Local Versus Global Link Information in the Web

PÁVEL CALADO, BERTHIER RIBEIRO-NETO, NIVIO ZIVIANI
Federal University of Minas Gerais
EDLENO MOURA
Akwan Information Technologies
and
ILMÉRIO SILVA
Federal University of Uberlândia

Information derived from the cross-references among the documents in a hyperlinked environment, usually referred to as link information, is considered important since it can be used to effectively improve document retrieval. Depending on the retrieval strategy, link information can be local or global. Local link information is derived from the set of documents returned as answers to the current user query. Global link information is derived from all the documents in the collection. In this work, we investigate how the use of local link information compares to the use of global link information. For the comparison, we run a series of experiments using a large document collection extracted from the Web. For our reference collection, the results indicate that the use of local link information improves precision by 74%. When global link information is used, precision improves by 35%. However, when only the first 10 documents in the ranking are considered, the average gain in precision obtained with the use of global link information is higher than the gain obtained with the use of local link information. This is an interesting result since it provides insight and justification for the use of global link information in major Web search engines, where users are mostly interested in the first 10 answers. Furthermore, global information can be computed in the background, which allows speeding up query processing.

Categories and Subject Descriptors: H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms: Theory, Experimentation

Additional Key Words and Phrases: Belief networks, link analysis, local and global information, World Wide Web

This work has been partially supported by SIAM project, grant MCT/FINEP/CNPq/PRONEX 76.97.1016.00, MCT/FCT scholarship SFRH/BD/4662/2001 (P. Calado), CNPq grant 300.188/95-1 (B. Ribeiro-Neto), and CNPq grant 520.916/94-8 (N. Ziviani).
Authors' address: P. Calado, B. Ribeiro-Neto, N. Ziviani, Department of Computer Science, Federal University of Minas Gerais, Av. Antonio Carlos 6627, 31270-010 Belo Horizonte, MG, Brazil; email: {pavel,berthier,nivio}@dcc.ufmg.br, edleno@akwan.com.br, ilmerio@ufu.br.

1. INTRODUCTION

Ranking based on link information consists of using the knowledge derived from the hypertextual structure of the Web to determine the relevance of the documents in the query answer set [Brin and Page 1998; Kleinberg 1998]. In this approach, the rank of a document is determined in terms of the number of other documents that contain references to it. A document that is referenced by many others is given a higher score than one that has little or no references. The use of link information has become important, since it allows improving document retrieval, as demonstrated by major Web search engines.

In this work, we compare the use of *local link information* versus the use of *global link information*. As proposed in Xu and Croft [1996], local information refers to the information in the set of documents returned as answers to a user query, whereas global information refers to information extracted from all the documents in the collection. Here we present a study on the use of local and global information extracted from the links between documents. We show that local link information can provide a significant improvement in precision figures and that global link information can provide a great increase in precision at low recall levels.

Using the framework of Bayesian belief networks, we study and compare various alternatives for combining local and global link information with information extracted from the contents of the documents. Following, we explicitly compare the results generated by the combined rankings with those yielded by a vectorial ranking [Salton and McGill 1983]. For this, we use a reference collection of about six million documents extracted from the World Wide Web. Our results indicate that this type of combination can improve retrieval performance considerably without requiring any extra information from the users at query time. Although the use of local information has led to better results on average, the use of global link information has yielded higher precision values for the top 10 ranked documents. Since global information is computed only once for the whole collection, it is preferable for use with Web search engines, for which high precision is important for the top-ranked documents and a short querying time is essential.

This work is organized as follows. Section 2 presents a review of several papers related to the subject of our proposal. Section 3 describes an algorithm to rank pages according to the hyperlink structure surrounding the documents in the collection and discusses the use of global versus local link information. Section 4 reviews the belief network model and discusses how to combine different types of evidence. In Section 5 we present experimental results on the use of different combinations of local and global link information. Finally, Section 6 presents some concluding remarks and suggestions for future work on the subject.

2. RELATED WORK

Cross-referencing information has been used in bibliometric science to measure document importance [Garfield 1972], to find related documents [Kessler

1963; Small 1973; Small and Koenig 1977], and to combine citation evidence with keyword evidence to improve document retrieval [Salton 1971]. In a hyperlinked environment, links among the documents can take the role of cross-references. This assumption has recently been used to rank documents in Web information systems.

In a seminal work, Brin and Page [1998] propose an algorithm, named PageRank, that uses the Web link structure to derive a measure of popularity for Web pages. A page is recursively defined as popular when it is linked by many other popular pages. Kleinberg [1998] later proposes the HITS algorithm, where pages assume two distinct functions: hub and authority. A good hub page links to many good authority pages. Recursively, a good authority page is linked by many good hub pages. In SALSA [Lempel and Moran 2000, 2001], the degrees of hub and authority for Web pages are computed by examining random walks through the Web graph, an approach that avoids some problems inherent to HITS. Following a different perspective, Cohn and Chang [2000] propose a a model that is able to group documents in categories and determine its authoritative degree within each category by performing factor analysis in the set of links and documents.

Kleinberg's [1998] proposal was restricted to documents in the answer set to a user query, together with documents that link to or are linked by them. On the other hand, in Brin and Page [1998], the proposed algorithm was applied to all the documents in the collection. We say that the first approach was based on *local information*, whereas the second was based on *global information*. Each approach has its own advantages and disadvantages and it is therefore important to evaluate where each of them can be applied. Comparisons on the use of local and global information have been performed in Xu and Croft [2000, 1996] and Salton et al. [1993], using only the textual content of documents. In this work we compare the use of global link information with the use of local link information.

Previous work has proposed combining link-based with content-based pieces of evidence in a single information retrieval model. The algorithm in Chakrabarti et al. [1998] proposes to combine the local link analysis described in Kleinberg [1998] with keyword-based evidence. The basic idea is to use the text surrounding the links to determine a weight for each link analyzed. The work in Bharat and Henzinger [1998] presents similar algorithms, but expanding the original user query with keywords obtained from the documents in the local answer set. The weight of each link is then computed based on the expanded query. This expansion process improves retrieval performance but can be somewhat expensive, since it greatly increases the number of terms to be processed.

In a different approach, Westerveld et al. [2001] and Gao et al. [2001] combine link- and content-based document rankings by using linear interpolation. In Kanungo and Zien [2001], the documents' link-based ranking is inserted into the *tf-idf* weights [Salton and McGill 1983] of the terms in the documents. These works have reported significant results for the task of finding site homepages. However, for document ranking, the improvements obtained with the use of link analysis were only marginal. In a manner more similar to our own, Dumais

and Jin [2001] propose a probabilistic model capable of combining link and content information. However, their model works by iteratively propagating the information through the link structure, whereas our model is based on a closed probabilistic formula for evidence combination. This can have obvious implications in efficiency.

Our work differs from previous studies in several directions. First, we adopt Bayesian networks [Pearl 1988] as a unifying modeling framework for the combination of link- and content-based evidence. Bayesian networks were first proposed as a modeling tool for information retrieval problems in Turtle and Croft [1990, 1991] and later used in Ribeiro-Neto and Muntz [1996] to model evidence derived from past queries and combine it with the vector space model [Salton and McGill 1983]. Second, we combine information on both authorities and hubs with content-based evidential information. This combination gives better retrieval results than strategies that use either authorities or hubs in isolation. Third, we explicitly compare the use of local link information with the use of global link information and analyze the advantages of each alternative. Both local and global link information is obtained through the HITS algorithm. This allows us to examine the behavior of HITS on the whole collection of documents, and provides both hub and authority global values, which would not be available through the use of PageRank. Finally, a direct comparison with PageRank is also provided.

3. COMPUTING LINK INFORMATION

One of the richest sources of information in a hyperlinked environment is the knowledge of its link structure. Such knowledge frequently encodes human judgment about the documents, which can be of critical importance in the generation of a good ranking. The HITS algorithm [Kleinberg 1998] uses this information to measure the importance of a document based on two metrics: a degree of *authority* and a degree of *hubness*. A good *authority* is defined as a document with a high number of incoming links from good *hubs*. Recursively, a good *hub* is defined as a document with a high number of outgoing links that point to good *authorities*.

The HITS algorithm computes a degree of goodness for hubs and authorities based on an analysis of the link structure surrounding the documents in the local answer set to a user query. In HITS, this set is called the *root set* of documents. This set is then expanded with its neighboring documents (i.e., documents that link to, or are linked by, the documents in the root set) thus forming the *base set* of documents. In the context of this work, we can say that both the root and base sets of documents are local. For this reason, we say that the algorithm computes a *local authority* value and a *local hub* value for each document. We can also apply the algorithm to the whole set of documents in the collection. In this case, we say that the algorithm computes a *global authority* value and a *global hub* value for each document.

The HITS algorithm interprets a collection of hyperlinked documents as a directed graph $\mathcal{G}$, where each document (page) is represented by a node and each link between two documents is represented by a directed edge. By assumption,

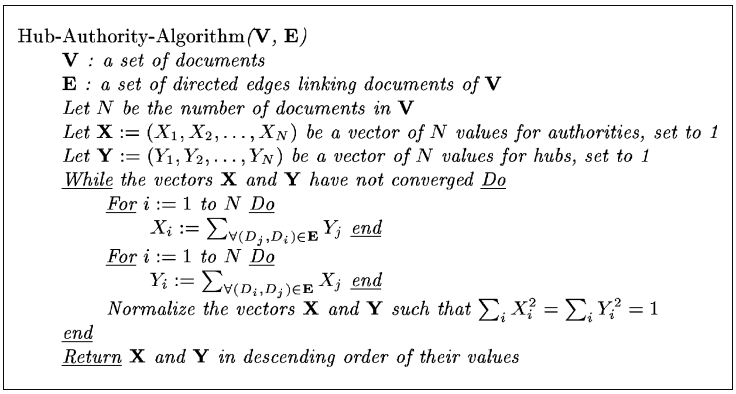

Fig. 1. Algorithm for computing the authority and the hub values of each node (or document).

a link from a document D to another document D' implies that the author of the document D endorses document D'.

Consider a collection of hyperlinked documents and its associated directed graph $\mathcal{G}$. Given a user query Q, the *local hub* and *local authority* values of each document can be computed using the link structure associated with the documents in the local answer set (the root set). To ensure that there are enough links among the documents, we form the base set by adding neighboring documents. Thus let $\mathcal{Q} = (\mathbf{V}, \mathbf{E})$ be a subgraph of $\mathcal{G}$ such that each node of $\mathbf{V}$ represents a document related to the query Q (a document in the base set) and the set of edges $\mathbf{E}$ represents a set of links related to the documents in $\mathbf{V}$. The *local hub* and *local authority* values of each document in $\mathbf{V}$ are computed by the algorithm presented in Figure 1.

The algorithm presented in Figure 1 can also be applied to the graph $\mathcal{G}$ representing all the documents and links in the full collection of hyperlinked documents, instead of a subgraph derived from the documents related to a topic of interest. In this case, we say that the algorithm uses *global information* and that the hub and authority values computed quantify degrees of *global hub* and of *global authority*. For each document, these values can then be combined with the corresponding vector rank to yield a new ordering.

Global information can also be obtained by the PageRank algorithm [Brin and Page 1998]. The main idea behind PageRank is the simulation of a random surfer on the Web. The surfer moves through the Web by randomly choosing a link from a page and eventually jumping to a random page. The PageRank of a document is the probability that the surfer has visited it during his random walk. The reader is directed to Page et al. [1998] for a more detailed description. In Section 5.3, we show a comparison of PageRank with the HITS algorithm, using global information. In this article, however, we focus on the HITS algorithm as a source of global link evidence, since its performance is quite similar to that of PageRank and it provides us with two distinct sources of evidence: hub and authority degrees. As shown in Section 5, both sources are useful in improving ranking results.

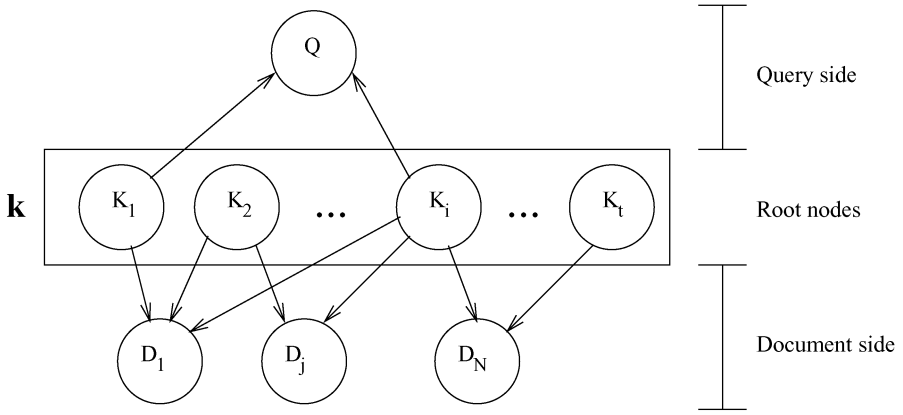

Fig. 2. Belief network for a query Q composed of the keywords K_1 and K_i.

4. THE BELIEF NETWORK MODEL

This section shows how to model content- and link-based evidential information using Bayesian networks. For this task we adopt the belief network model introduced in Ribeiro-Neto and Muntz [1996]. This model takes an epistemological view, as opposed to a frequentist view, of the information retrieval problem and interprets probabilities as degrees of belief devoid of experimentation, as also done in Turtle and Croft [1991] and Wong and Yao [1995].

The belief network model adopts Bayesian networks as its basic foundation. Bayesian networks are useful because they provide a graphical formalism for explicitly representing independencies among the variables of a joint probability distribution. The probability distribution is represented through a directed acyclic graph, whose nodes represent the random variables of the distribution. The relationships among these variables are modeled as directed edges that represent causal dependencies among the linked variables. The strengths of these dependencies are expressed by conditional probabilities. The fundamental principle is that the known independencies among the random variables of a domain are declared explicitly and that a joint probability distribution is synthesized from this set of declared independencies.

4.1 Modeling Content-Based Evidence on a Belief Network

In a traditional content-based information retrieval system, the documents and the user queries are usually represented as sets of keywords. As a result, queries and documents are treated analogously, as proposed in Ribeiro-Neto and Muntz [1996]. Figure 2 illustrates a belief network that reflects this symmetry. Following the Bayesian theory presented in Pearl [1988], the instantiation of the root nodes *separates* the document nodes from the query node, making them mutually independent. We say that the query is on the *query side* of the network, and the documents are on the *document side* of the network.

In the network of Figure 2, each node D_j models a document, the node Q models the user query, and the K_i nodes model the keywords in the collection. The vector $\mathbf{k}$ is used to refer to any of the possible states of the *root* nodes K_i.

A *binary* random variable is associated with the node Q, which is also denoted by Q. In this notation it should always be clear whether we are referring to the query, to the node in the network, or to its associated binary variable. The variable Q is 1, denoted by q, to indicate that Q is *on* and Q is 0, denoted by $\overline{q}$, to indicate that Q is *off*. Analogously, a *binary* random variable D_j is associated with the document node D_j. The variable D_j is 1, denoted by d_j, to indicate that D_j is *on* and D_j is 0, denoted by $\overline{d}_j$, to indicate that the variable D_j is *off*. A *binary* random variable K_i is also associated with each keyword K_i. All of these variables are binary since they provide enough semantics for modeling the information retrieval problem. Varying degrees of relevance are represented in the model as conditional probabilities, as we discuss in the immediate following.

In the network of Figure 2, the ranking computation is based on quantifying the similarity between a document D_j and the query Q by the probability $P(d_j|q)$, that is, the probability that the variable D_j is *on* given that the variable Q is *on*. By the rule of total probabilities and the independencies modeled in the network we can write:

$$P(d_j|q) = \eta \sum_{\mathbf{k}} P(d_j|\mathbf{k})\, P(q|\mathbf{k})\, P(\mathbf{k}), \tag{1}$$

where η is a normalizing constant [Pearl 1988]. In our belief network model, Equation (1) represents the generic expression for computing the rank of a document D_j with regard to the query Q.

Modeling the Vector Space Model. Equation (1) can be used to represent any of the classic models in IR, namely, the Boolean, the vector, and the probabilistic models, as demonstrated in Ribeiro-Neto et al. [2001]. Here we review how to use a belief network to compute a ranking generated by the vector space model [Salton and McGill 1983]. Since this is a very popular keyword-based model for IR and one that can be computed efficiently, it was chosen as the reference of the content-based IR model in our study (i.e., the vector space model is our baseline).

To compute a vectorial ranking in our belief network we specify the probabilities $P(\mathbf{k})$, $P(q|\mathbf{k})$, and $P(d_j|\mathbf{k})$. First we define the prior probabilities $P(\mathbf{k})$ associated with the root nodes as follows.

$$P(\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i\ g_i(\mathbf{q}) = g_i(\mathbf{k}) \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $g_i(\mathbf{u})$ is a function that returns the value of the ith variable in the vector $\mathbf{u}$. Equation (2) establishes that the only state $\mathbf{k}$ of the set $\mathbf{K}$ of root nodes that is taken into account is the one for which the active keywords are exactly those in the query Q. For $P(q|\mathbf{k})$, we write:

$$P(q|\mathbf{k}) = \begin{cases} 1 & \text{if } \forall_i\ g_i(\mathbf{q}) = g_i(\mathbf{k}) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and for $P(d_j|\mathbf{k})$, we write:

$$P(d_j|\mathbf{k}) = \frac{\sum_{i=1}^{t} w_{ij} \cdot w_{i\mathbf{k}}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}\sqrt{\sum_{i=1}^{t} w_{i\mathbf{k}}^2}}, \tag{4}$$

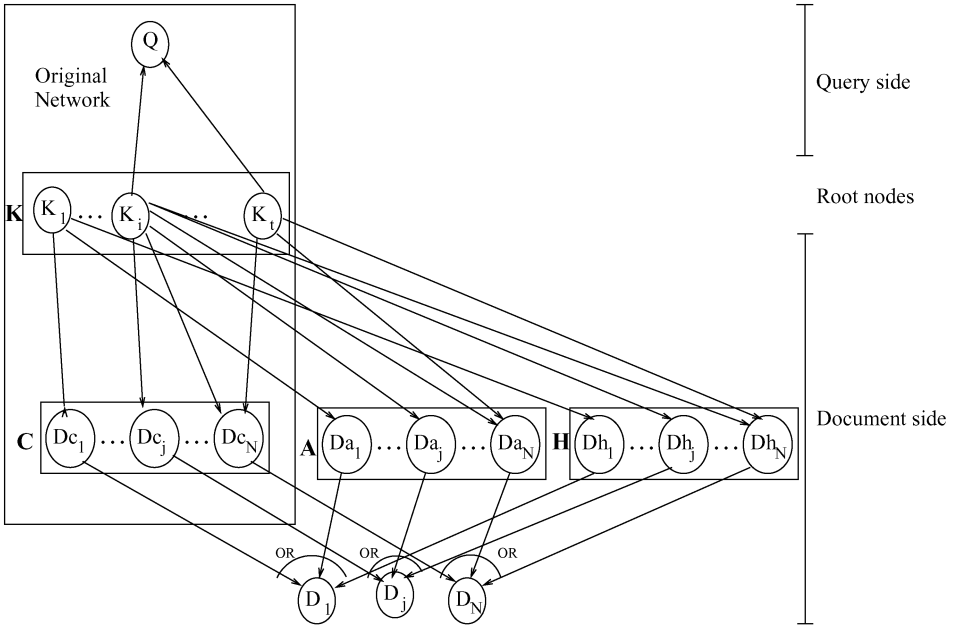

Fig. 3. Belief network expanded with link-based evidence.

where $w_{i\mathbf{k}}$ and w_{ij} are *tf-idf* weights [Salton and McGill 1983] used in the vector model. We define $w_{ij} = TF_{ij} \times \log(1/DF_i)$, where TF_{ij} is the number of times term i appears in document D_j and DF_i is the number of documents where term i appears. Similarly, we define $w_{i\mathbf{k}} = 1 \times \log(1/DF_i)$. This specification is valid and consistent because $P(d_j|\mathbf{k})$ measures the cosine of the angle between two vectors, which is a number between 0 and 1. By substituting Equations (2) through (4) into Equation (1), we obtain a ranking for the D_j documents, expressed as $P(d_j|q)$, which preserves the ordering dictated by a vectorial ranking.

4.2 Modeling Link-Based Evidence on a Belief Network

We expand the belief network model discussed above to also include evidential information extracted from the link structure of the environment, as proposed in Silva et al. [2000]. This is accomplished by adding new edges, nodes, and probabilities to the original network presented in Figure 2, as illustrated in Figure 3. We say that this expansion is modular in the sense that it preserves all the properties of the previous network. This strategy allows us to combine the keyword-based evidence associated with the documents content with the link-based evidence obtained from the surrounding hypertextual environment, in a natural and convenient way.

In the belief network of Figure 3, the left-hand side of the network represents the original network of Figure 2 with the adaptations: each document node D_j is renamed Dc_j, for content-based. The right-hand side of the network models the link-based sources of evidence. These can be obtained either from the link

structure associated with the set of documents in the answer set to a query, thus being sources of *local evidence*, or from the link structure associated with the whole collection of documents, thus being sources of *global evidence*. In Silva et al. [2000] only local link information was considered. Here we consider both local and global link information and directly compare them.

To represent link-based evidential knowledge in the network, we associate two new nodes Dh_j and Da_j with each document D_j in the answer set for query Q. We associate a binary random variable Dh_j with the node Dh_j to model evidence associated with the document D_j as a hub. This evidence is computed from the link structure associated with the local or global set of documents, as presented in Section 3. Hub values are represented in our network as the conditional probability of Dh_j being *on* given the keywords in the query Q and given an implicit knowledge of the surrounding link structure. Analogously, we associate a binary random variable Da_j with the node Da_j to model evidence associated with the document D_j as an authority. Thus we now have three sets of nodes representing evidential knowledge associated with the documents in the network: the set **H**, composed of nodes representing hub evidence, the set **A**, composed of nodes representing authority evidence, and the set **C**, composed of nodes representing content-based evidence. The state of the associated random variables is given by **h**, **a**, and **c**, respectively.

The set of nodes **K** is used to model the occurrence of keywords in the query Q and, once instantiated, induces beliefs on each of the nodes in the sets **C**, **H**, and **A**. The propagation of these beliefs in the network is done according to the conditional probabilities governing the relationships between the set **K** and each of the sets **C**, **H**, and **A**. The specification of the conditional probabilities is based on the vector space model and on the HITS algorithm, as we later discuss.

The binary random variable Dh_j associated with each node Dh_j of **H** is 1 to indicate that the hub evidence associated with the document D_j is to be considered in the ranking computation. Also, the binary random variable Da_j associated with each node Da_j of **A** is 1 to indicate that the authority evidence associated with the document D_j is to be considered in the ranking computation. The node D_j represents the combination of content-based and link-based evidential knowledge from the left- and right-hand sides of the network. The conditional probabilities, discussed below, define how these evidences are combined.

4.3 General Equation for Ranking Computation

In Figure 3, the rank $P(d_j|q)$ associated with a document D_j can be computed using Equation (1). However, the conditional probability $P(d_j|\mathbf{k})$ now depends on link- and content-based pieces of evidence, combined through a disjunctive operator *or*. This is accomplished as follows.

$$P(d_j|\mathbf{k}) = 1 - (1 - P(dc_j|\mathbf{k})) \times (1 - P(dh_j|\mathbf{k})) \times (1 - P(da_j|\mathbf{k})). \quad (5)$$

Substituting Equation (5) into Equation (1), we can write:

$$P(d_j|q) = \eta \sum_{\mathbf{k}} [1 - (1 - P(dc_j|\mathbf{k})) \times (1 - P(dh_j|\mathbf{k})) \times (1 - P(da_j|\mathbf{k}))] \times P(q|\mathbf{k}) \times P(\mathbf{k}). \quad (6)$$

The computation of the probability $P(d_j|\mathbf{k})$ depends on the states of the nodes Dc_j, Da_j, and Dh_j. The probability $P(q|\mathbf{k})$ can be computed through the proper specification of the states of the root nodes K_i, establishing interesting alternatives for computing the rank of a document D_j with regard to a query Q, as we now discuss.

4.4 Ranking Computation

The belief network model can represent the vector model through proper specification of the conditional probabilities in the network, as discussed in Section 4.1. To simplify our notation, let R_{jq} be a reference to the vectorial score of the document D_j with regard to a query Q, computed according to our network model using Equation (4). Thus:

$$R_{jq} = \frac{\sum_{i=1}^{t} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \sqrt{\sum_{i=1}^{t} w_{iq}^2}}. \tag{7}$$

Furthermore, let H_{jq} and A_{jq} be the hub and authority values, respectively, associated with document D_j, computed by Kleinberg's algorithm presented in Section 3. These values are computed using either local or global analysis.

Case 1: *Content-Based Ranking.* For representing a ranking based solely on document content, we ignore the knowledge derived from the local link structure. This is accomplished in our network model by defining:

$$P(dc_j|\mathbf{k}) = R_{jq}; \qquad P(dh_j|\mathbf{k}) = 0; \qquad P(da_j|\mathbf{k}) = 0. \tag{8}$$

Applying Equations (2), (3), (4), and (8) into Equation (6), we obtain:

$$P(d_j|q) = \eta \times R_{jq}. \tag{9}$$

Therefore, the general network of Figure 3 naturally subsumes a ranking dictated by the vector space model.

Case 2: *Ranking Based on Hub Evidential Knowledge.* To represent a ranking that depends only on hub-based knowledge, we redefine the probabilities as follows.

$$P(dc_j|\mathbf{k}) = 0; \qquad P(dh_j|\mathbf{k}) = H_{jq}; \qquad P(da_j|\mathbf{k}) = 0, \tag{10}$$

which allows ignoring information associated with content- and authority-based evidence. Notice that the hub evidence associated with D_j is modeled as the conditional probability $P(dh_j|k)$, whose value is set to H_{jq}, the hub value of the document D_j with regard to the query Q.

Applying Equations (2), (3), and (10) to Equation (6), we obtain

$$P(d_j|q) = \eta \times H_{jq}. \tag{11}$$

In this case, our network simply reproduces a ranking based on hub values.

It is important to note that the use of local or global information depends only on the way H_{jq} is computed by the hub-authority algorithm presented in Figure 1. Local information is obtained from the local set of documents, whereas global information is obtained from the global set of documents.

Table I. Alternative Rankings Modeled in Our Belief Network Model

Case	Ranking	$P(d_j \mid q)$
1	Vector	$\eta \times R_{jq}$
2	Hub	$\eta \times H_{jq}$
3	Authority	$\eta \times A_{jq}$
4	Vector-Hub	$\eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq})]$
5	Vector-Authority	$\eta \times [1 - (1 - R_{jq}) \times (1 - A_{jq})]$
6	Vector-Hub-Authority	$\eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq}) \times (1 - A_{jq})]$

Case 3: *Ranking Based on Authority Evidential Knowledge.* In this case, we write:

$$P(dc_j \mid \mathbf{k}) = 0; \qquad P(dh_j \mid \mathbf{k}) = 0; \qquad P(da_j \mid \mathbf{k}) = A_{jq}, \tag{12}$$

where A_{jq} is the authority value of the document D_j with regard to the query Q.

Applying Equations (2), (3), and (12) to Equation (6) we obtain

$$P(d_j \mid q) = \eta \times A_{jq}. \tag{13}$$

As a result, our network simply reproduces a ranking based on authority values. As in Case 2, local information is obtained from the local set of documents, and global information is obtained from the global set of documents.

Case 4: *Combining Content-Based and Hub-Based Pieces of Evidence.* We now discuss how our network model can be used to naturally combine keyword-based evidential knowledge with link-based evidential knowledge.

Making $P(dc_j \mid \mathbf{k}) = R_{jq}$, $P(dh_j \mid \mathbf{k}) = H_{jq}$, and $P(da_j \mid \mathbf{k}) = 0$ and applying Equations (2) and (3) to Equation (6), we obtain:

$$P(d_j \mid q) = \eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq})]. \tag{14}$$

Case 5: *Combining Content-Based and Authority-Based Pieces of Evidence.* Making $P(dc_j \mid \mathbf{k}) = R_{jq}$, $P(dh_j \mid \mathbf{k}) = 0$, and $P(da_j \mid \mathbf{k}) = A_{jq}$ and applying Equations (2) and (3) to Equation (6), we obtain:

$$P(d_j \mid q) = \eta \times [1 - (1 - R_{jq}) \times (1 - A_{jq})]. \tag{15}$$

Case 6: *Combining Content-Based, Hub-Based, and Authority-Based Pieces of Evidence.* Making $P(dc_j \mid \mathbf{k}) = R_{jq}$, $P(dh_j \mid \mathbf{k}) = H_{jq}$, and $P(da_j \mid \mathbf{k}) = A_{jq}$ and applying Equations (2) and (3) to Equation (6), we obtain:

$$P(d_j \mid q) = \eta \times [1 - (1 - R_{jq}) \times (1 - H_{jq}) \times (1 - A_{jq})]. \tag{16}$$

Summary of Ranking Alternatives. Table I summarizes the six alternative rankings modeled in our network. In all cases, the values H_{jq} and A_{jq} might be derived either from local or global information. Notice that we do not consider the combination of only authority-based and hub-based pieces of evidence. Experimental results indicated that this combination, without the use of content-based evidential information, is not promising.

Table II. Characteristics of the Database

Number of Pages	Number of Distinct Words	Number of Words per Page	Number of Queries	Average Number of Words per Query	Average Number of Pages per Query Pool	Average Number of Relevant Pages per Query Pool
5,939,061	2,669,965	413.5	50	1.78	83.26	36

It is interesting to note that, using our Bayesian Network model, a weighted combination of evidence can naturally be done. A Noisy-OR model [Pearl 1988] could also be used as an alternative, instead of a standard disjunction. The noise parameters would serve as weights for each evidence. Weight adjusting can be used to refine the model, in order to increase the quality of the results, or to adapt it to the needs of specific users. In this article, however, we are interested only in comparing the use of local and global information and, therefore, leave this type of refinement for future work.

5. EXPERIMENTAL RESULTS

In this section we evaluate each of the six ranking alternatives presented in Table I, considering the adoption of either global or local hub and authority values. Our experiments are based on a collection of documents extracted from the World Wide Web.

5.1 The Reference Collection

Our reference collection is composed of a database of Web pages, a set of example Web queries, and a set of relevant documents associated with each example query. The database is composed of 5,939,061 pages of the Brazilian Web, under the domain ".br". The pages were automatically collected by the document collector described in Silva et al. [1999], and indexed using inverted lists [Witten et al. 1999].

A total of 50 example queries were selected from a log of 100,000 queries submitted to the *TodoBR* search engine (http://www.todobr.com.br). The queries selected were the 50 most frequent ones. Some frequent queries related to sex were not considered. The mean number of keywords per query is 1.78. Of the selected queries, 28 were quite general, such as "tabs," "movies," or "mp3." Following, 14 queries were more specific, but still on a general topic, such as "transgenic food," or "electronic commerce." Finally, 8 queries were quite specific, consisting mainly of music band names. Similarly to Hawking and Craswell [2001], for all queries, a description of what documents should be considered relevant was given to the users. For instance, for query "employment," users were instructed to consider relevant only the sites dedicated to employment ads.

For each of our 50 example queries, we composed a query pool formed by the top 20 documents generated by each of our 6 types of network ranking. Hub and authority values were computed using either global or local information, which yielded a total of 11 different ranking strategies (the vectorial ranking is unaffected by the type of information considered). The characteristics of the database used are summarized in Table II. Each query pool contained an

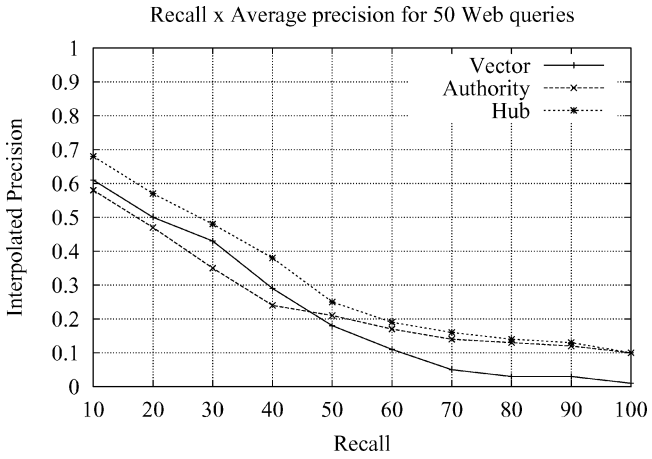

Fig. 4. Precision figures for vector, authority, and hub rankings, using local link information.

average of 83.26 pages. All documents in each query pool were submitted to a manual evaluation by a group of 29 users, all of them familiar with Web searching. Users were allowed to follow links and evaluated the pages according not only to their textual content, but also to their linked pages and graphical content. The average number of relevant pages per query pool was 36. We adopted the same pooling method used for the Web-based collection of TREC [Hawking et al. 1998, 1999].

In our experiments, the local answer sets used consisted of the first 200 documents obtained with the vectorial ranking. For the local HITS algorithm, this set was expanded with its neighboring documents, as explained in Section 3. Also, to avoid self-reinforcement problems, that is, a page getting high hub and authority values because it is linked with many pages within the same site [Lempel and Moran 2000; Bharat and Henzinger 1998], only links to pages belonging to outside sites were considered.

5.2 Ranking Using Local Link Information

In this section we compare and evaluate the six alternative rankings in Table I, considering only local link information. All comparisons were made in terms of precision-recall figures [Baeza-Yates and Ribeiro-Neto 1999], averaged over the 50 test queries. Note that the recall values are relative to the set of evaluated documents, since we are not able to evaluate the entire collection.

Figure 4 illustrates the retrieval performance for the vector, hub, and authority rankings presented in Section 4.4. We observe that the hub ranking is superior for our set of queries. This happens because good hubs are generally pages with a great number of links to other pages covering a particular subject, such as pages from Web directories. Our test users considered these pages relevant when searching for information on the subject. Thus almost all pages highly ranked as hubs were taken as relevant pages. We also observe that the vector and authority rankings both have good precision values. This indicates

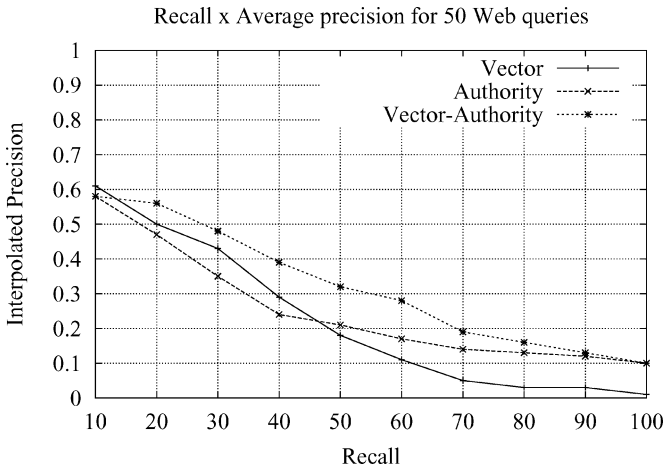

Fig. 5. Precision figures for vector, authority, and vector-authority network rankings, using local link information.

that they should not be disregarded as important sources of evidence when ranking Web pages.

In Figure 5 we investigate the impact of combining the vector and authority rankings in our belief network model. The results indicate that this combination yields precision figures that are superior to those provided by each ranking in isolation. At low recall levels, the vector-authority ranking shows a small decrease in precision when compared to the vector ranking. This happens because some pages, although pointed by many others (thus being good authorities), are unrelated to the query topic and therefore not relevant to the users. Nevertheless, the authority ranking contributes to improve the overall precision for all recall levels above 20%, where the vector ranking is not as good as it is at lower recall levels.

Figure 6 shows the impact of combining the vector and hub rankings in our belief network model. Again we observe that this combination yields higher precision figures than those obtained by each ranking in isolation.

Finally, in Figure 7, we show the impact of combining the vector, authority, and hub rankings. This three-way combination of evidence yields superior results. At recall levels below 30%, the vector-hub combination has slightly better performance, due to high relevance given by users to pages with many links. At middle and high recall levels the vector-hub-authority combination shows a large improvement over the remaining rankings, showing that both hub and authority values are useful for determining document ranks.

These results confirm the preliminary results presented in Silva et al. [2000], but now considering a much larger test collection. They demonstrate that the belief network model is able to take advantage of the distinct nature of each of our three types of evidential knowledge to provide improved overall retrieval performance. This is an interesting result that indicates the strength of belief networks as a framework for consistently combining distinct pieces of evidence on support of a relevance ranking, a characteristic also observed

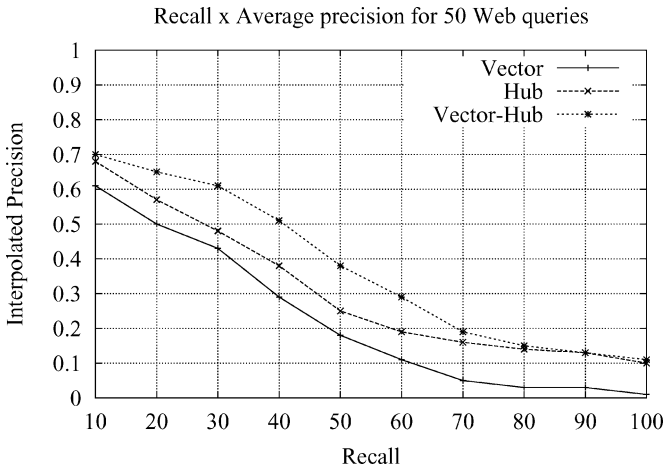

Fig. 6. Precision figures for vector, hub, and vector-hub network rankings, using local link information.

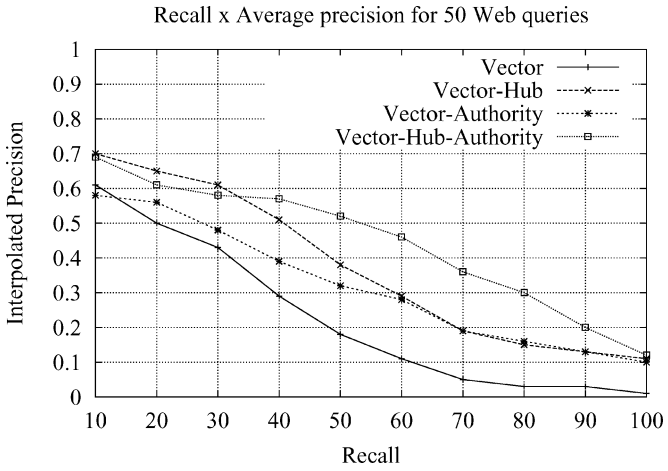

Fig. 7. Precision figures for vector, vector-hub, vector-authority, and vector-hub-authority network rankings, using local link information.

in Ribeiro-Neto and Muntz [1996] and Turtle and Croft [1991] in distinct scenarios.

5.3 Local Versus Global Link Information

In this section we examine the use of global link information and compare its results to those obtained by the use of local link information.

Figure 8 shows results for the vector-hub-authority ranking, when global and local information are considered. We see that the global ranking has a gain in precision, at low recall levels, superior to that of the local ranking. At high recall, precision for the global ranking drops below that of the local ranking, approaching the performance obtained by the vector space model.

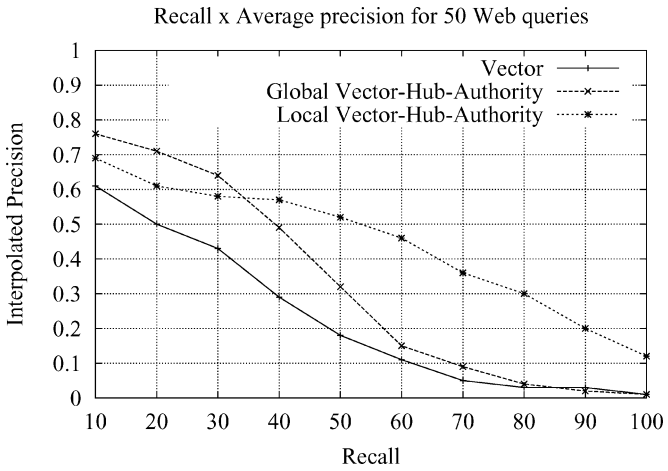

Fig. 8. Precision figures for vector, global vector-hub-authority, and local vector-hub-authority rankings.

The high precision at low recall levels provided by the use of global information shows that this type of information can be effectively used to improve the ranking provided by Web search engines, whose users are mostly interested in the first 10 answers. On the other hand, due to the nature of the HITS algorithm [Kleinberg 1998], global hub and authority values are somewhat divided among all pages in the collection. Therefore, for large collections, many pages get equal global hub and authority values that are very close to zero. At high recall levels, this makes the final combined global vector-hub-authority value dependent mainly on the vector part of the combination. The combined global ranking becomes, thus, very similar to the vectorial ranking, which explains the lower precision at high recall values.

Another problem happens if the collection consists of several disconnected components. In this case, HITS will assign zero hub and authority values to all but the main group of pages. In our collection, HITS attributed hub and authority values to roughly 170,000 pages. Although a small percentage of the total collection, this is enough to give significant improvement in the precision of the results. Since a very small number of pages (about 1%) accounts for the great majority of in- and out-links (about 80%), the most important pages (i.e., the pages containing the answers to the most frequent queries) are represented in this group. Notice that this distribution is not unusual in the Web, as can be seen in Kumar et al. [2000].

Table III shows the average gains in precision, compared to the vector space model, for the local and global vector-hub-authority rankings, measured at the 10, 20, and 30 top ranked documents. We observe that the global vector-hub-authority ranking provides more relevant documents within the first 20 documents. For the first 10 documents, the gain in precision obtained with the use of global information is much higher than the gain obtained with the use of local information. Therefore the use of global evidence is useful mainly at low recall values. Also, unlike local information, global information has the advantage of

Table III. Precision Figures for the Top 10, 20, and 30 Ranked Documents, when Local and Global Information Is Used with-the Vector-Hub-Authority Ranking

Number of Pages	Vector Precision	Local Information		Global Information	
		Precision	Gain (%)	Precision	Gain (%)
10	0.541	0.582	8	0.671	24
20	0.403	0.566	40	0.617	53
30	0.297	0.523	76	0.478	61

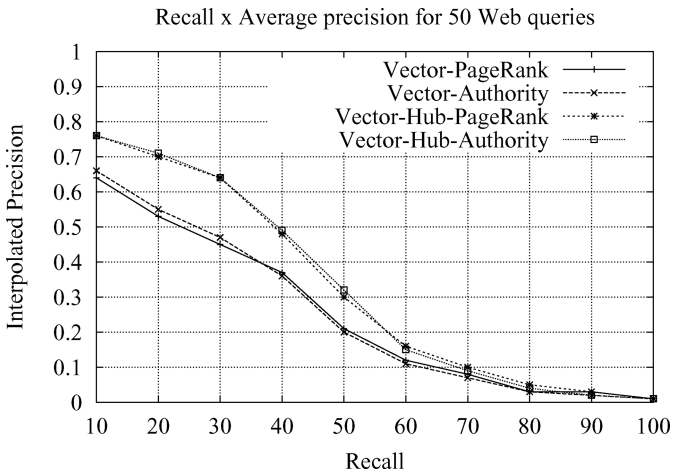

Fig. 9. Precision figures for the PageRank algorithm.

not requiring any extra computation at query time. This makes it an interesting alternative for systems where high precision is especially important for the first documents in the ranking and where query answer time is critical, such as Web search engines [Spink et al. 2001].

Comparison with PageRank. Since PageRank is a popular method of link analysis based on global information, it is interesting to compare its performance with that of the HITS algorithm. The PageRank score of a document was combined with its vectorial score through Equation (15), used for the Vector-Authority evidence combination. Figure 9 shows the resulting precision/recall curves.

Interestingly, the PageRank algorithm performs quite similarly to the HITS algorithm. When combined only with content-based information, its performance is very close to the performance of HITS when only authorities and document content are considered. By combining PageRank with hub and authority evidence, through Equation (16), the curve is still close to that of the vector-hub-authority combination for the HITS algorithm. In fact, if we observe the documents returned by both algorithms, we see that most of the documents returned as authorities by HITS are also present in the PageRank results, although in different positions.

This behavior can be explained as follows. Since the majority of links are concentrated in only 1% of the whole collection, we can expect many of the

Table IV. Precision Figures for the Vector, Vector-Authority, Vector-Hub, and Vector-Hub-Authority Network Rankings, Using Local Information

	Vector	Vec-Aut		Vec-Hub		Vec-Hub-Aut	
Recall	Precision	Precision	Gain (%)	Precision	Gain (%)	Precision	Gain (%)
10	0.610	0.582	−5	0.701	15	0.689	13
20	0.501	0.556	11	0.647	29	0.614	23
30	0.427	0.477	12	0.613	44	0.582	36
40	0.286	0.394	38	0.511	79	0.566	98
50	0.185	0.322	74	0.378	104	0.524	183
60	0.114	0.281	146	0.295	159	0.456	300
70	0.054	0.195	261	0.187	246	0.357	561
80	0.031	0.160	416	0.150	384	0.299	865
90	0.026	0.132	408	0.134	415	0.204	685
100	0.010	0.103	930	0.106	960	0.117	1070
Average	0.267	0.351	31	0.405	52	0.466	74

Table V. Precision Figures for the Vector, Vector-Authority, Vector-Hub, and Vector-Hub-Authority Network Rankings, Using Global Information

	Vector	Vec-Aut		Vec-Hub		Vec-Hub-Aut	
Recall	Precision	Precision	Gain (%)	Precision	Gain (%)	Precision	Gain (%)
10	0.610	0.659	8	0.748	23	0.756	24
20	0.501	0.549	10	0.688	37	0.713	42
30	0.427	0.472	11	0.617	44	0.637	49
40	0.286	0.364	27	0.411	44	0.492	72
50	0.185	0.201	9	0.232	25	0.323	75
60	0.114	0.112	−2	0.142	25	0.150	32
70	0.054	0.073	35	0.072	33	0.092	70
80	0.031	0.029	−6	0.035	13	0.043	39
90	0.026	0.024	−8	0.026	0	0.024	−8
100	0.010	0.010	0	0.010	0	0.010	0
Average	0.267	0.292	10	0.345	29	0.360	35

pages designated as good authorities by HITS to have a high degree of in-links. This would also make them very likely to be visited by a surfer randomly following links. PageRank introduces random jumps to avoid circular links, where the random surfer would get caught in an endless cycle. However, in our collection, we removed links between pages within the same site, which greatly reduces the effects of circular linking. Thus we can expect PageRank and HITS to perform similarly.

5.4 Summary of Evaluation Results

Table IV summarizes the results of our experiments using local sources of evidence. The table shows the gains in precision of the combined rankings relative to the vector space model. Although the vector-authority ranking provides a gain of 31% in precision, the vector-hub ranking yields a gain of 52%. Furthermore, the combined vector-hub-authority ranking leads to a higher gain in precision, close to 74%.

Table V summarizes the results of our experiments using global sources of evidence. The largest gain in precision is now 35%, obtained by the vector-hub-authority combination. The vector-authority and vector-hub combinations yield

Table VI. T-Test Significance for Each Recall Level, for 50 Test Queries

Recall	Local	Global
10	0.094	0.003
20	0.054	<0.001
30	0.016	<0.001
40	<0.001	<0.001
50	<0.001	0.001
60	<0.001	0.074
70	<0.001	0.027
80	<0.001	0.177
90	<0.001	0.165
100	<0.001	0.084

gains of 10 and 29%, respectively. Even though average values are smaller than those obtained using local information, we see an improvement in precision for recall values below 40%. As discussed before, this means that relevant pages retrieved using global information are more concentrated at the top of the ranking, even if the answer set contains a smaller number of relevant pages.

To confirm the significance of the results, a *t-test* [Anderson and Finn 1997] was performed for the 50 test queries, comparing the means of the precision values for the vectorial and vector-hub-authority rankings. The test was performed considering global and local sources of evidence. Table VI shows the significance attained at each recall level from 10 to 100%. We observe that all measures present a high reliability, except the precision values for the global ranking at high recall levels. This lower confidence can be explained. As was discussed in Section 5.3, at high recall levels hub and authority values are close to zero for many pages. This makes the combined rankings similar to the vectorial ranking. Any differences between the two are, therefore, more likely to be due to chance, thus leading to a lower confidence level.

6. CONCLUSIONS AND FUTURE WORK

In this work we compared the use of local and global link information in a hyperlinked environment. Link-based information was combined with information on document content through Bayesian belief networks. For the comparison we ran experiments using a reference collection of about six million documents extracted from the Web. Our results showed that combining content-based and link-based sources of evidence yields better retrieval results than using any of them separately. In fact, local link information yielded a gain in precision of 74% when compared with the results of the vector space model. Global link information yielded a gain in precision of 35% for our test collection. These results suggest that, in general, the use of local link information is more promising.

Interestingly, global information was shown to be useful in improving retrieval results at low recall values. For the first 10 documents in the ranking, the use of global link information produced an average gain in precision of 28%, whereas the use of local information showed a gain of only 8%. Also, global information required no extra processing at query time. These characteristics make the use of global sources of evidence a valuable alternative whenever high

precision at low recall is important and query processing efficiency is essential, such as in Web search engines.

As an added result, this work confirms the effectiveness of our Bayesian Network model for combining evidence from different sources to improve document ranking, as was indicated before in Silva et al. [2000].

It is important to note that previous results in the TREC Web track [Hawking and Craswell 2001] seem to indicate that the use of link analysis brings few gains to the task of document ranking. However, several fundamental points in the TREC experiments make the testing environment different from ours. First, documents in the TREC collection were judged only according to their text, and judges were not allowed to follow links. In our collection, on the other hand, the judges had access to the real site, including its multimedia content, and were allowed to follow links. Thus most pages classified as relevant were hub pages and site homepages were only about 36% of all the relevant documents. Second, link analysis algorithms are expected to work better for more general queries. Since the most frequent queries in the Web are, usually, quite general, we can expect good results from link-based retrieval. TREC Web queries, however, tend to be very specific, thus harming the effectiveness of such an approach. Third, our collection was collected by Web crawlers, whereas TREC was constructed as a subset of a larger collection. Although care was taken in TREC to assure good linkage among pages, it is very likely that the link distribution is quite different from the Web, thus causing some disparity in the results.

Future work on this subject includes new applications for the combination of link- and content-based evidential information. Besides the ranking of Web pages, link-based evidence can be combined with content-based evidence for different retrieval tasks, such as document clustering or information filtering. Document clustering has already been approached in several works [Dean and Henzinger 1999; Kumar et al. 1999], but an approach that combines both types of evidence is still lacking. Also, link-based evidence can be combined with other sources of information, besides document content, such as user relevance feedback or past queries information. In general, the use of link-based sources of evidence in document collections, and in particular in the Web, is still a recent subject and offers many different paths to be explored.

REFERENCES

ANDERSON, T. W. AND FINN, J. D. 1997. *The New Statistical Analysis of Data*, 1st ed. Springer-Verlag, New York.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*, 1st ed. Addison-Wesley-Longman, Reading, MA.

BHARAT, K. AND HENZINGER, M. R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia), 104–111.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, 107–117.

CHAKRABARTI, S., DOM, B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D., AND KLEINBERG, J. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the Seventh International World Wide Web Conference* (Brisbane), 65–74.

COHN, D. AND CHANG, H. 2000. Learning to probabilistically identify authoritative documents. In *Proceedings of the Seventeenth International Conference on Machine Learning* (Stanford, CA), 167–174.

DEAN, J. AND HENZINGER, M. R. 1999. Finding related pages in the World Wide Web. *Comput. Netw. 31*, 11–16 (May), 1467–1479. Also in *Proceedings of the Eighth International World Wide Web Conference*.

DUMAIS, S. T. AND JIN, R. 2001. Probabilistic combination of content and links. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans), 402–403.

GAO, J., CAO, G., HE, H., ZHANG, M., NIE, J.-Y., WALKER, S., AND ROBERTSON, S. 2001. TREC—10 Web track experiments at MSRA. In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)* (Gaithersburg, MD), 384–392.

GARFIELD, E. 1972. Citation analysis as a tool in journal evaluation. *Science 178*, 4060, 471–479.

HAWKING, D. AND CRASWELL, N. 2001. Overview of TREC-2001 Web track. In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)* (Gaithersburg, MD), 61–67.

HAWKING, D., CRASWELL, N., AND THISTLEWAITE, P. B. 1998. Overview of TREC-7 very large collection track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* (Gaithersburg, MD), 91–104.

HAWKING, D., CRASWELL, N., THISTLEWAITE, P. B., AND HARMAN, D. 1999. Results and challenges in Web search evaluation. *Comput. Netw. 31*, 11–16 (May), 1321–1330. Also in *Proceedings of the Eighth International World Wide Web Conference.*

KANUNGO, T. AND ZIEN, J. Y. 2001. Integrating link structure and content information for ranking web documents. In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)* (Gaithersburg, MD), 237–239.

KESSLER, M. M. 1963. Bibliographic coupling between scientific papers. *Amer. Doc. 14*, 1 (Jan.), 10–25.

KLEINBERG, J. M. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (San Francisco), 668–677.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. Trawling the Web for emerging cyber-communities. *Comput. Netw. 31*, 11–16 (May), 1481–1493. Also in *Proceedings of the Eighth International World Wide Web Conference.*

KUMAR, S. R., RAGHAVAN, P., RAJAGOPALAN, S., SIVAKUMAR, D., TOMKINS, A., AND UPFAL, E. 2000. The Web as a graph. In *Proceedings of the Nineteenth Symposium on Principles of Database Systems* (Dallas), 1–10.

LEMPEL, R. AND MORAN, S. 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Comput. Netw. 33*, 1–6 (June), 387–401. Also in *Proceedings of the Ninth International World Wide Web Conference.*

LEMPEL, R. AND MORAN, S. 2001. Salsa: The stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst. 19*, 2 (April), 131–160.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1998. The PageRank citation ranking: Bringing order to the Web. Tech. Rep., Stanford Digital Library Technologies Project.

PEARL, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 2nd ed. Morgan-Kaufmann, San Francisco.

RIBEIRO-NETO, B. AND MUNTZ, R. 1996. A belief network model for IR. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich), 253–260.

RIBEIRO-NETO, B., SILVA, I., AND MUNTZ, R. 2000. *Soft Computing in Information Retrieval: Techniques and Applications*, 1st ed. Springer Verlag, New York, Chapter 11—Bayesian Network Models for IR, 259–291.

SALTON, G. 1971. Automatic indexing using bibliographic citations. *J. Doc. 27*, 2 (June), 98–110.

SALTON, G. AND MCGILL, M. J. 1983. *Introduction to Modern Information Retrieval*, 1st ed. McGraw-Hill, New York.

SALTON, G., ALLAN, J., AND BUCKLEY, C. 1993. Approaches to passage retrieval in full text information systems. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pittsburgh), 49–58.

SILVA, A., VELOSO, E., GOLGHER, P., RIBEIRO-NETO, B., LAENDER, A., AND ZIVIANI, N. 1999. CobWeb—A crawler for the Brazilian Web. In *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE'99)* (Cancun), 184–191.

SILVA, I., RIBEIRO-NETO, B., CALADO, P., MOURA, E., AND ZIVIANI, N. 2000. Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece), 96–103. Best Student paper.

SMALL, H. G. 1973. Co-citation in the scientific literature: A new measure of relationship between two documents. *J. Amer. Soc. Inf. Sci. 24*, 4 (July), 265–269.

SMALL, H. G. AND KOENIG, M. E. D. 1977. Journal clustering using a bibliographic coupling method. *Inf. Process. Manage. 13*, 5, 277–288.

SPINK, A., WOLFRAM, D., JANSEN, B. J., AND SARACEVIC, T. 2001. Searching the Web: The public and their queries. *J. Amer. Soc. Inf. Sci. Technol. 52*, 3 (Feb.), 226–234.

TURTLE, H. AND CROFT, W. B. 1990. Inference networks for document retrieval. In *Proceedings of the Thirteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Brussels), 1–24.

TURTLE, H. AND CROFT, W. B. 1991. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst. 9*, 3 (July), 187–222.

WESTERVELD, T., KRAAIJ, W., AND HIEMSTRA, D. 2001. Retrieving Web pages using content, links, URLs and anchors. In *Proceedings of the Tenth Text REtrieval Conference (TREC-2001)* (Gaithersburg, MD), 663–672.

WITTEN, I. H., MOFFAT, A., AND BELL, T. C. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd ed., Morgan-Kaufmann, San Francisco.

WONG, S. K. M. AND YAO, Y. Y. 1995. On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst. 13*, 1 (Jan.), 38–68.

XU, J. AND CROFT, W. B. 1996. Query expansion using local and global document analysis. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zurich), 4–11.

XU, J. AND CROFT, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst. 18*, 1 (Jan.), 79–112.

Received November 2001; revised July 2002; accepted October 2002