The Database and Information Retrieval Research Group at UFAM

João M. B. Cavalcanti, Marco Cristo, David Fernandes, Edleno S. de Moura and Altigran S. da Silva

Instituto de Computação
Universidade Federal do Amazonas, Brazil
{john,marco,david,edleno,alti}@dcc.ufam.edu.br

Abstract. We present in this article a summary of the main research efforts and results produced by the Database and Information Retrieval group of the Universidade Federal do Amazonas in the last years. The main focus of the research carried out by the group is on the areas of Data Management, Information Retrieval and Data Mining applied to the context of the World-Wide Web. The ultimate goal is developing cutting-edge technologies for a number of important Web applications. The complementary expertise of the researchers in the group allows the development of interdisciplinary work related to these three areas and gives the group a singular profile and a solid basis for providing high impact solutions to practical problems in several research topics.

Categories and Subject Descriptors: H. Information Systems [**H.2 DATABASE MANAGEMENT**]: H.2.4 Systems; H. Information Systems [**H.2 DATABASE MANAGEMENT**]: H.2.8 Database Applications; H. Information Systems [**H.3. INFORMATION STORAGE AND RETRIEVAL**]: H.3.0 General

Keywords: BDRI/UFAM group, Information Retrieval, Data Management, Data Mining

1. INTRODUCTION

The Database and Information Retrieval Group (BDRI) from the Universidade Federal do Amazonas (UFAM) has made strong contributions to the areas of Data Management, Information Retrieval and Data Mining applied to the context of the World-Wide Web in the past years. This is evidenced not only by the research articles published in high quality venues, but also by the software prototypes built in our labs and our cooperations with companies through which we convert research results into economical and social development.

The group is composed of researchers with a strong academic background on these areas, and focuses its research efforts on providing cutting-edge technologies for a number of Web applications. The complementary expertise of the researchers has led the development of interdisciplinary work related to these three areas and assigns the group a singular profile and a solid basis for presenting high impact solutions to practical problems in the several research topics addressed within the group.

The development of World-Wide Web applications, along with all the fast pace growing statistics related to them (number of users, pages, servers, on-line services, etc.), raises several interesting and high economical impact research problems. Among them, our efforts in the BDRI group are concentrated in Web Data Management and Web Information Retrieval problems.

The term *Web Data Management (WDM)* has been used in the recent literature to refer to the area that studies problems related to fetching, extracting, querying, modeling, storing, transforming, and integrating semistructured data available on the Web. These issues have been growing in importance in the scientific community in the last years, as it can be seen by the considerable space devoted to research on this topic in important publication venues related to disciplines such as Databases, Information Retrieval, and Artificial Intelligence.

The efforts in the *Web Information Retrieval* area include research topics such as: information retrieval models for Web search, efficiency issues in information retrieval, noise removal, recommendation and advertising systems, and multimedia information retrieval.

Importantly, the two areas are intrinsically mixed in our research work, with both Information Retrieval techniques being used as a tool to solve Web Data Management problems and also Web Data Management techniques being used to improve Information Retrieval systems. This symbiosis is one of the main competitive advantages of our research group when compared to other groups in which these complementary expertises are not present. As a result, the group was able to produce high level academic research results in the past years. In addition, we have also worked in alternatives to transfer the knowledge produced by the group to the industry, which is done through projects in cooperation with established companies and also through the creation of companies to explore our research results in practical applications.

This article is organized as follows. Section 2 presents general statistics and data about the group. Section 3 presents the main research topics addressed by us in the past years, including also current work and future directions. Section 4 presents an overview about the relationship between the group and industry. Section 5 presents final remarks.

2. GROUP PROFILE

This section presents the researchers, the main research areas and the projects of the Database and Information Retrieval Research group at UFAM.

2.1 Researchers

The BDRI group has five researchers, three of them are recipients of research and technology scholarships from the Brazilian Research Council (CNPq). They all share common specific areas of interest allowing a strong cooperation within the group in particular co-supervision of M.Sc. and Ph.D. students. It is a common practice in the group to have two or more researchers involved in each thesis or dissertation research work. Next we briefly described the researcher's background and specific areas of interest.

—João Marcos Bastos Cavalcanti: received his Ph.D. degree from the University of Edinburgh (Scotland) in 2003. His main research areas are content-based image retrieval, image classification and Web data management.

—Marco Antônio Pinheiro de Cristo: received his Ph.D. degree from UFMG in 2006. His main research areas include information retrieval, data mining and digital libraries.

—David Braga Fernandes de Oliveira: received his Ph.D. degree from UFMG in 2010. His main research areas are search using semi-structured content and Web data management.

—Edleno Silva de Moura: received his Ph.D. degree from UFMG in 1999. His main research areas are information retrieval and Web data management.

—Altigran Soares da Silva: received his Ph.D. degree from the Federal University of Minas Gerais (UFMG) in 2002. His main interests include databases and Web data management.

Currently the group includes 13 Ph.D. students and 17 M.Sc. students. It also includes some undergraduate students which participate in research projects. The group has already concluded the supervision of 50 M.Sc. dissertations.

2.2 Funding

The BDRI group has continuously received research grants through several projects, either individually or in cooperation with other research groups in Brazil and abroad. Next we briefly describe some ongoing research projects in which the group is involved.

Advertising using Information Retrieval and Datamining Technologies (MCT/CNPq 474099/2007-5). This work concerns the identification and combination of evidences for proper advertising on the Web and Digital TV, using machine learning techniques.

National Institute for Web Research - INCT for the Web. This project is conducted in cooperation with the Federal University of Minas Gerais (UFMG), the Federal University of Rio Grande do Sul (UFRGS) and the Federal Institute of Technology of Minas Gerais.

Image Search and Classification Combining Multiple Features. In this project we study and propose solutions for search and classification of images. In particular we are interested in search on large and heterogeneous image databases, image version detection and remote sensing image classification. More specifically, it involves the development of image descriptors for search and classification tasks.

Information Retrieval on the Hidden Web. This project is carried out under the cooperation of the BDRI group and the Data Management Group of the University of Utah. It involves the investigation of problems related to the hidden Web, including generation of data collectors for feeding data extractors and automated organization of Web forms aiming at meta-search.

Automated Context-based Selection of Advertisements - Sabaco. This project has the goal of creating new algorithms for systems of advertisement selection, evaluating the proposed algorithms and constructing prototypes of context-based advertisement selection systems. It also has the goal of finding a partner company that can exploit the research results in a business environment.

2.3 Cooperation

The group maintains cooperation with several research groups in Brazil and abroad. Some of the institutions with which we had recent or ongoing cooperation are: University of Utah (USA), University of Alberta (Canada), Virginia Polytechnic Institute and State University (USA), Instituto Superior Técnico (Portugal), Federal University of Minas Gerais - UFMG, Federal University of Rio Grande do Sul - UFRGS, State University of Campinas - UNICAMP, State University of São Paulo - USP.

2.4 Other Activities

The members of the BDRI group participate as reviewers of scientific journals such as ACM Transactions on Information Systems, IEEE Multimedia, Journal of The Brazilian Computer Society (online), IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems (online), Information Systems (Oxford), Information Processing & Management, Software, Practice & Experience (Print), Data & Knowledge Engineering, and also have integrated the program committee of conferences, such as ACM SIGIR (2005-2007), World Wide Web Conference (2006-2010), SPIRE (2003-2007), ACM CIKM (2006-2007), ACM WSDM (since 2008), SBBD (since 2005).

Additionally, the members of the group have served in several roles and positions in such institutions as the Brazilian Academy of Sciences, the Brazilian Computing Society board,the Committee for Research and Development Activities in the Amazon, the Superior Council of the Amazonas State Research Foundation (FAPEAM), the Vice-presidency of Research and Graduate Studies at UFAM, and the coordination of the Graduate Program in Informatics.

3. MAIN RESEARCH TOPICS

In this section we present a sample of the research results produced by the group in the past years. While this is not an exhaustive list, the topics described in this section present the most important results obtained and a list of what we are currently developing. Thus, such a list of research topics provides a good overview about the diversity and quality of the research performed by the group.

Also note that such a list exemplifies our collaborative work with the companies and research groups described in Section 2.3.

3.1 Information Extraction from Textual Sources

The Web is abundant in textual content that holds implicit (semi) structured data. In many cases, these data are present not only in HTML pages, but also in *blog posts*, *tweets*, *RSS feeds*, etc. Moreover, they often occur without the presence of explicit markers and are organized in a structure also implied. The problem of extracting data from textual sources has been studied in the context of Web Data Management for more than one decade [Laender et al. 2002].

Our group has been working in the problem of information extraction by text segmentation (IETS), which consists of extracting semi-structured data records by identifying attribute values occurring in continuous text such as bibliographic citations, product descriptions, classified ads, etc. Currently, the most successful IETS methods are based on learning sequential models, such as Hidden Markov Models (HMM) [Borkar et al. 2001] and Conditional Random Fields (CRF) [Lafferty et al. 2001]. In fact, methods based on CRF are the state-of-the-art, outperforming HMM-based methods in experimental evaluations. In these models, input texts are considered as sequences of tokens or strings (composed by more than one token) to which labels must be assigned, so that these tokens/strings can then be identified as values of attributes.

In [dos Santos et al. 2006] we presented a Hidden Markov Models (HMMs) based approach to extracting data from semi-structured texts. Distinctly from previous proposals in the literature that also use HMMs, this 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 through a learning process, by 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.

Later on, we developed FLUX-CiM (Flexible Unsupervised Extraction of Citation Metadata) [Cortez et al. 2007; 2009], a knowledge-base approach to help extracting the correct components of bibliographic citations in any format. Differently from related approaches that rely on manually built knowledge-bases (KBs) for recognizing the components of a citation, in our case, such a KB is automatically constructed from an existing set of sample metadata records from a given area (e.g., computer science or health sciences). Our approach does not rely on patterns encoding specific delimiters of a particular citation style. It is also unsupervised, in the sense that it does not rely on a learning method that requires a training phase. These features assign to our technique a high degree of automation and flexibility. Results of experiments we have carried out indicate precision and recall levels above 94% and perfect extraction for the large majority of citations tested. We compared our approach against an information extraction method based on CRF, showing that ours produced superior results without the training phase.

We have also developed ONDUX (On Demand Unsupervised Information Extraction) [Cortez et al. 2010], a new unsupervised probabilistic approach for IETS. As other unsupervised IETS approaches, ONDUX relies on information available on pre-existing data to associate segments in the input string with attributes of a given domain. Unlike other approaches, we rely on very effective matching strategies instead of explicit learning strategies. The effectiveness of this matching strategy is also exploited to disambiguate the extraction of attributes through a reinforcement step that explores sequencing and positioning of attribute values directly learned on-demand from test data, with no previous human-driven training, a feature unique to ONDUX. This assigns to ONDUX a high degree of flexibility and results in superior effectiveness, as demonstrated by experimental evaluation with textual sources from different domains, in which ONDUX is compared with a state-of-art IETS approach.

Our newest work in this research topic is JUDIE (Joint Unsupervised Structure Discovery and Information Extraction) [Cortez et al. 2011]. While in state-of-the-art Information Extraction methods the structure of the data records is manually supplied by the user as a training step, JUDIE is capable of detecting the structure of each individual record being extracted without any user assistance. This is accomplished by a novel Structure Discovery algorithm that, given a sequence of labels representing attributes assigned to potential values, groups these labels into individual records by looking for frequent patterns of label repetitions among the given sequence. We also show how to integrate this algorithm in the information extraction process by means of successive refinement steps that alternate information extraction and structure discovery. Through an extensive experimental evaluation with different datasets in distinct domains, we compare JUDIE with state-of-the-art information extraction methods and conclude that, even without any user intervention, it is able to achieve high quality results on the tasks of discovering the structure of the records and extracting information from them.

An interesting application of our research on information extraction techniques is a method, called *iForm* [Toda et al. 2009; Toda et al. 2010], which we have developed for automatically using data-rich text for filling form-based input interfaces. Our method takes a text as input, extracts implicit data values from it and fills appropriate fields. For this task, it relies on knowledge obtained from values of previous submissions for each field, which are freely obtained from the usage of the interfaces. Our method exploits features related to the content and the style of these values, which are combined through a Bayesian framework. Through extensive experimentation, we show that our approach is feasible and effective, and that it works well even when only a few previous submissions to the input interface are available.

3.2 Keyword-Based Queries over Databases

Among the research topics of interest to our group, maybe the most representative of our hybrid profile, mixing data management and information retrieval, is the work on how to allow IR-style keyword-based queries to be issued and processed on structured databases. The motivation for this work is allowing users to specify queries in a more flexible and intuitive way. Indeed, the traditional approach for querying structured databases requires the user to know not only the syntax of some query language, but also to be aware of the details regarding the database schema.

In [Mesquita et al. 2007], we present LABRADOR, a method for efficiently publishing relational databases on the Web by using a simple text box query interface. This method takes an unstructured keyword-based query posed by a user and automatically derives 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 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.

More recently, we extended the LABRADOR approach for automatically deriving structured XML queries from keyword-based queries. This method, called StruX [Hummel et al. 2010], users specify a schema-independent unstructured keyword-based query and it automatically generates a top-k ranking of schema-aware queries based on a target XML database. Then, one of the top ranked structured queries can be selected, automatically or by a user, to be executed by an XML query engine. The generated structured queries are XPath expressions consisting of an entity path (e.g., `dblp/article`) and predicates (e.g., `/dblp/article[author="john" and title="xml"]`). We use the concept of entity, commonly adopted in the XML keyword search literature, to define suitable root nodes for the query results. Also, StruX uses IR techniques to determine in which elements a term is more likely to occur.

This research topic is currently been explored in two different ways: (1) allowing keyword-based queries being entirely processed using the resources available on a typical relational DBMS; and (2) allowing keyword-based queries being processed over XML streams.

3.3 Data Integration

Our main research initiatives within this topic is on problems related to data cleaning. In data integration tasks, records from a single dataset or from different sources must often be compared to identify records that represent the same real world entity. This is a hard problem whose solutions is important to improve the quality of the data resulting from integration process.

In [de Freitas et al. 2010] we propose a new method that uses Genetic Programming (GP) to automatically generate similarity functions to identify record replicas in a given dataset. The generated similarity functions properly combine and weight the best features available among the record fields in order to tell when two distinct records represent the same real-world entity. This method improves the one presented in [de Carvalho et al. 2006], significantly reducing the need for training data. This is achieved by using a semi-supervised GP process based on the active learning paradigm. In our method, a committee of multi-attribute functions votes for classifying record pairs as duplicates or not. When the committee majority voting is not enough to predict the class of the data pairs, a user is called to solve the conflict. Experimental results show that our method guarantees the quality of the deduplication process while reducing the number of labeled examples needed.

Another cleaning-related problem we have been addressing is *blocking*. During the deduplication process, the cost of searching for duplicate records grows quadratically as the number of records available in the data sources increases and, for this reason, direct approaches, such as comparing all record pairs, must be avoided. In this context, blocking methods are used to create groups of records that are likely to correspond to the same real world entity, so that the deduplication can be applied to these blocks only. In the recent literature, machine learning processes are used to find the best blocking function, based on a combination of low cost rules, which define how to perform the record blocking. In [Evangelista et al. 2010] we present a new blocking method based on machine learning. Different from previous methods, our method allows the use of a larger number of rules for defining blocking functions, leading to a more effective process for the identification of duplicate records. Experimental results with real and synthetic data show that our method achieves over 95% of correctness when generating blocks of potential duplicate. Due to this successful result, this work received the *SBBD Best Paper* award in 2009.

Another problem related to data integration we have been working on is *data exchange*, that is, translating data from one *source* collection into data that conforms to the schema of a *target* collection. Specifically, we propose in [Mesquita et al. 2007] a lightweight framework for data exchange that is suitable for non-expert and casual users sharing data on the Web or through peer-to-peer systems. Unlike previous work, we consider a simplistic data model and schema formalism that are suitable for describing typical on-line data, and propose algorithms for mapping such schemas as well as for translating the corresponding instances. Our solution requires minimal overhead and setup costs compared to existing data exchange systems, making it very attractive in the Web data exchange setting. We report experimental results indicating that our method works well with real Web data from various domains.

3.4 Focused Crawling

Focused crawling has emerged as an effective and efficient alternative to locate specific concepts on the Web. A concept can be a general topic (e.g., molecular biology), a specific subject (e.g., HIV virus) or object (e.g., call for papers, scientist biographies), etc. Focused crawling is crucial for many important applications, such as community information systems, vertical search engines and digital libraries.

For some of these applications, the criteria to determine whether a page is relevant to be collected 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 developed a new structure-driven approach for generating Web agents that requires a minimum effort from the users [Vidal et al. 2006b; 2006a] 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, our structure-driven crawlers have been able to collect all pages that match the samples given, including those pages added after the crawlers have been generated.

In [Vidal et al. 2008], we extended this approach to deal with the so-called *hidden Web*. The hidden Web consists of data that is generally hidden behind form interfaces, and as such, it is out of reach for traditional crawlers.

With the goal of leveraging the high-quality information in this largely unexplored portion of the Web, in [Vieira et al. 2008] we proposed a new strategy for automatically retrieving data hidden behind keyword-based form interfaces. Unlike previous approaches to solve this problem, our strategy adapts the query generation and selection by detecting features of the index. We describe an extensive experimental evaluation which shows that: our strategy is able to derive appropriate queries to obtain high coverage while, at the same time, avoiding the retrieval of redundant data; and it obtains higher coverage and is more efficient than approaches that use a fixed strategy for query generation.

Also on the context of the hidden Web, in [Barbosa et al. 2007] we addressed the problem of organizing hidden-web databases, in a joint work with the DB group at the University of Utah. Given a heterogeneous set of Web forms that serve as entry points to hidden-Web databases, our goal is to cluster the forms according to the database domains to which they belong. We propose a new clustering approach that models Web forms as a set of hyperlinked objects and considers visible information in the form context—both within and in the neighborhood of forms—as the basis for similarity comparison. Since the clustering is performed over features that can be automatically extracted, the process is scalable. In addition, because it uses a rich set of metadata, our approach is able to handle a wide range of forms, including content-rich forms that contain multiple attributes, as well as simple keyword-based search interfaces. An experimental evaluation over real Web data shows that our strategy generates high-quality clusters—measured both in terms of entropy and F-measure. This indicates that our approach provides an effective and general solution to the problem of organizing hidden-Web databases.

3.5 Information Retrieval Models

3.5.1 *Structure-aware IR systems.* Unlike plain text documents, Web pages are commonly composed of distinct segments such as main content, service channels, decoration skins, navigation bars, copyright and privacy announcements. These different segments, commonly referred to in the literature as *blocks*, can be automatically identified in Web pages and can be used to improve information retrieval tasks such as ranking, Web link analysis, and Web mining. However, the task of manually assigning weights to blocks requires specialized knowledge about the ranking function adopted in the search system, as well as a complex and expensive human effort to evaluate the relative importance of all regions found in the target Web site. For Web collections containing a large number of Web pages spread across multiple Web sites, which is a common scenario, this manual assignment of block weights may be unfeasible in practice. To avoid some of these practical problems, we proposed methods for automatically computing block weight factors and for using such weights to rank documents in Web search systems [Fernandes et al. 2007; de Moura et al. 2010]

In [Fernandes et al. 2011] we proposed a fully automatic method for page segmentation that adopts a DOM tree alignment strategy to solve the problem of template detection [Vieira et al. 2006]. The results indicated that the proposed method produces better segmentation results when compared to the best segmentation method we found in literature. Further, when applied as input to the segment aware Web search method proposed by us in [de Moura et al. 2010], it produces results close to those produced when using a manual page segmentation method.

We have also worked on the design of new algorithms to detect and remove template from Web pages. Templates are pieces of HTML code shared by a set of pages of a same Web site. They are usually generated using authoring/publishing tools or by programs that build HTML pages to publish content from a database. In spite of their usefulness, the information available in templates is redundant and thus processing and storing such information just once for a set of pages may save computational resources. Recognizing this problem, we proposed fast and accurate methods for detection of templates [Vieira et al. 2006; Vieira et al. 2009]. The proposed method works in two steps. Initially, templates are detected using a set of sample pages, and then this information is used to remove the templates present in the other pages in the collection. The experimental evaluation conduced show that this approach is effective for identifying terms occurring in templates, obtaining F-measure values around 0.9. In [Vieira et al. 2009] we presented methods for detecting templates considering a scenario where multiple templates can be found in a collection of Web pages. The proposed idea was to efficiently partition the input collection into clusters of pages that contain a common template, and then apply a single-template detection procedure over each cluster. The experimental results presented in this paper, that was conduced over a representative set of Web pages, show that the proposed approach is efficient and scalable while obtaining accurate results. This paper presents experiments over a representative set of Web pages, and shows that the proposed approach is efficient and scalable while obtaining accurate results.

3.5.2 *Hypergraph Model.* In another research effort related to Web search, we proposed a representation of the Web as a directed hypergraph, instead of a graph, where links can connect not only pairs of pages, but also pairs of disjoint sets of pages [Berlt et al. 2010; Berlt et al. 2007]. In our model, the Web hypergraph is derived from the Web graph by dividing the set of pages into non-overlapping sets and using the links between pages of distinct sets to create hyperarcs. Each hyperarc connects a set of pages to a single page and is created with the goal of providing more reliable information for link analysis methods. We used the hypergraph structure to compute the reputation of Web pages by experimenting hypergraph versions of two previously proposed link analysis methods, Pagerank and Indegree. The hypergraph versions of the two methods are referred to as HyperPagerank and HyperIndegree, respectively.

We experimented the methods combining the page reputation with the textual content of the pages and with the anchor text information available on the collection adopted in the experiments. Two combination strategies previously proposed in literature were adopted. The experiments were performed dividing the query sets according to their types, into navigational and informational, and according to their popularity. We present experiments which indicate the hypergraph versions of Pagerank and Indegree produce better results when compared to their original graph versions.

3.5.3 *Ranking Learning.* Modern Web search engines use different strategies to improve the overall quality of their document rankings. Usually the strategy adopted involves the combination of multiple sources of relevance into a single ranking. Considering this scenario, we have studied the use of Genetic Programming (GP) to derive good evidence combination functions using different sources of evidence of relevance [Silva et al. 2009]. The idea is to select a set of representative queries that is then evaluated by users to produce examples of how documents should be ranked in the system for each of the sample queries in the set. We then use this training set as input for learning methods which try to generalize the ranking strategy to apply it in future queries.

In [Silva et al. 2009] we proposed the use of evolutionary techniques to derive good evidence combination functions using three different sources of evidence of relevance: the textual content of documents, the reputation of documents extracted from the connectivity information available in the processed collection and the anchor text concatenation. The combination functions discovered by our evolutionary strategies were tested using a collection containing 368 queries extracted from a real nation-wide search engine query log with over 12 million documents. The experiments performed indicate that our proposal is an effective and practical alternative for combining sources of evidence into a single ranking. We also show that different types of queries submitted to a search engine can require different combination functions and that our proposal is useful for coping with such differences.

3.6 Efficiency Issues in Web Search

Information retrieval systems need to be not only highly effective but also extremely efficient, since query throughput is a central problem in these systems. In this section we describe some of the results obtained in this topic by the two groups.

3.6.1 *Locality-Based Pruning.* One way to address query processing efficiency without losing effectiveness is to reduce the amount of data to be processed at query time. In [de Moura et al. 2008], we address the issue of compressing the search engine's textual database from an atypical perspective. We take the occurrence position of words in the text as an input for determining its importance in the collection. The novelty of this work arises from the fact that positional information can improve the quality of the final pruning, allowing further reduction in the indexes sizes without loss of quality in the ranking.

We used this new approach to propose and experiment simple and effective pruning methods that allow a fast construction of the pruned index. The proposed methods are specially useful for pruning in environments where the document database changes continuously, such as large scale Web search engines. Extensive experiments are presented showing that the proposed methods can achieve high compression rates while maintaining the quality of results for the most common query types present in modern search engines, i.e. conjunctive and phrase queries. In the experiments, our locality based pruning approach allowed reducing search engine indexes to 30% of their original size, with almost no reduction in precision at the top answers. Furthermore, we concluded that even an extremely simple locality based pruning method can be competitive when compared to complex methods that do not rely on locality information. We are now working to further developing this idea by proposing new dynamic pruning methods that also use locality information as a heuristic to select indexes entries to be discarded at query processing time.

3.6.2 *Removing Replicated Web Sites.* Identifying replicated sites is an important task for search engines. 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 user. We proposed in this topic a new approach to select Web sites that are likely to be replicas in a Web search engine database. Our first method uses the websites' structure and the content of their pages to detect possible replicas [da Costa Carvalho et al. 2007]. In a following work, we applied genetic programming to this problem [Carvalho et al. 2008].

Besides the waste of resources, replicated information also affects the quality of the answers provided to the search engine users. This is mainly because replicas insert duplicated connectivity information in Web collections, causing anomalies on the functioning of search engine ranking algorithms. This is a major problem, since connectivity information is one of the most important pieces of evidence to compute the score of Web pages given a query in modern because search engine technology [Calado et al. 2003]. Another obvious problem is the possibility of having different answers representing the same content, giving the user the feeling that answers are repeated. Therefore, the removal of replicated information from a Web collection not only reduces costs but also improves the quality of the service provided by the search engine.

The task of detecting replicated Web sites seems to be simple at first. For instance, one could compare all the pages of Web sites in the database in a pairwise fashion in order to identify possible replicas. However, in large scale search engines the cost to perform such comparison would be prohibitive. A general solution to this problem found in the literature [Bharat et al. 2000] is to adopt heuristics to previously select pairs of sites that are more likely to be replicas, and then perform the detailed page comparison only for these pairs. The quality of a replica detection method is then directly related to the accuracy of heuristics in finding such promising pairs. Previous solutions to find replicated Web sites do not take the content of pages in account, with the claim that the content turns the process unacceptably expensive. We proposed an alternative method that finds replica candidate pairs by taking the advantages of using the content, while not increasing the processing times. Our method improves the quality of the replica candidate detection task in 47.23% when compared to previously proposed methods, being extremely useful in practice [da Costa Carvalho et al. 2007].

3.6.3 *Removing Noise Links.* The currently booming search engine industry has determined many online organizations to attempt to artificially increase their ranking in order to attract more visitors to their Web sites. At the same time, the growth of the Web has also inherently generated several navigational hyperlink structures which have a negative impact on the importance measures employed by current search engines. We studied this problem and have proposed and evaluated algorithms for identifying all these noisy links over the Web graph, may them be spam or simple relationships between real world entities represented by sites, replication of content, etc. Unlike prior work, we target a different type of noisy link structures, residing at the site level, instead of the page level.

We investigated three main types of site level relationships: mutual reinforcement (in which many links are exchanged between the two sites), abnormal support (where most of one site's links are pointing to the same target site), and link alliances (in which several sites create complex link structures that boost the PageRank score of their pages) [Carvalho and Silva 2003]. When the relation between such sets of sites is considered suspicious, we assume that the links between them are noisy and penalize them accordingly. Our experiments show a very strong increase in the quality of the output rankings after having applied our techniques.

3.7 Advertising

An important topic related to web information retrieval and query processing is the problem of selecting ads in content-based advertisement systems. Content-based advertising constitutes the key web monetization strategy nowadays. We have been working on this topic in cooperation with the LATIN laboratory at UFMG. We regard this specific problem as one of the most important research directions for our group. More specifically, our first work was related to ad selection [Ribeiro-Neto et al. 2005]. We presented approaches in which external sources of information are used to improve ranking on content-based advertising. We proposed and evaluated several methods for matching pages and advertisements, and determined how accurate they were in picking the most relevant ads to the content of a Web page. In a following work [Lacerda et al. 2006], we proposed a new framework for associating ads with web pages based on Genetic Programming (GP). Our GP method aimed to learn functions that select the most appropriate ads, given the contents of a Web page. These ranking functions were designed to optimize overall precision and minimize the number of misplacements.

3.8 Digital Libraries

Our group have been working in several topics related to Digital Libraries in close cooperation with groups from UFMG. Some of this topics are detailed above.

3.8.1 *Topic Classification.* A Digital Library (DL) system encompasses a range of services to collect, manage and preserve digital content. A common task of such a system is to classify data according to its content in order to make it more convenient management operations such as data storing and presentation. In particular, we have studied many aspects of the classification task, such as the methods and evidence to be used.

Regarding the study of evidences, we have explored how linkage information inherent to different collections can be used to enhance the effectiveness of classification algorithms [Couto et al. 2010; Calado et al. 2006; Couto et al. 2006; Calado et al. 2003; Cristo et al. 2003; Zhang et al. 2005; Zhang et al. 2004]. We have studied three link-based bibliometric measures (co-citation, bibliographic coupling and Amsler) on different collections, using several algorithms and learning strategies. We found that both hyperlink and citation information can be used to learn reliable and effective classifiers. In one of the test collections, we obtained improvements of about 70% over a traditional text-based kNN classifier. We also studied alternative ways of combining bibliometric based classifiers with text based classifiers. Specifically, regarding methods, we have proposed a novel approach for classifying documents that combines different pieces of evidence (e.g., textual features of documents, links, and citations) transparently, through a data mining technique which generates rules associating these pieces of evidence to predefined classes [Veloso et al. 2006]. The proposed approach employs a lazy method which delays the inductive process until a document is given for classification, therefore taking advantage of better qualitative evidence coming from the document. We found that our approach was able to outperform traditional classifiers based on the best available evidence in isolation as well as state-of-the-art multi-evidence classifiers.

3.8.2 *Quality Evaluation.* An important issue regarding DLs is to determine the relative quality of its contents. We have addressed this problem by suggesting automatic methods to infer content quality of information provided by users in Web 2.0 sites. For instance, in [Hasan Dalip et al. 2009], we have studied many pieces of evidence, some of them proposed by us, to assess its usefulness as content quality indicators in Wikipedia. To accomplish this, we used machine learning techniques and observed that the most important ones are the easiest to extract, namely, textual features related to length, structure and style. We have further shown that our method outperformed previous approaches in terms of effective quality prediction. We also studied the quality of the information provided by users in web sites centered around active web communities, such as Youtube, YahooVideo, LastFM and CiteULike [Figueiredo et al. 2009]. More specifically, we investigated textual attributes such as title, description, tags, and comments provided by the users considering three aspects: utilization; discriminative and descriptive power. As a result, we have found that (1) collaborative textual attributes, although not significantly exploited in some applications, contain the largest amount of information when present, (2) there is a significant diversity of information between the textual attributes, and (3) the title and tags of the objects seem to be the most promising attributes for IR services, whereas the former is almost always present and has a high power specification, and second, when used, has a high discriminative and descriptive power.

3.8.3 *Tools.* A consequence of our research is the development of tools useful for DLs. For instance, in [Silva et al. 2009], we proposed a process to automatically retrieve metadata related to documents in a DL. The process uses results from queries submitted to Web search engines for finding the full text or any related material corresponding to the documents. After an empirical analysis, we found that a simple re-ranking strategy over the combined results from Scholar and Google would allow us to retrieve most of the information missing in the DL.

3.9 Query Processing

3.9.1 *Related Queries.* In a joint work with the LATIN laboratory at UFMG, we developed a method which aims to improve the quality of query results provided by search engines. To accomplish this, we use previous queries submitted to such system to determine the relation among them. Then, we make use of this information to improve the results for new queries. In that 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 our model reaches a precision of 90.5% in 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. 2005; Fonseca et al. 2004; Fonseca et al. 2003].

3.9.2 *Query Topic.* We have also worked in the topic of determining user goals behind a query in order to produce specialized ranking functions [Herrera et al. 2010]. Queries submitted to search engines can be classified according to the user goals into three distinct categories: navigational, informational and transactional. Such classification may be useful, for instance, as additional information for advertisement selection algorithms and for search engine ranking functions, among other possible applications. We have studied the impact of using several features extracted from the document collection and query logs on the task of automatically identifying the users' goals behind their queries. We propose the use of new features not previously reported in the literature and study their impact on the quality of the query classification task. Experimental results indicate that the new proposed set of features improves the quality of the classification task when compared to previous proposals. We report experiments with two web collections where we were able to obtain 82.5% and 77.67% of overall accuracy when classifying queries according to the three distinct user goals studied.

3.10 Content-Based Image Retrieval

In this area we are particularly interested in investigating and developing image descriptors for searching on large and heterogeneous image databases. Several image descriptors proposed in the literature achieve high levels of efficiency and accuracy. However most of them run experiments using small image databases (less than 20,000 images). Most of these image databases are composed of well defined categories, which facilitates the search task resulting in high precision levels. It is reported in the literature that the overall effectiveness of image descriptors is relatively low in large and heterogeneous databases with no knowledge or previous categorization of the images.

The main goal of our work is to define and show that our visual descriptors outperforms state-of-the-art descriptors for applications with large and non-specific domain image collections. We have carried out experiments using several known image databases such as Wang (1,000 images with 10 categories, each with 100 images) and MPEG-7 CCD (5,466 images having 50 evaluated queries) in the small image database size category. We are also working with a large image database collected from the Yahoo! directory with over 100,000 images. In addition we also have the CoPhIR Test Collection with 106 million images extracted from Flickr.

4. INDUSTRY RELATIONS

The activities of technology transfer to the society are among the main goals of the BDRI group. We keep permanent contact with the industry and other society segments in an attempt to identify opportunities to develop cooperation projects and try to apply the knowledge produced by our research in the solution of practical problems. Further, we motivate and provide support to alumni interested in create startups, besides our own initiative to create startups in cooperation with other groups.

Among the companies and research institutes we have developed cooperation projects we include: (a) The Tropico Sistemas, where we developed software to improve the management of information in the company; (b) Instituto Nokia de Tecnologia, where we have worked in recommendation systems for mobile digital tv environments; (c) UOL S/A, where we have developed several research projects, including solutions to classification, ranking, advertising and data management problems; (d) Fabriq S/A, where we have worked to improve Intranet search systems developed by the company. Besides the cooperation with established companies, our group has also participated in the creation of Internet startups.

Examples of company startups with which the group has cooperated include: (a) Akwan Information Technologies S/A, a startup that developed search systems and that was sold to Google in 2005; (b) Zunnit technologies, a company created in 2008 focused on the development of technology to recommendation systems; (c) Nhemu Tecnologias de Internet LTDA, a company created in 2010 that develops technologies to online shopping stores.

These examples illustrate the constant effort of the group to maintain its research activities focused on practical problems and also on developing high quality solutions with high economical and social value.

5. FINAL REMARKS

In less than 10 years of activities, our group has built a solid reputation as an engaged research group in the areas of Information Retrieval, Data Management and Data Mining, with a focus on Web based applications, such as search engines, digital libraries, and social networks.

The research efforts we have developed in these areas have brought significant contributions and resulted in articles presented in conferences such as VLDB, SIGMOD, SIGIR, WWW and ICDE, as well as articles published in high prestigious journals such as *Transactions on Information Systems*, *Information Systems* and *Data & Knowledge Engineering*. For two years, 2007 and 2009, we have received the *Best Paper Award* from SBBD.

As for human resources, our group has formed 50 master students so far, all of them playing key roles in the industry, academia and on the government. Three of these students had their work classified among the top ten Computer Science master thesis in Brazil by the Brazilian Computer Society in its yearly graduate work competition, in 2009, 2010 and 2011. The first Doctorate students are expected to finish their work in 2012.

We continue to work in most of the research topics described here, and we expect to present new contributions in the near future. One of the main current efforts of our group is concentrated on the use of Information Retrieval, Data Management and Data Mining techniques to improve topics of high economical impact, such as advertising and recommender systems. Such research topics have introduced new challenging technical problems and raise interesting questions, such as: (a) to determine how to take advantage of social networks and collaboratively created content to improve web systems, (b) to investigate how to effectively organize the large amount of semi-structured data presented in web-based applications, and (c) to design ranking functions able to satisfy conflicting goals between relevance and revenue are examples of new challenges presented in our research areas.

Finally, it is very important to highlight that the research we develop and the results we have achieved are only possible due to the invaluable resources we have received from national agencies such as CNPq, CAPES and FINEP, as well as from regional agencies such as FAPEAM and SUFRAMA.

REFERENCES

Barbosa, L., Freire, J., and da Silva, A. S. Organizing hidden-web databases by clustering visible web documents. In *Proc. of the International Conference on Data Engineering.* Istanbul, Turkey, pp. 326–335, 2007.

Berlt, K., de Moura, E. S., Carvalho, A., Cristo, M., Ziviani, N., and Couto, T. Modeling the web as a hypergraph to compute page reputation. *Inf. Syst.* vol. 35, pp. 530–543, July, 2010.

Berlt, K., de Moura, E. S., da Costa Carvalho, A. L., Cristo, M., Ziviani, N., and Couto, T. A hypergraph model for computing page reputation on web collections. In *Proc. of the Brazilian Symposium on Databases.* Joao Pessoa, PB, Brazil, pp. 35–49, 2007.

Bharat, K., Broder, A. Z., Dean, J., and Henzinger, M. R. A comparison of techniques to find mirrored hosts on the www. *IEEE Data Eng. Bull.* 23 (4): 21–26, 2000.

Borkar, V. R., Deshmukh, K., and Sarawagi, S. Automatic segmentation of text into structured records. In *Proc. of the ACM International Conference on Management of Data*. Santa Barbara, CA, USA, pp. 175–186, 2001.

Calado, P., Cristo, M., Gonçalves, M. A., de Moura, E. S., Ribeiro-Neto, B., and Ziviani, N. 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.

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

Calado, P., Moura, E. S., Ribeiro-Neto, B., Reis, I., and Ziviani, N. Local versus global link information. *ACM Transactions of Information Systems* 21 (1): 1–22, 2003.

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

Carvalho, M. G., Laender, A. H. F., Gonçalves, M. A., and da Silva, A. S. Replica identification using genetic programming. In *Proc. of the ACM symposium on Applied computing*. Fortaleza, Ceara, Brazil, pp. 1801–1806, 2008.

Cortez, E., da Silva, A. S., de Moura, E. S., and Laender, A. H. F. Joint unsupervised structure discovery and information extraction. In *Proc. of the ACM International Conference on Management of Data*. Athens, Greece, pp. 541–552, 2011.

Cortez, E., da Silva, A. S., Gonçalves, M. A., and de Moura, E. S. ONDUX: on-demand unsupervised learning for information extraction. In *Proc. of the ACM International Conference on Management of Data*. Indianapolis, IN, USA, pp. 807–818, 2010.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., and de Moura, E. S. FLUX-CIM: flexible unsupervised extraction of citation metadata. In *Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*. pp. Vancouver, BC, Canada, 2007.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., and de Moura, E. S. A flexible approach for extracting metadata from bibliographic citations. *Journal of the American Society for Information Science and Technology* 60 (6): 1144–1158, 2009.

Couto, T., Cristo, M., Gonçalves, M. A., Calado, P., Ziviani, N., Moura, E., and Ribeiro-Neto, B. A comparative study of citations and links in document classification. In *Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*. Chapel Hill, NC, USA, pp. 75–84, 2006.

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

Cristo, M., Calado, P., de Moura, E. S., Ziviani, N., and Ribeiro-Neto, B. A. Link information as a similarity measure in web classification. In *Proc. of the International Symposium on String Processing and Information Retrieval*, M. A. Nascimento, E. S. de Moura, and A. L. Oliveira (Eds.). Lecture Notes in Computer Science, vol. 2857. Springer, Manaus, Brazil, pp. 43–55, 2003.

da Costa Carvalho, A. L., de Moura, E. S., da Silva, A. S., Berlt, K., and Bezerra, A. A cost-effective method for detecting web site replicas on search engine databases. *Data Knowl. Eng.* vol. 62, pp. 421–437, September, 2007.

de Carvalho, M. G., Gonçalves, M. A., Laender, A. H. F., and da Silva, A. S. Learning to deduplicate. In *Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*. Chapel Hill, NC, USA, pp. 41–50, 2006.

de Freitas, J., Pappa, G., da Silva, A., Gonçalves, M., Moura, E., Veloso, A., Laender, A., and de Carvalho, M. Active learning genetic programming for record deduplication. In *Proc. of the IEEE Congrees on Evolutionary Computation*. Barcelona, Spain, pp. 1–8, 2010.

de Moura, E. S., dos Santos, C. F., de Araujo, B. D. S., da Silva, A. S., Calado, P., and Nascimento, M. A. Locality-based pruning methods for web search. *ACM Transactions of Information Systems* 26 (2), 2008.

de Moura, E. S., Fernandes, D., Ribeiro-Neto, B., da Silva, A. S., and Gonçalves, M. A. Using structural information to improve search in web collections. *Journal of the American Society for Information Science and Technology* vol. 61, pp. 2503–2513, December, 2010.

dos Santos, R. O., de Sá Mesquita, F., da Silva, A. S., and Vilarinho, E. C. C. Extração de dados e metadados em textos semi-estruturados usando hmms. In *Proc. of the Brazilian Symposium on Databases*. Forianopolis, SC, Brazil, pp. 117–131, 2006.

Evangelista, L. O., Cortez, E., da Silva, A. S., and Jr., W. M. Adaptive and flexible blocking for record linkage tasks. *JIDM* 1 (2): 583–597, 2010.

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., and Braga, E. A site oriented method for segmenting web pages. In *Proc. of the ACM Conference on Research and Development in Information Retrieval*. Beijing, China, 2011.

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., and Gonçalves, M. A. Computing block importance for searching on web sites. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Lisbon, Portugal, pp. 165–174, 2007.

Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M., Fernandes, D., Moura, E., and Cristo, M. Evidence of quality of textual features on the web 2.0. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Hong Kong, China, pp. 909–918, 2009.

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

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

Fonseca, B. M., Golgher, P., Pôssas, B., Ribeiro-Neto, B., and Ziviani, N. Concept-based interactive query expansion. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Bremen, Germany, pp. 696–703, 2005.

Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*. Austin, TX, USA, pp. 295–304, 2009.

Herrera, M. R., de Moura, E. S., Cristo, M., Silva, T. P., and da Silva, A. S. Exploring features for the automatic identification of user goals in web search. *Information Processing and Management* vol. 46, pp. 131–142, March, 2010.

Hummel, F., da Silva, A. S., Moro, M., and Laender, A. H. F. Automatically generating structured queries in xml keyword search. In *Proc. of the International Workshop of the Initiative for the Evaluation of XML Retrieval*. Vugh, The Netherlands, pp. 194–205, 2010.

Lacerda, A., Cristo, M., Gonçalves, M. A., Fan, W., Ziviani, N., and Ribeiro-Neto, B. A. Learning to advertise. In *Proc. of the ACM Conference on Research and Development in Information Retrieval*. Seattle, WA, USA, pp. 549–556, 2006.

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

Lafferty, J., McCallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the International Conference on Machine Learning*. Williamstown, MA, USA, pp. 282–289, 2001.

Mesquita, F., Barbosa, D., Cortez, E., and da Silva, A. S. FleDEx: flexible data exchange. Lisbon, Portugal, pp. 25–32, 2007.

Mesquita, F., da Silva, A. S., de Moura, E. S., Calado, P., and Laender, A. H. F. Labrador: Efficiently publishing relational databases on the web by using keyword-based query interfaces. *Information Processing and Management* 43 (4): 983–1004, 2007.

Ribeiro-Neto, B. A., Cristo, M., Golgher, P. B., and de Moura, E. S. Impedance coupling in content-targeted advertising. In *Proc. of the ACM Conference on Research and Development in Information Retrieval*. Salvador, Bahia, Brazil, pp. 496–503, 2005.

Silva, A. J. C., Gonçalves, M. A., Laender, A. H. F., Modesto, M. A. B., Cristo, M., and Ziviani, N. Finding what is missing from a digital library: A case study in the computer science field. *Information Processing and Management* 45 (3): 380–391, 2009.

Silva, T. P. C., de Moura, E. S., Cavalcanti, J. a. M. B., da Silva, A. S., de Carvalho, M. G., and Gonçalves, M. A. An evolutionary approach for combining different sources of evidence in search engines. *Inf. Syst.* vol. 34, pp. 276–289, April, 2009.

Toda, G. A., Cortez, E., da Silva, A. S., and de Moura, E. A probabilistic approach for automatically filling form-based web interfaces. *Procedings of the VLDB Endowment* 4 (3): 151–160, December, 2010.

Toda, G. A., Cortez, E., Mesquita, F., da Silva, A. S., Moura, E., and Neubert, M. Automatically filling form-based web interfaces with free text inputs. In *Proc. of the International Conference on World Wide Web*. Madrid, Spain, pp. 1163–1164, 2009.

Veloso, A., Meira, Jr., W., Cristo, M., Gonçalves, M., and Zaki, M. Multi-evidence, multi-criteria, lazy associative document classification. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Arlington, Virginia, USA, pp. 218–227, 2006.

Vidal, M. L. A., da Silva, A. S., de Moura, E. S., and Cavalcanti, J. a. M. B. Gogetit!: a tool for generating structure-driven web crawlers. In *Proc. of the International Conference on World Wide Web*. Edinburgh, Scotland, pp. 1011–1012, 2006a.

Vidal, M. L. A., da Silva, A. S., de Moura, E. S., and Cavalcanti, J. a. M. B. Structure-driven crawler generation by example. In *Proc. of the ACM Conference on Research and Development in Information Retrieval*. Seattle, WA, USA, pp. 292–299, 2006b.

Vidal, M. L. A., da Silva, A. S., de Moura, E. S., and Cavalcanti, J. M. B. Structure-based crawling in the hidden web. *Journal of Universal Computer Science* 14 (11): 1857–1876, 2008.

Vieira, K., Barbosa, L., Freire, J., and da Silva, A. S. Siphon++: a hidden-webcrawler for keyword-based interfaces (short paper). In *Proc. of the ACM International Conference on Information and Knowledge Management*. Napa Valley, CA, USA, pp. 1361–1362, 2008.

Vieira, K., Costa Carvalho, A. L., Berlt, K., Moura, E. S., Silva, A. S., and Freire, J. On finding templates on web collections. *World Wide Web* vol. 12, pp. 171–211, June, 2009.

Vieira, K., da Silva, A. S., Pinto, N., de Moura, E. S., Cavalcanti, J. M. B., and Freire, J. A fast and robust method for web page template detection and removal. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Arlington, USA, pp. 258–267, 2006.

Zhang, B., Chen, Y., Fan, W., Fox, E. A., Gonçalves, M., Cristo, M., and Calado, P. Intelligent gp fusion from multiple sources for text classification. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Bremen, Germany, pp. 477–484, 2005.

Zhang, B., Gonçalves, M. A., Fan, W., Chen, Y., Fox, E. A., Calado, P., and Cristo, M. Combining structural and citation-based evidence for text classification. In *Proc. of the ACM International Conference on Information and Knowledge Management*. Washington, D.C., USA, pp. 162–163, 2004.