We have designed an environment that allows the extraction of geospatial data from Web pages, converts it to XML format, 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 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. In this work, we have studied and proposed strategies for identifying and finding similar identities among objects from such Web sources. In our approach, the object identification works like the relational join operation where a similarity function takes the place of the equality condition. This similarity function is based on information retrieval techniques. Our approach differs from others in the literature since it can be used to identify objects more complexly structured (e.g., XML documents) and not only objects with a flat structure such as relations. The effectiveness of our approach was verified through experiments, where the results with real Web data sources from different domains reached precision levels above 75% (Carvalho and Silva, 2003).

Extracting data from the Web has been a challeging 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 work, we propose the application of 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).

Another interesting work related to data extraction we have carried out is the *development of agents for data extraction* (Lage et al., 2004). 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. In our work, we have created a method for automatically generating agents to collect hidden Web pages. This method uses a pre-existing data repository for identifying the contents of these pages and takes the advantage of some patterns that can be found among Web sites to identify the navigation paths to follow. To demonstrate the accuracy of our method, we have carried out experiments with sites from different domains (Lage et al., 2004).

Regarding query processing, we have addressed two distinct problems. The first problem deals with *automatic structuring of keyword-based queries when searching Web databases* (Calado et al., 2004b; 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., 2004b). Moreover, considering that structured or filded 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).

The second query processing problem 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. This can happen in a database in which the instances result from a Web data extraction and when the user query conditions can have misspelling errors (Dorneles et al., 2004). This problem 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 work, we are proposing and studying methods for querying XML documents that use *textual* similarity metrics. Further, as in XML we handle a nested structure - i.e, collections of values – it would

be useful to have similarity metrics that support them. So, for this purpose we are also working on some *aggregated* metrics for the nested structure. The results we have obtained in this work are an similarity search stretegy to deal with vagueness and a set of metrics for comparing elements of different types in XML documents.

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). Using query trees to capture the notions of selection, projection, nesting, grouping, and heterogeneous sets found throughout 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 view 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 Information Retrieval Models

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

Our first work in this topic is the development of a model that combines techniques from the data mining with traditional information retrieval models. As a result, we have a new technique for computing term weights for index terms, which leads to a new ranking mechanism, referred to as the *set-based model*. The components of our 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 IR 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., 2004).

Another work related to IR 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 IR systems developed for the site modeled. In Web site design, a principle accepted by many authors is the separation between in-

formation content, navigation structure and visualization. This idea promotes a better understanding of the data requirements (content), the underlying architecture of the site (navigation) and an appropriate user interface (visualization). Furthermore it makes maintenance tasks easier as each of those components can be managed separately (Cavalcanti and Robertson, 2003; Vasconcelos and Cavalcanti, 2004). Recent technologies such as XML, XSL and stylesheets also promote separation between content and visualization, encouraging and facilitating the development of methods for Web site construction based on those concepts. Our proposed approach to Web site development is based on these ideas but innovates by modeling IR aspects of the application. According to our assumption, by modeling specific IR attributes 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 search engine development (Ahnizeret et al., 2004). We are now working on evolving this idea to automatically discover the structure of a Web site and apply our structured IR model using this structure.

We have also worked on the development of models that use taxonomies for classifying and retrieving information. We have firstly tested this idea in the medical domain (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. 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. We are now working on a generalization of this strategy in order to apply it to other applications, such as classification of Web news, processing of juridical information (Silveira and Ribeiro-Neto, 2004), and classification of office and clerical documents found in many company intranets.

A final work related to IR models aims to improve the quality of query results provided by search engines by using previous queries submitted to such system to determine the relation among them and then make use of this information to improve the results for new queries. In this work, we have proposed and studied a method to automatically generate suggestions of related queries submitted to Web search engines. The method extracts information from the log of past queries submitted to search engines using algorithms for mining association rules. Experimental results performed with a commercial searching engine indicate that we can obtain with our model correct suggestions in 90.5% of the top 5 suggestions presented for common queries extracted from a real log. Further, the related queries can also be used as information for a query expansion model, resulting

in an improvement in the final quality of the answers provided by the systems (Fonseca et al., 2004, 2003).

3.4 Efficiency Issues in IR

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, 2003). 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. We adopt two distinct strategies of index partitioning in the distributed system, namely local index partitioning and global index partitioning. In both strategies, documents are ranked using the vector space model along with a document filtering technique for fast ranking. We have evaluated and compared the impact of these two index partitioning strategies on query processing performance. Experimental results on retrieval efficiency show that, within our framework, the global index partitioning outperforms the local index partitioning.

Another research direction in this topic is the development of new *pruning methods for search engines* (Fernandes, 2004). One way to address query processing efficiency without losing effectiveness is to reduce the amount of data to be processed at query time. In this work we are tackling the issue of compressing the search engine's textual database from an atypical perspective. We are using text summarization as a compression tool. The novelty of this work arises from the fact that standard text summarization techniques stem from the domain of natural language processing, which in turn pay a premium for maintaining the summarized text readable. Our concern on the other hand is to maintain the summarized text retrievable. That is, we aim at imposing a much lower overhead at query processing time (and also resource consumption) while still keeping the loss in retrieval effectiveness at a minimum. Extensive experiments with this new proposed method using a 22 Gb database from an actual search engine has shown that it is possible to reduce the size of the database to approximately 20% of its original size with minimum loss in the quality of the query results.

We have also been working on the use of *data compression algorithms to reduce the size of text and indexes* (Ziviani, 2004; 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). 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.

3.5 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 time required. For this reason, several approaches were 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).

4 Conclusions and Future Directions

Developing core technologies for managing and processing information on electronic documents was the focus of the Gerindo project in its first two years. Many 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 the group, the project has also made other significant achievements. First, it has provided a stimulating environment for collaboration in five 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 seach engines, digital libraries, and geographical information systems. Therefore, we expect the project will keep its course of action and produce even stronger results in the next two years.

References

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

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.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., 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.

Botelho, F. C. (2004). Aplicação de hashing perfeito mínimo para representar o vocabulário de um índice invertido. Dissertação de mestrado a ser defendida em outubro de 2004, Orientador Nivio Ziviani.

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 Database*, Toronto, Canada. To appear.

Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B., and Ziviani, N. (2004a). Linkage similarity measures for the classification of web documents. *Submitted to the Journal of the American Society for Information Science and Technology*.

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, USA.

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

Camillo, S. D., Mello, R. S., and Heuser, C. A. (2003). Querying heterogeneous xml sources through a conceptual schema. In *Proceedings of the 22nd International Conference on Conceptual Modeling*, volume 2813 of *Lecture Notes in Computer Sciency*, pages 186–199.

Carvalho, J. C. P. and 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.

Coelho, T., Calado, P., Souza, L., Ribeiro-Neto, B., and Muntz, R. (2004). Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417.

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*, pages 43–55, Manaus, Brazil.

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

Fernandes, D. R. (2004). Extração de informação irrelevante de máquinas de busca. Master's thesis, Universidade Federal do Amazonas, Manaus. Orientador Edleno Silva de Moura.

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*. To appear.

Fonseca, B., Golgher, P., Moura, E. S., and Ziviani, N. (2003). Using association rules to discover related queries on search engines. In *Latin American Web Conference (LA-WEB)*, pages 66–71, Santiago, Chile. IEEE Computer Society.

Goncalves, M. A., Fox, E. A., Krowne, A., Calado, P., Laender, A. H. F., 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.

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. Orientador Alberto H. F. Laender.

Lage, J. P., 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.

Moraes, M. S. (2004). Localização geográfica de páginas web. Dissertação de mestrado em andamento, orientador Edleno Silva de Moura, DCC/UFAM.

Otey, M., Veloso, A., Wang, C., Parthasarathy, S., and Meira Jr, W. (2004). Parallel and distributed methods for incremental frequent itemset mining. *Systems, Man and Cybernetics, Part B*. To appear.

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

Pereira Jr, A. R. and Ziviani, N. (2004a). Geração de impressão digital para recuperação de documentos similares na web. In *Anais do II Workshop de Tecnologia da Informação e Linguística, XXIV Congresso da Sociedade Brasileira de Computação*, pages 1569–1578, Salvador, Bahia.

Pereira Jr, A. R. and Ziviani, N. (2004b). Retrieving similar documents from the web. *Submitted to the Journal of Web Engineering*.

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*, Padova, Itália.

Reis, D. C., Golgher, P. B., 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.

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.

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.

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.

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, Amazonas, Brazil.

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*, Washington DC. Poster. To appear.

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

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.