

Vertical Searching in Juridical Digital Libraries

Maria de Lourdes da Silveira^{1,2,5}, Berthier Ribeiro-Neto^{1,3,6}, Rodrigo de Freitas Vale³, and Rodrigo Tôres Assumpção⁴

¹ Computer Science Department of the Federal University of Minas Gerais

² Prodabel - Information Technology Company for the City of Belo Horizonte, Brazil

³ Akwan Information Technologies

⁴ Central Bank of Brazil

⁵ Partially supported by Brazilian CNPq scholarship grant 141294/2000-0

⁶ Partially supported by Brazilian CNPq Individual Grant 300.188/95-1, Finep/MCT/CNPq Grant 76.97.1016.00 (project SIAM), and CNPq Grant 17.0435/01-6 (project I³DL)

Abstract. In the world of modern digital libraries, the searching for juridical information of interest is a current and relevant problem. We approach this problem from the perspective that a new searching mechanism, specialized in the juridical area, will work better than standard solutions. We propose a specialized (or *vertical*) searching mechanism that combines information from a juridical thesaurus with information generated by a standard searching mechanism (the classic vector space model), using the framework of a Bayesian belief network. Our vertical searching mechanism is evaluated using a reference collection of 552,573 documents. The results show improvements in retrieval performance, suggesting that the study and development of vertical searching mechanisms is a promising research direction.

Keywords: vertical searching, thesaurus, digital library, juridical area, ranking.

1 Introduction

The juridical literature is always expanding. A good part of this literature is now available online and accessible through the Internet. Such accessibility facilitates the access to specialized juridical information, but introduces new problems on its own. Lawyers have limited time for bibliographical searching and frequently have limited access to important information sources. Also they have great difficulty in identifying the most relevant information within the vast juridical collections of today. As a result, people involved with the law and its use are usually unable to take full advantage of computer devices for accomplishing the tasks of juridical analysis and searching.

A standard approach to this problem is to directly apply Information Retrieval (IR) techniques to a full text collection from the juridical domain. We investigate this issue and evaluate its effectiveness. While this approach does provide a solution to the problem of finding relevant information in a large juridical collection (or digital library), it does not take into account any specialized knowledge from the juridical arena. That is, there is small effort into using knowledge from the law field to improve the searching for specialized juridical information. This clearly seems to be a strong limitation.

An alternative approach is to combine specialized juridical knowledge, encoded in the form of a thesaurus, with evidence generated by standard IR techniques. In this case, the focus is on building a specialized searching mechanism (or algorithm) for juridical digital libraries. Since the search space is restricted to juridical documents and to the thesaurus concepts, we say that this new algorithm implements a type of *vertical searching*.

Investigating the problem of vertical searching for juridical documents is the path we follow here. We consider a specific form of juridical knowledge, i.e., information provided by a controlled vocabulary of juridical concepts and the relationships among them, encoded in the form of a thesaurus. Given this juridical thesaurus, we study the problem of how to improve the quality of the answers generated by the system, i.e., how to improve retrieval performance.

To combine evidence from standard IR techniques (i.e., keyword-based searching) with knowledge derived from the juridical thesaurus, we adopt the framework of Bayesian networks [10].

Bayesian networks are useful because they allow combining distinct sources of evidence in consistent fashion. Our Bayesian framework leads to a new ranking formula. Through experimentation, we show that this new formula yields improvements in retrieval performance.

The paper is organized as follows. In Section 2, we discuss related work. In Section 3 we present the extended belief network model for juridical digital libraries. In Section 4 we briefly introduce the Federal Justice Council (CJF) Thesaurus, which we adopted, present the reference collection, and discuss our experimental results. Our conclusions follow.

2 Related Work

Traditionally, thesauri are constructed manually for use in a specific area of knowledge. The objective of a thesaurus is to represent concepts and specify their relationships. The more commonly represented relationships are those of equivalence, hierarchy and associativity. In equivalence relationships, the concepts are synonyms, quasi-synonyms, and partial synonyms. In hierarchy relationships, the concepts have a subordination relationship, like broad and narrow terms. In associativity relationships, the concepts have a horizontal relationship, different from the previous ones, defined by specialists in the knowledge domain [6].

A thesaurus can be used both for indexing a collection and for searching information of interest in the collection. In the task of indexing, the thesaurus is used as a list of authorized words that normalize the indexing language. In the task of searching, the thesaurus is used to show to the user associations between its concepts. These associations might lead the user to concepts related to the query concept that were not foreseen by him and that might be useful in the query formulation process [1]. Another use of a thesaurus is as a tool to assist the users with the specification of the query subject [15].

Many authors have used thesaurus in automatic query expansion [2, 3, 5, 7, 11]. Greenberg investigated the use of a business thesaurus for query expansion [2, 3]. The experiments were executed over a collection of business related documents. The queries were real queries selected from questions formulated by business students. The author considered only queries that could be mapped to a concept in the thesaurus. She concluded that, if the focus is on improving precision figures, synonyms and narrow terms should be used for query expansion. Otherwise, if the focus is on improving recall figures, related terms and broad terms should be used for query expansion.

Kristensen [5] studied the use of synonym, narrow and related terms in query expansion in a full text newspaper database. She found enhancements in recall and reduction in precision, and concluded that a thesaurus is “clearly a recall-enhancing tool”. Her research showed specifically that synonyms and related terms are equivalent in recall possibilities and that synonyms degrade precision more than related terms.

Mandala et al. [7] used three types of thesauri to select terms for automatic query expansion. The types of thesauri were hand-crafted, generated automatically using co-occurrence patterns, and generated automatically using predicate-argument logic to gather linguistic relations. They developed algorithms to estimate the probability that the terms of each thesaurus are good candidates for query expansion. They showed that the combination of the three types of thesauri yields good results.

Qiu et al. [11] proposed a query expansion algorithm to select and weight the terms to be used in automatic query expansion. They considered the query as a concept and showed how to find good candidate terms for automatic query expansion, using the similarity of the whole query concept to the terms of the collection. They argued that consideration of a query concept is a superior approach.

Ribeiro-Neto et al. [13] used belief networks to combine evidence obtained from past queries with keyword-based evidence. Using four distinct reference collections, they report improvements in precision of at least 59% over keyword-based searching, showing that past queries can be used to improve retrieval performance when combined with keyword-based searching.

Silva et al. [16] used belief networks to combine evidence from the link structure of the Web with keyword-based evidence. The link information is composed of hub evidence (document with

a high number of outgoing links) and authority evidence (documents with a high number of incoming links). They report experiments combining keyword-based, hub, and authority evidential information that show gains in precision of 74% over keyword-based searching.

In here, we investigate the use of the concepts and the relationships of a juridical thesaurus as a source of evidential knowledge for improving query results. Contrary to the work in [2, 3], we use the framework of Bayesian networks as a formal underpinning. This provides for modularity and extensibility to naturally incorporate new sources of evidence into the model.

3 Extended Belief Network for a Juridical Digital Library

We adopt the belief network model proposed in [12] as a framework for combining distinct sources of evidential information in support of a ranking of the documents. This network model is extended with new nodes, edges, and probabilities to fit the information encoded in the juridical thesaurus. We say that this expansion is modular in the sense that it preserves all properties of the previous network. Figure 1 illustrates our extended belief network for a juridical digital library. The left hand side represents the original belief network proposed in [12] and the right hand side represents the extension due to the juridical thesaurus.

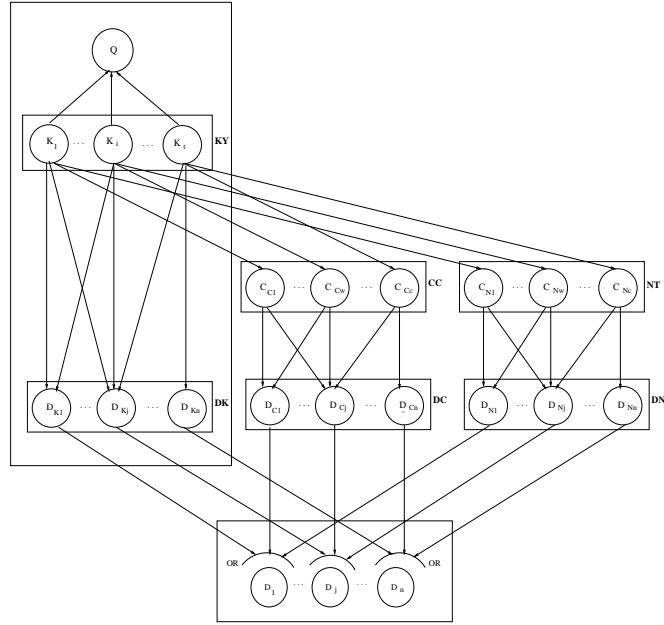


Fig. 1. Extended belief network for a juridical digital library.

In Figure 1, the keywords that compose the user query \mathbf{Q} are represented by nodes from the set \mathbf{KY} . The concepts, which are represented by nodes from the set \mathbf{CC} , are obtained automatically by directly mapping the query \mathbf{Q} into the concepts of the thesaurus. From the mapped concepts, the thesaurus allows inferring related narrow terms, represented by nodes from the set \mathbf{NT} . We defer to represent broad terms, related terms, and synonyms to simplify the model and facilitate comprehension.

The nodes D_{Kj} , D_{Cj} , and D_{Nj} , represent the document D_j in distinct contexts. The node D_{Kj} is used to represent the document D_j when it appears as an answer to a keyword-based retrieval process, i.e., the user query is considered as composed of keywords and is processed using the standard cosine formula of the vector model. The node D_{Cj} is used to represent the document

D_j when it appears as an answer to a concept-based retrieval process, i.e., the document must contain a reference to a concept of the thesaurus related to the user query. The node D_{Nj} is used to represent the document D_j when it appears as an answer to a query composed solely of a narrower concept associated with the original user query, i.e., the document must contain a reference to the narrower concept. We model D_{Kj} , D_{Cj} , and D_{Nj} separately in our network to allow evaluating the impact of concept-based retrieval (versus keyword-based retrieval) in the quality of the results. Evidence provided by the set of documents D_{Kj} , D_{Cj} , and D_{Nj} is combined through a disjunctive operator, as done in [12].

With each node in the belief network we associate a binary random variable. Each of these random variables is labeled with the same label of its corresponding node (because it should always be clear whether we are referring to the node or to its associated variable). The variables are all binary (i.e., each variable is *on* or *off*) because this simplifies modeling and provides enough semantics for modeling the problem of IR ranking. Variable degrees of relationship between any pair of related variables are represented by conditional probabilities associated with the edges of the network.

The documents in the set **DK** are ranked according to the standard vector model (i.e., the cosine formula applied to keywords). The documents in the sets **DC** and **DN** are ranked using the cosine formula applied to concepts. These sets of ranked documents represent additional evidence that can be accumulated to yield a new ranking formula (which, we presume, will lead to higher precision).

In the extended belief network, the rank of a document D_j is computed as follows:

$$P(d_j | q) = \eta \sum_{\forall k} P(d_j | k) P(q | k) P(k) \quad (1)$$

Assuming that the unique keywords of interest are the query keywords, we define $P(q|k)$ as:

$$P(q|k) = \begin{cases} 1 & \text{if } \mathbf{KY} = k_q \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where k_q is a state of the variables in **KY** in which the unique nodes that are *on* are those present in the query. This allows rewriting Equation (1) as:

$$P(d_j | q) \sim P(d_j | k_q) P(q | k_q) P(k_q) \quad (3)$$

where we deleted the normalizing constant η , because it has no effect on the ranking.

We observe that $P(d_j|k_q)$ depends on the evidence obtained from the thesaurus. These evidences are used to enrich the network with distinct representations of the original query. For each such representation, a ranking of documents is generated (D_{Cj} and D_{Nj}). These rankings are viewed as distinct sources of evidence on the final relevance of the documents to the original query. To combine these sources of evidence we use a disjunctive operator because it yields better results [12, 13, 16]. Thus, the document node D_j accumulates all the ranking evidence through a disjunction of the beliefs associated with the nodes D_{Kj} , D_{Cj} , and D_{Nj} . This allows rewriting Equation (3) as:

$$P(d_j | q) \sim [1 - P(\bar{d}k_j | k_q) \times P(\bar{d}c_j | k_q) \times P(\bar{d}n_j | k_q)] \times P(q | k_q) \times P(k_q) \quad (4)$$

Evaluating each term of this equation in isolation, for example $P(\bar{d}k_j | k_q)$, we obtain:

$$P(\bar{d}k_j | k_q) = \frac{P(\bar{d}k_j \wedge k_q)}{P(k_q)} \quad (5)$$

As $P(k_q)$ is constant for all documents, define $\alpha = 1/P(k_q)$. Equation (5) can then be rewritten as follows:

$$\begin{aligned} P(\bar{d}k_j | k_q) &= \alpha P(\bar{d}k_j \wedge k_q) = \alpha \sum_{\forall k_i} P((\bar{d}k_j \wedge k_q) | k_i) \times P(k_i) \\ &= \alpha \sum_{\forall k_i} P(\bar{d}k_j | k_i) \times P(k_q | k_i) \times P(k_i) \end{aligned} \quad (6)$$

where k_i is a state of the set of variables \mathbf{KY} . We can further write:

$$\begin{aligned} P(\bar{d}k_j | k_q) &= \alpha \sum_{\forall k_i} P(\bar{d}k_j | k_i) \times P(k_q \wedge k_i) \\ &= \alpha \sum_{\forall k_i} P(\bar{d}k_j | k_i) \times P(k_i | k_q) \times P(k_q) \end{aligned} \quad (7)$$

Assuming that the unique keywords of interest are the query keywords, we define:

$$P(k_i | k_q) = \begin{cases} 1 & \text{if } k_i = k_q \\ 0 & \text{otherwise} \end{cases}$$

which yields:

$$P(\bar{d}k_j | k_q) = 1 - P(dk_j | k_q) \quad (8)$$

where k_q is a state of the variables in \mathbf{KY} where the unique active nodes (i.e., nodes whose associated variables are *on*) are those that correspond to the keywords in the original user query.

The same reasoning can be applied to the other terms of Equation (4), which yields:

$$\begin{aligned} P(d_j | q) &\sim [1 - (1 - P(dk_j | k_q)) \times (1 - P(dc_j | k_c)) \times (1 - P(dn_j | k_n))] \times \\ &P(q | k_q) \times P(k_q) \end{aligned} \quad (9)$$

where k_c and k_n represent states of the sets of random variables \mathbf{CC} and \mathbf{NT} , respectively, in which the only active nodes are those associated with the concepts and narrow terms mapped from the original user query.

We define the prior probabilities $P(k_q)$, associated with the root nodes, as $P(k_q) = (1/2)^t$, where t is the number of terms in the collection, which means that all terms are equally likely.

For $P(q|k_q)$, we write:

$$P(q|k_q) = \begin{cases} 1 & \text{if } \forall_i g_i(q) = g_i(k_q) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $g_i(u)$ is a function that returns the value of the i -th variable in the vector u . Equation (10) establishes that the only state k_q of the set K that is taken into account is the one for which the active keywords are exactly those in the query q .

The probability $P(dk_j | k_q)$ is defined as:

$$P(dk_j | k_q) = \frac{\sum_{i=1}^t w_{ij} \times w_{ik_q}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \times \sqrt{\sum_{i=1}^t w_{ik_q}^2}} \quad (11)$$

where $w_{i\mathbf{k}}$ and w_{ij} are *tf-idf* weights [14], as those used in the vector model. This definition preserves the ordering dictated by a vectorial ranking.

Equation (9) is the ranking formula of our Bayesian model for a juridical digital library. Our ranking formula allows combining evidence in several ways. For instance, consider, for a moment, that we are interested only on the results yielded by the vector model. To obtain this effect, we define $P(dc_j | k_c) = 0$; $P(dn_j | k_n) = 0$. As a result, the ranking $P(d_j|q)$ becomes $P(d_j|q) \sim P(dk_j | k_i) \times P(q | k_q) \times P(k_q)$ which computes a vectorial ranking.

To consider the combination of keyword-based and concept-based retrieval, we define: $P(dn_j | k_n) = 0$. As a result, $P(d_j | q) \sim [1 - (1 - P(dk_j | k_q)) \times (1 - P(dc_j | k_c))] \times P(q | k_q) \times P(k_q)$, which yields a ranking that combines keyword-based and concept-based retrieval. Further, the combination of evidences two by two, can be evaluated by properly defining the related conditional probabilities.

4 Experimental Results

In this section, we briefly introduce the CJF Thesaurus and the reference collection used in our experiments. After we analyze the results in two steps. First we show the results for our extended belief network model. Following we compare these results with the procedure of automatic query expansion.

4.1 The CJF Thesaurus

In this work, we use a manually constructed juridical thesaurus, the CJF Thesaurus, to find concepts related to the original query. This thesaurus was designed and constructed by the Brazilian Federal Justice Council in Brazil (CJF), a regulatory institution that supervises the system of Federal Courts in Brazil and their operations [4]. The motivation was to construct a tool to control the vocabulary used by manual indexers when constructing the Indexing Section⁷ of the juridical document. The CJF Thesaurus comprises many fields of Law, as for example, Criminal Law, Civil Law, Public Law, Commercial Law, Administrative Law, Constitutional Law, International Law, among others.

The CJF Thesaurus has 8,357 juridical concepts (CC) organized in lexicographical order. From these concepts, 6,103 are classified as Narrow Terms (NT), 891 as Broad Terms (BT), 7,301 as Related Terms (RT), and 1,702 as synonyms (SY). Among the synonyms, 674 are classified as preferred and 1,208 as non-preferred. We note that a concept classified as Narrow Term of a given concept may be classified as Broad Term of another concept and as a Related Term of a third one.

4.2 The Juridical Reference Collection

Our evaluation is carried out using a reference collection composed of Brazilian juridical documents, of a set of real queries, and of a set of relevant documents for each of them, determined by a Brazilian lawyer. The collection is composed of 552,573 juridical documents from the Supreme Federal Court (STF) [17], the Superior Court of Justice (STJ) [9], and the 5 Federal Courts of Appeal [8]⁸. The average document size is 164.68 words.

The test queries were selected from a set of real requests made by users of the CJF Internet site [4]. The queries selected, in number of 38, satisfied the following conditions: they had concepts of the CJF Thesaurus associated with them, and these concepts included Narrow Terms relationships⁹. The average number of keywords per query is 8.8.

The relevant document evaluation was done using the well known *pooling* method [18]. The pool of candidate documents was composed of the top-50 ranked documents generated by each of our network rankings. The pool was evaluated by a lawyer, who classified each document as relevant or non-relevant to the query.

4.3 The Results for Our Extended Belief Network Model

We compare the results of a keyword-based (KY) ranking, the baseline of our analysis, with a concept-based (CC) ranking, and with the results generated by a Bayesian combination of them (KY+CC). The results in terms of precision-recall figures are shown Figure 2. The keyword-based ranking yields already a good result, with an average precision of 0.399.

The results yielded by a concept-based (CC) ranking are superior, with an average precision of 0.479. The gain in average precision is of 20.05% and is due to the use of the knowledge encoded in the thesaurus that is related to the test queries.

⁷ After judgment, the juridical document is enriched by a section called Indexing Section created with the purpose to improve information retrieval.

⁸ URL's for the other four Federal Court sites can be obtained from [4].

⁹ These decisions were made in order to verify the thesaurus contribution.

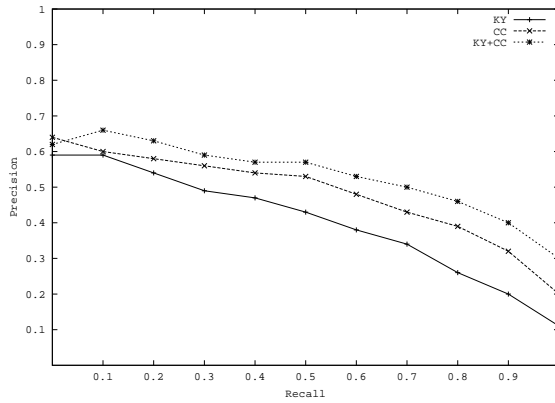


Fig. 2. Results of Bayesian Combination with Keywords and Concepts.

Finally, we observe that the result generated by a Bayesian combination of keyword-based and concept-based evidences (KY+CC) exceeds the results generated by each of them in isolation. The average precision is 0.529. The gain in average precision is now 32.58%, relative to the keyword-based ranking. This indicates the usefulness of our approach to improve the quality of the results.

4.4 Comparing with Automatic Query Expansion

To illustrate the differences between our extended belief network model and the more common procedure of automatic query expansion (using concepts of the CJF Thesaurus related to query), we consider the usage of Narrow Terms (NT). In the case of the network model, this is done by modifying the ranking to consider Narrow Terms, as described in Section 3. In the case of automatic query expansion, this is done by expanding the original query with Narrow Terms related to it.

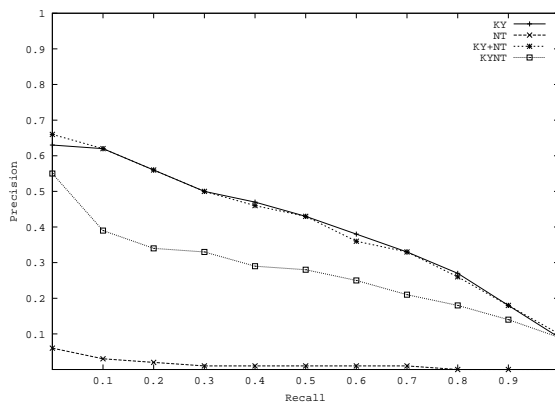


Fig. 3. Comparison of Bayesian Combination and Automatic Query Expansion.

Figure 3 illustrate the results. We first observe that Narrow Terms (NT) by themselves yield very poor results. The reason is that the test queries are long (8.8 keywords per query) and have many Narrow Terms related to them (29.5 terms per query generated by 6.1 concepts per query). This generates excessive noise that degrades the results. Despite this negative effect, the extended belief network model is able to filter the noisy NT terms and generate a combined

ranking (KY+NT) that is very similar to the ranking produced by keywords only (KY). The ranking generated by automatic query expansion with NT (KYNT), however, suffers the negative impact of the noisy NT terms and yields poor results (relative to our baseline).

5 Conclusions

In this paper, we discussed the design and evaluation of a vertical searching algorithm for juridical digital libraries. Our vertical algorithm allowed combining specialized information obtained from a juridical thesaurus, the CJF Thesaurus, with evidence generated by the classic vector space model. To allow implementing this combination in consistent fashion we relied on Bayesian belief networks.

Our extended belief network model for juridical digital libraries was evaluated using a juridical reference collection of 552,573 documents and 38 real test queries. Our results showed good improvements in retrieval performance. Further, they showed that our model is distinct from the standard technique of automatic query expansion. Such results suggest that the study and development of vertical searching mechanisms is a promising research direction.

References

1. Alan Gilchrist. *The Thesaurus in Retrieval*. Aslib, London, 1971.
2. Jane Greenberg. Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology*, 52(5):402–415, 2001.
3. Jane Greenberg. Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. *Journal of the American Society for Information Science and Technology*, 52(6):487–498, 2001.
4. Federal Justice Council (CJF) in Brazil. <http://www.cjf.gov.br>, 2002.
5. Jaana Kristensen. Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management*, 29(6):733–744, 1993.
6. Frederick Wilfrid Lancaster. *Indexing and Abstracting in Theory and Practice*. Library Association Publishing, London, 1991.
7. Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. Query expansion using heterogeneous thesauri. *Information Processing & Management*, 36(2000):361–378, 2000.
8. Federal Court of Appeals First Region (TRF1). <http://www.trf1.gov.br>, 2002.
9. Superior Court of Justice (STJ). <http://www.stj.gov.br>, 2002.
10. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Los Altos, California, 1988.
11. Yonggang Qiu and H. P. Frei. Concept based query expansion. In *Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 160–169, 1993.
12. Berthier Ribeiro-Neto and Richard R. Muntz. A belief network model for IR. In *ACM Conference on Research and Development in Information Retrieval - SIGIR96*, pages 253–260, 1996.
13. Berthier Ribeiro-Neto, Ilmério Silva, and Richard Muntz. Bayesian network models for IR. In Fabio Crestani and Gabriella Pasi, editors, *Soft Computing in Information Retrieval Techniques and Applications*, pages 259–291. Springer Verlag, 2000.
14. G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983.
15. Ali Asghar Shiri and Crawford Revie. Thesauri on the web: current developments and trends. *Online Information Review*, 24(4):273–279, 2000.
16. Ilmério Silva, Berthier Ribeiro-Neto, Pável Calado, Edleno Moura, and Nivio Ziviani. Link-based and content-based evidential information in a belief network model. In *ACM Conference on Research and Development in Information Retrieval - SIGIR2000*, pages 96–103, 2000. Best Student paper.
17. Supreme Federal Court (STF). <http://www.stf.gov.br>, 2002.
18. Ellen M. Voorhees and Donna Harman. Overview of the fifth Text REtrieval Conference. In *Proceedings of the fifth Text REtrieval Conference (TREC-5)*, pages 1–28. National Institute of Standards and Technology, Gaithersburg, MD 20899, 1996.