UPGRADE is the European Journal for the Informatics Professional, published bimonthly at <http://www.upgrade-cepis.org/>

Publisher
UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by Novática <http://www.ati.es/novatica/>, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática,* <http://www.ati.es/>)

UPGRADE monographs are also published in Spanish (full version printed; summary, abstracts and some articles online) by Novática

UPGRADE was created in October 2000 by CEPIS and was first published by Novática and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>)

UPGRADE is the anchor point for UPENET (UPGRADE European NETwork), the network of CEPIS member societies' publications, that currently includes the following ones:
• Informatik-Spektrum, journal published by Springer Verlag on behalf of the CEPIS societies GI, Germany, and SI, Switzerland
• ITNOW, magazine published by Oxford University Press on behalf of the British CEPIS society BCS
• Mondo Digitale, digital journal from the Italian CEPIS society AICA
• Novática, journal from the Spanish CEPIS society ATI
• OCG Journal, journal from the Austrian CEPIS society OCG
• Pliroforiki, journal from the Cyprus CEPIS society CCS
• Pro Dialog, journal from the Polish CEPIS society PTI-PIPS

ISSN 1684-5285

Monograph of next issue (April 2007)
"Information Technologies for Visually Impaired People"
(The full schedule of **UP**GRADE is available at our website)

CEPIS UPGRADE
The European Journal for the Informatics Professional
http://www.upgrade-cepis.org
Vol. VIII, issue No. 1, February 2007

Monograph: Next Generation Web Search
(published jointly with Novática*)
Guest Editors: *Ricardo Baeza-Yates, José-María Gómez-Hidalgo, and Paolo Boldi*

2 Presentacion. The Future of Web Search — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*

5 Efficient Sparse Linear System Solution of the PageRank Problem — *Gianna M. Del Corso, Antonio Gullì, and Francesco Romani*

12 Learning to Analyze Natural Language Texts — *Giuseppe Attardi*

19 SNAKET: A Personalized Search-result Clustering Engine — *Paolo Ferragina and Antonio Gullì*

27 The Multimodal Nature of the Web: New Trends in Information Access — *Luis-Alfonso Ureña-López, Manuel-Carlos Díaz-Galiano, Arturo Montejo-Raez, and Mª Teresa Martín-Valdivia*

33 Adversarial Information Retrieval in the Web — *Ricardo Baeza-Yates, Paolo Boldi, and José-María Gómez-Hidalgo*

41 GERINDO: Managing and Retrieving Information in Large Document Collections — *Nivio Ziviani, Alberto H. F. Laender, Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, and Wagner Meira Jr.*

49 Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web — *Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras*

57 Yahoo! Research Barcelona: Web Retrieval and Mining — *The Yahoo! Research Team*

UPENET (UPGRADE European NETwork)

59 From **Novática** (ATI, Spain)
Informatics Profession
The Maturity of IT Professionalism in Europe — *Sean Brady*

68 From **Pro Dialog** (PTI-PIPS, Poland)
Graphical Interfaces
Portable Declarative Format for Specifying Graphical User Interfaces — *Zbigniew Fryźlewicz and Rafał Gierusz*

75 From **Novática** (ATI, Spain)
Next-generation Web
Blogs: On the Cutting Edge of the Next-generation Web — *Antonio Miguel Fumero-Reverón and Fernando Sáez-Vacas*

CEPIS NEWS

83 Harmonise Project: Building up to the Final Report—*François-Philippe Draguet*
84 News & Events: European Funded Projects and News Updates

* This monograph will be also published in Spanish (full version printed; summary, abstracts, and some articles online) by **Novática**, journal of the Spanish CEPIS society ATI (*Asociación de Técnicos de Informática*) at <http://www.ati.es/novatica/>.

GERINDO: Managing and Retrieving Information in Large Document Collections

Nivio Ziviani, Alberto H. F. Laender, Edleno Silva de Moura, Altigran Soares da Silva, Carlos A. Heuser, and Wagner Meira Jr.

We present in this article a summary of some of the main results produced in the five years of the GERINDO research project. This project aims to address the increasing demand for software tools capable of dealing with information available in large document collections, such as the World Wide Web, and involves the participation of several researches from three Brazilian universities. The project efforts have been focused on a number of research topics on web information retrieval and management, such as information retrieval models, searching techniques, document categorization, semi structured data management, generation of agents for document collection, and efficiency issues. In addition to its specific research contributions, the project has stimulated the interaction among the researchers of the three universities and has promoted other collaborations with research groups from North America and Europe.

Authors

Nivio Ziviani has a PhD in Computer Science from the University of Waterloo, Canada, 1982. He is a Professor of the Department of Computer Science of the Federal University of Minas Gerais (UFMG), Brazil, where he coordinates the Laboratory for Treating Information (LATIN). He is a Professor Emeritus in Computer Science at UFMG. He is a co-founder of Miner Technology Group, sold to Folha de São Paulo/UOL group in 1999, and Akwan Information Technologies, sold to Google Inc. in 2005. He has co-authored over 100 refereed papers and 4 books in the areas of algorithm design and information retrieval, the latter his primary area of research. He was General Co-Chair of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval and co-founder of the International Conference on String Processing and Information Retrieval (SPIRE). <nivio@dcc.ufmg.br>.

Alberto H. F. Laender holds a PhD degree in Computing from the University of East Anglia, UK, 1984. He joined the Computer Science Department of the Federal University of Minas Gerais in 1975, where he is currently a Full Professor and the head of the Database Research Group. He is also a member of the Brazilian National Research Council Computer Science Advisory Committee and of the Advisory Board of the ACM Special Interest Group on Management of Data. Prof. Laender has served as a program committee co-chair, as well as a program committee member, for several national and international conferences on databases and Web-related topics. He is the author of more than 100 refereed journal and conference papers, and was one of the co-founder of Akwan Information Technologies, a Brazilian search technology company acquired by Google Inc. in July 2005. Prof. Laender´s research interests include conceptual modeling and database design, web data management, web information systems, and digital libraries. <laender@dcc.ufmg.br>.

Edleno Silva de Moura is an Associate Professor of Computer Science at the Federal University of Amazonas (UFAM) in Brazil, where he heads the Information Technology Research Group (GTI). He received a PhD in Computer Science from the Federal University of Minas Gerais (UFMG), Brazil, in 1999, where his research activities were focused on applications of data compression for information retrieval systems. After finishing his PhD, he worked as an associate researcher at UFMG and as Chief Technology Officer for Akwan Information Technologies, a company specialized in developing information retrieval systems for the Web. He is the author of several papers in journals and conference proceedings covering topics in the areas of information retrieval, text indexing, text searching, text compression, and related areas. <edleno@dcc.fua.br>.

Altigran Soares da Silva received a Ph.D. (2002) in Computer Science from the Federal University of Minas Gerais (UFMG), Brazil. Currently, he is an Associate Professor of Computer Science at the Federal University of Amazonas (UFAM) and participates as an associate researcher of the UFMG Database Group. He has been working on a number of research projects founded by Brazilian agencies such as CNPq and FINEP. He has served as a program committee member for conferences on databases and web technology worldwide and also as an external reviewer for international Computer Science Journals. In 2007, he is the Program Committee chair of 22th Brazilian Symposium on Databases. His main research interests include extraction and management of web semistructured data, web information retrieval and digital libraries. Currently, he is serving as the Director of Marketing and Communications of the Brazilian Computer Society. <alti@dcc.fua.br>.

Carlos A Heuser received his Dr. in Computer Science from the University of Bonn, Germany, in 1986. He also holds an Electrical Engineer title from the Universidade Federal do Rio Grande do Sul in Brazil (1973) and a M.Sc. in Computer Science from the same University (1976). He is a Full Professor of the Instituto de Informática at the Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil. He has published a book on database design (in Portuguese) and more than 90 conference and journal papers in the areas of databases and software engineering. He has chaired latin american conferences and workshops, and has actively served in the program committees of international conferences and has been reviewer for journal papers. He is a member of the Brazilian Computer Society (SBC). <heuser@inf.ufrgs.br>.

Wagner Meira Jr. obtained his PhD from the University of Rochester in 1997 and is currently Associate Professor at the Computer Science Department at Universidade Federal de Minas Gerais, Brazil. His research focuses on scalability and efficiency of large scale parallel and distributed systems, from massively parallel to Internet-based platforms, and on data mining algorithms, their parallelization, and application to areas such as information retrieval, bioinformatics, and e-governance. <meira@dcc.ufmg.br>.

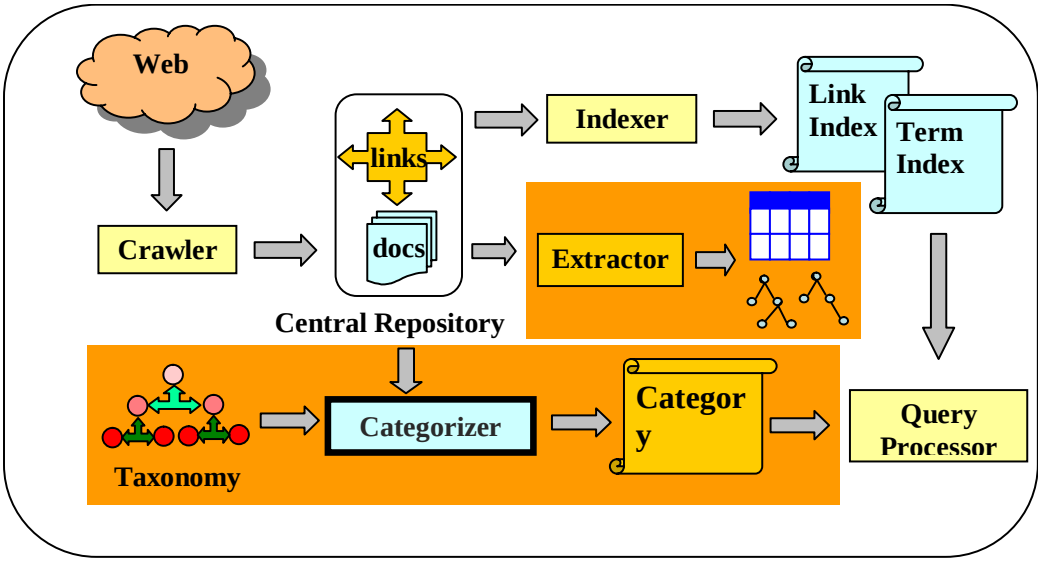

Figure 1: Architecture of an Advanced Environment for Web Information Retrieval and Management.

Keywords: GERINDO, Information, Information Retrieval, Management, Search Engines, World Wide Web.

1 Introduction

GERINDO[1] is a research project, funded by the Brazilian National Council for Scientific and Technological Development (CNPq/CT-INFO grant number 55.2087/02-5), that aims to address the increasing demand for software tools capable of dealing with information available in large document collections, such as the World Wide Web. In addition to the virtually limitless amount of documents available on the Web, it is now common to find institutions (e.g., companies and governmental departments) where most of the documents produced are stored in electronic media available in their intranets. It is, therefore, not surprising that there is an increasing demand for new technologies capable of efficiently managing and retrieving information available in electronic documents.

The project[2] involves the participation of more than 30 researchers (professors and research students) 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 project efforts have been focused on a number of research topics on web information retrieval and management, such as information retrieval models, searching techniques, document categorization, semistructured data management, generation of agents for web page collection, and efficiency issues. The groups from these three universities cooperate very intensively to develop integrated solutions. The project has fomented the collaboration with research groups from several universities from North America and Europe, including the Virginia Technology Institute and State University, 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. The project also established a strong and successful collaboration with Akwan Information Technologies[3], a Brazilian search technology company acquired by Google Inc. in July 2005.

This article reports on the main research results that have been produced by the project over the last five years and is organized as follows. Section 2 briefly describes the infrastructure and methodological aspects regarding the coordination of the project's activities. Section 3 summarizes the main results produced so far in some of the main research topic addressed by the project. Finally, Section 4 presents conclusions and directions for future research.

2 Infrastructure and Methodology

The project adopts as its main methodological strategy the use of a unified repository to store the work produced by the research groups of each university. The repository uses the Savannah environment[4], a GNU Public License (GPL) software that provides facilities for project manage-

[1] GERINDO means *managing* in Portuguese and is an acronym for "Gerência e Recuperação de INformação em DOcumentos" (Managing and Retrieving Information in Documents).
[2] <http://www.dcc.ufmg.br/gerindo>.
[3] <http://www.akwan.com.br>.
[4] <http://www.dcc.ufmg.br/repositorio>.

ment, 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 regular 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

Figure 1 describes the architecture of an advanced environment for web information retrieval and management, which we will use as a reference for discussing some of our main research results. As we can see, this architecture includes a number of modules, for instance the **Crawler**, the **Indexer** and the **Query Processor**, that are also part of any search engine available on the Web. Briefly, the **Crawler** is a software agent that navigates through the Web and collects all documents it is able to reach, storing them in a repository for further processing. The **Indexer** processes the collected documents and builds indexes that are then used by the **Query Processor** to retrieve those documents that are more relevant to the users' queries. Traditionally, such indexes are built based only on the content of the documents, i.e., considering only the list of terms (words) found in the documents. However, new generation search engines, such as Google[5], also build other indexes that consider additional information on the Web topology, i.e., information on how the web documents are linked to each other.

Two additional modules in this environment enhance the functionality of a traditional search engine. The **Categorizer** improves the effectiveness of the retrieved information by automatically categorizing the collected documents according to some predefined taxonomy. The **Extractor** provides facilities for extracting data from the documents, storing them according to some specific format (e.g., relational tables or XML), in order to allow for further processing such as mining, database-like querying, application-oriented publishing, and integration to other data applications (e.g., digital libraries, e-brokers, etc.).

In the following, we summarize the main results produced so far in some of research topics addressed by the GERINDO project. A more complete and detailed description of these results can be found in [25].

3.1 Information Retrieval Models and Searching Techniques

A major aim of the GERINDO project is to develop novel techniques for constructing 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 (see Figure 1). Therefore, in this project we have spent significant effort in developing new information retrieval models [17] [19] [21] .

A major result in this topic is a model that combines data mining techniques with traditional information retrieval models. This has originated a new technique for computing term weights for index terms, which leads to a new ranking mechanism, referred to as the *set-based model* [17]. 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[6], 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 [2], of 37% in average precision 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 [17].

We have also worked on the development of models that use taxonomies for categorizing and retrieving information. As shown in Figure 1, taxonomy-based models are key to improve the effectiveness of the information retrieved. We have firstly tested this idea in the medical domain, using the *International Code of Diseases* (ICD) as the taxonomy to categorize and retrieve information available in medical document collections [19]. 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 con-

[5] <http://www.google.com>.
[6] <http://trec.nist.gov/data.html>.

verted 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 [9]. 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 [21], and categorization of office and clerical documents found in many company intranets.

Query refinement techniques [2] are also 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 [8]. 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.

We have also proposed a novel method for removing noisy links from the collection of web documents indexed by a search engine [6]. Unlike prior works on this topic, our method detects and removes noisy link structures residing at the site level, instead of at the page level. Thus, we have proposed site level versions for existing noise detection algorithms. With this slight change in the noise detection strategy, we have obtained an improvement in the ranking quality and a significant reduction in the number of links between pages in the experiments we have conducted with a collection crawled from the Brazilian Web. Our algorithms have identified as noisy up to 16.7% of the links from our test collection, thus demonstrating that searching for noisy links in search engine document collections is more important than searching only for spam.

3.2 Document Categorization

In addition to the categorization model proposed in [19], we have investigated other issues on automatic categorization of web documents [5] [7]. Thus, 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 or not the documents belong to a specific category.

In [5], we have investigated how link information can be accurate in predicting document categories. Tests performed in a web directory show that link information alone allows the classifying of documents with an average precision of 86%. Further, when combined with a traditional text-based classifier precision increases to values up to 90%, representing gains that range from 63 to 132% over the use of text-based classification alone. Also, 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 category of each page.

We have also investigated whether techniques that are used on the Web can be applied to collections of documents containing citations between scientific papers [7]. In this work we have conducted a comparative study of digital library citations and web links in the context of automatic text classification. We have shown that there are in fact differences between citations and links in this context. For the comparison, we have 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.

A practical, but very important categorization problem we have addressed 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. In [11] 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 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.

3.3 Semistructured Data Management

Traditional information retrieval systems provide only two ways of handling web documents: querying and browsing. However, the Web is a vast repository of data that might be useful for a huge number of applications [13]. Thus, the main aim of the research in this topic is to develop methods and tools for dealing with data available on the Web, and

also in other non-structured data sources (e.g., XML documents), in order to provide facilities similar to those available in traditional database systems for managing such data, as suggested in Figure 1. Specific problems addressed in this topic include data extraction [12] [18] [23], query processing [4] [10] [15] [20], and XML views [3].

3.3.1 Data Extraction

Extracting data from the Web has been a challenging problem over the past years and several techniques and tools have been developed to address it [13]. In the GERINDO project, we have proposed an approach to extracting data from web pages, called DEByE (Data Extraction By Example), that is based on a small set of examples specified by the user [12]. A major difference of our approach when compared with other ones is the fact that the user specifies examples according to a structure of his liking and that this structure is described at example specification time. For the specification of the examples, the user interacts with a tool we developed which adopts nested tables as its visual paradigm (see Figure 2). Nested tables are simple, intuitive, and allow the user to be shielded from technical details (such as HTML tags, formatting operators, and learning automata) related to the extraction problem. The examples provided by the user indicate the structure and the textual surroundings of the data to be extracted, and are then used to generate patterns that allow extracting data from new pages.

Despite the large number of general purpose techniques and tools for extracting data from the Web, in some practical situations the best solution is to develop domain-oriented methods. Thus, in this project we have also proposed a domain oriented approach to automatically extracting news from web sites [18], which 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. This approach has been partially implemented as part of a commercial tool for automatic news clipping.

In some 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 they negatively impact the performance and resource usage of 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 [23]. Our method works in two steps. Firstly, 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. An experimental evaluation has shown that our approach is quite effective, being able to correctly detect and extract around 90% of the information available on templates of real web pages from a representative sample we have used for tests.

3.3.2 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 [4]. 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. Firstly, 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 presented to the user as ranked answers. A simple prototype web search system has been developed to demonstrate the feasibility of this approach. 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. 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 [10].

This approach has been further extended to deal with valuable information stored in relational databases. There has been 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 [15].

As illustrated in Figure 2, this system operates by taking an unstructured keyword-based query posed by a user and generating a set of ranked *candidate structured queries* that fits this user's information needs, as expressed by the original query. Then, an SQL query is derived for one of the structured queries (the system can automatically pick up this query or the user can choose it from the top ranked ones) and sent to a relational DBMS for execution, being its results 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

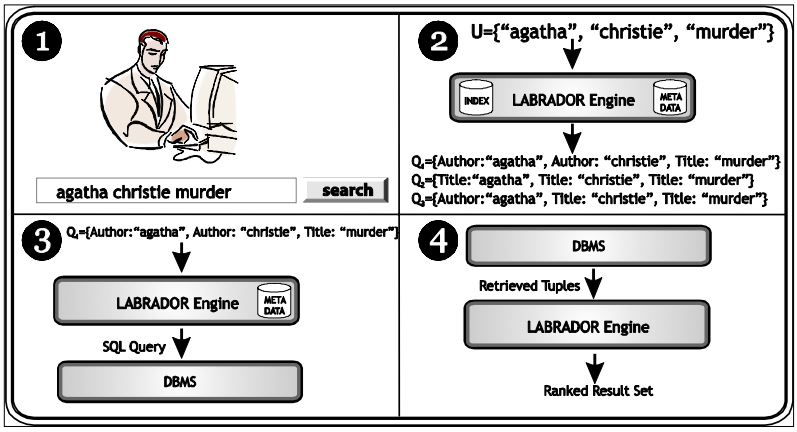

Figure 2: Query Processing in LABRADOR.

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. This problem appears in applications accessing data whose representation the user is unaware of. Some typical scenarios are a database that stores data extracted from the Web (see Figure 1) or when the issued queries include conditions that may contain misspelling errors. In such scenarios, queries having equality operators can lead to empty results. A solution is the use of similarity functions for comparing data. In our research we have tackled several problems related to the application of similarity functions. One such a problem is the determination of the threshold to be applied when processing *range-queries* in a database. We have developed a semi-automatic process for the estimation of precision and recall values for several thresholds for a database attribute and a specific similarity function. Another problem is that of evaluating the quality of similarity functions.

In [20] we introduce the concept of discernability, a measure to evaluate similarity functions for range queries. We propose two different methods for estimating this measure, one algorithmic and the other one statistical. Experiments empirically corroborate the effectiveness of both methods and show that they produce similar results.

3.3.3 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 [3]. 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. Then, we 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.

3.4 Agents for Web Page Collection

Another research topic addressed in the GERINDO project is the automatic generation of agents (or crawlers) for collecting web pages. However, we have focused our work on this topic upon very specific types of agent. For instance, 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 such situations, the integration of this data relies on the fast generation of agents that can automatically fetch these pages for further processing. As a result, there is an increasing need for tools that can help users generate these agents.

Thus, we have created a method for automatically generating agents to collect hidden web pages for data extraction [14]. 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 show that our method has been able to successfully generate a complete agent for 80% percent of the sites considered.

Agents are also used in many web applications (e.g., web directories and digital libraries) to build collections of similar pages required to accomplish their tasks. Usually, the criteria to determine whether 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 better criteria than their content to guide the crawling process. With this in mind, we have developed a new structure-driven approach for generating web agents that requires a minimum effort from the users to construct them [22]. 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 based on *navigation patterns*, i.e., sequences of link patterns that should be followed to reach pages that are 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 given samples, including those pages added after the agents have been generated.

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. Thus, the development of efficient query processing strategies as well as of methods that reduce the amount of data handled at query time is key to improve the performance of the query processor (see Figure 1).

A major effort in this research topic has been the development of new distributed query processing strategies for 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 investigated performance issues in this context using simulation, analytical modeling, experimentation, or a combination of them. However, 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 [1]. 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 [16]. 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 worked on the use of data compression algorithms to reduce the size of text and indexes [24]. The technique combines several data compression features to provide economical storage, faster indexing and accelerated searches. Compressing both the index and the complete text cuts the total space in half the size of the non-compressed text. The time required to build the index and answer a query is far less than if the index and text had not been compressed. This illustrates a rare case where there is no space-time trade-off.

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 groups and gives to Brazil an excellent opportunity to be seen in the near future as a leading country in software development in this area.

Besides these important research contributions, the GERINDO project has also made other significant achievements. First, it has provided a stimulating environment for collaboration in distinct research topics that have produced solutions for a number of problems using a combination of different approaches. Second, the project has been an important source of new and challenging problems that have served as research topics for several MSc and PhD students. Third, the project 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, especially the interaction with the Brazilian company Akwan Information Technology, which is now part of Google Brazil. Therefore, we expect the project will keep its course of action in order to consolidate the undergoing work. Future work will include additional topics such as focused crawling, integration of data from heterogeneous web sources, web data mining, information retrieval in sensor network environments, generation of minimal perfect hash functions for improving index structures, and detection of duplicate and near-duplicate web pages.

References

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

[2] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press/Addison-Wesley, New York, 1999. ISBN: 020139829X. <http://sunsite.dcc.uchile. cl/irbook/>.

[3] V. P. Braganholo, S. B. Davidson, and C. A. Heuser. PATAXÓ: A framework to allow updates through XML views. ACM Transactions on Database Systems, 31(3):839—886, 2006.

[4] P. Calado, A. S. da Silva, R. C. Vieira, A. H. F. Laender, and B. Ribeiro-Neto. A Bayesian network approach to searching Web databases through keyword-based queries. Information Processing & Management,

40(5):773-790, 2004.

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

[6] A. L. da Costa Carvalho, P.-A. Chirita, E. S. de Moura, P. Calado, and W. Nejdl. Site Level Noise Removal for Search Engines. In Proceedings of the 15th International World Wide Web Conference. Edinburgh, Scotland, 2006, pp. 73-82.

[7] T. Couto, M. Cristo, M. A. Gonçalves, P. Calado, N. Ziviani, E. S. de Moura, and B. Ribeiro-Neto. A Comparative Study of Citations and Links in Document Classification. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, Chapel Hill, North Carolina, USA, 2006, pp. 75-84.

[8] B. Fonseca, P. B. Golgher, E. S. de Moura, B. Pôssas, and N. Ziviani. Discovering Search Engine Related Queries Using Association Rules. Journal of Web Engineering, 4(2):215-227, 2004.

[9] H. R. Freitas-Junior, B. Ribeiro-Neto, R. F. Vale, A. H. F. Laender, and Luciano R. S. de Lima. Categorization-driven cross-language retrieval of medical information. Journal of the American Society for Information Science and Technology, 57(4}:501—510, 2006.

[10] M. A. Gonçalves, E. A. Fox, A. Krowne, P. Calado, A. H. F. Laender, A. S. da Silva, and B. Ribeiro-Neto. The Effectiveness of Automatically Structured Queries in Digital Libraries. In Proceedings of the 4th IEEE/ACM Joint Conference on Digital Libraries, Tucson, Arizona, USA, 2004, pp. 98-107.

[11] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, Learning to Advertise. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006, pp. 549-556.

[12] A. H. F. Laender, B. A. Ribeiro-Neto, and A. S. da Silva. DEByE - Data Extraction By Example. Data and Knowledge Engineering, 40(2):121-154, 2002.

[13] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A Brief Survey of Web Data Extraction Tools. SIGMOD Record, 31(2):84-93, 2002.

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

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

[16] E. S. de Moura, C. F. dos Santos, D. R. Fernandes, A. S. da Silva, P. Calado, and M. A. Nascimento. Improving Web Search Efficiency via a Locality Based Static Pruning Method. In Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, 2005, pp. 235-244.

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

[18] D. C. Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender. Automatic Web News Extraction Using Tree Edit Distance. In Proceedings of the 13th International World Wide Web Conference, New York, NY, USA, 2004, pp. 502-511.

[19] B. A. Ribeiro-Neto, A. H. F. Laender, and L. R. S. de Lima. An experimental study in automatically categorizing medical documents. Journal of the American Society for Information Science and Tecnology, 52(5):391-40, 2001.

[20] R. da Silva, Raquel Stasiu, V. M. Orengo, and C. A. Heuser. Measuring quality of similarity functions in approximate data matching. Journal of Informetrics, 1(1): 35-46, 2007.

[21] M. L. Silveira and B. Ribeiro-Neto. Concept-based ranking: a case study in the juridical domain. Information Processing & Management, 40(5):791-805, 2004.

[22] M. L. A. Vidal, A. S. da Silva, E. S. de Moura, and J. M. B. Cavalcanti. Structure-driven Crawler Generation by Example. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006, pp. 292-299.

[23] K. M. Vieira, A. S. da Silva, E. S. de Moura, J. M. B. Cavalcanti, N. Pinto, and J. Freire. A Fast and Robust Method for Template Detection and Removal. In Proceedings of the ACM Fifteenth Conference on Information and Knowledge Management, Arlington, Virginia, USA, 2006, pp. 258-267.

[24] N. Ziviani, E. Silva de Moura, G. Navarro, R. Baeza-Yates. Compression: A Key for Next-Generation Text Retrieval Systems. IEEE Computer, 33(11):37-44, 2000.

[25] N. Ziviani, A. H. F. Laender, E. Silva de Moura, A. S. da Silva, C. A. Heuser, and W. Meira Jr. GERINDO: Managing and Retrieving Information in Large Document Collections. Technical Report 01/2007, Departamento de Ciência da Computação, UFMG, Belo Horizonte, 2007.

