Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER International Journal of Approximate Reasoning 34 (2003) 163–179

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Bayesian belief networks for IR

Marco Antônio Pinheiro de Cristo [a,d], Pável Pereira Calado [a,1], Maria de Lourdes da Silveira [a,e,2], Ilmério Silva [b], Richard Muntz [c,3], Berthier Ribeiro-Neto [a,*,4]

[a] *Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil*
[b] *Department of Informatics, Federal University of Uberlândia, Uberlândia, MG, Brazil*
[c] *Department of Computer Science, University of California, Los Angeles, CA, USA*
[d] *Fucapi, Technology Foundation, Manaus, Brazil*
[e] *Prodabel, Information Technology Company for the City of Belo Horizonte, Minas Gerais, Brazil*

Received 3 March 2003; accepted 3 June 2003

Abstract

We review the application of Bayesian belief networks to several information retrieval problems, showing that they provide an effective and flexible framework for modeling distinct sources of evidence in support of a ranking. To illustrate, we explain how Bayesian networks can be used to represent the classic vector space model and demonstrate how this basic representation can be extended to naturally incorporate new evidence from distinct information sources. These models have been shown useful in several text collections, where the combination of evidential information derived from past queries, thesauri, and the link structure of Web pages has led to significant improvements in retrieval performance.

* Corresponding author.
E-mail address: berthier@dcc.ufmg.br (B. Ribeiro-Neto).
[1] Author supported by MCT/FCT scholarship SFRH/BD/4662/2001.
[2] Author supported in part by CNPq Individual Grant 141.294/2000-0.
[3] Author supported by NSF grants IIS-0086116, ANI-0085773, and EAR-9817773.
[4] Author supported in part by I3DL project grant 680154/01-9 and Finep/MCT/CNPq Grant 76.97.1016.00, project SIAM under program Pronex.

0888-613X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2003.07.006

Keywords: Bayesian networks; Information retrieval; Ranking; Evidence combination

1. Introduction

The quest towards high quality results in information retrieval has considered distinct research avenues. Of these, one of the most promising is the idea of combining relevance information generated by distinct sources of evidential knowledge. For instance, one popular source of evidential knowledge in the Web is its link structure. This ranking can be combined with information on keywords (in the form of a vector ranking, for instance) to yield improved results [4]. We say that evidence from the Web link structure has been combined with evidence from the contents of the documents. Since these two sources of evidence are distinct and somewhat independent, their contributions tend to result in an effect that improves the accuracy of the ranking.

Combining distinct sources of evidential knowledge in support of a ranking is, therefore, a current and important research direction. To accomplish such a combination in a consistent fashion, a formal framework is highly desirable. One such framework that has gained acceptance in the research community utilizes Bayesian belief networks [11]. Indeed, from the point of view of IR research, Bayesian networks are attractive because they provide a modeling framework that is powerful enough to represent various types of evidential information, such as keyword-based knowledge, Web linkage knowledge, thesauri-based knowledge, etc.

In this work, we discuss the application of Bayesian belief networks to IR. We review the basic Bayesian network model proposed in [12] and discuss how this model can be extended to accommodate distinct sources of evidential information, in particular, knowledge of past queries, knowledge of the link structure of the Web, and knowledge of the concepts of a thesaurus and their relationships. We briefly review experimental results reported in the literature. In general, we observe that the use of Bayesian belief networks to combine distinct evidential information results in advanced IR models that consistently outperform their traditional counterparts.

The paper is organized as follows. In Section 2, we introduce the graph-based framework of Bayesian networks. In Section 3, we briefly present different approaches to IR using Bayesian networks. In Section 4, we describe in more detail a Bayesian network model for document ranking. In Section 5, we show that this model can be extended to combine evidence from distinct sources, to improve retrieval performance. Finally, in Section 6, we present some conclusions.

2. Bayesian networks

Bayesian networks provide a graphical formalism for explicitly representing the independencies among the variables of a domain, thus providing a concise specification of a joint probability distribution [11]. This representation is based on a directed acyclic graph where a set of random variables makes up the nodes of the network and a set of directed links connects pairs of nodes. In this graph, an edge from one node to another one means that the first has a *direct influence* on the second. This influence is quantified through a conditional probability distribution function correlating the states of each node with the states of its parents.

To illustrate, let X and Y be two random variables and let x and y be two of their respective values. We use X and Y to refer to the random variables as well as to the nodes in the network associated with these variables. An edge directed from Y, the *parent* node, to X, the *child* node, represents the influence of the variable Y on the variable X, which is quantified by the conditional probability $P(x|y)$.

In general, let $\mathbf{P}$ be the set of all parent nodes of a node X, as illustrated in Fig. 1. Further, let $\mathbf{p}$ be a set of values for all the variables in $\mathbf{P}$ and let x be a value of the variable X. The influence of $\mathbf{P}$ on X can be modeled by any function $\mathbf{F}$ such that $\sum_x \mathbf{F}(x,\mathbf{p}) = 1$ and $0 \leqslant \mathbf{F}(x,\mathbf{p}) \leqslant 1$. The function $\mathbf{F}(x,\mathbf{p})$ provides a numerical quantification for $P(x|\mathbf{p})$.

A Bayesian network for a joint probability distribution $P(x_1,x_2,x_3,x_4,x_5)$ is shown in Fig. 2. Node X_1, the *root node*, is a node without parents whose

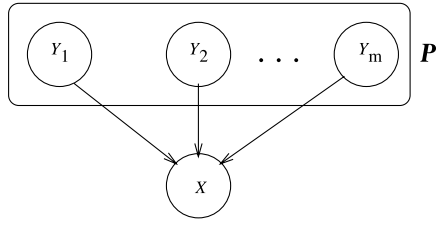

Fig. 1. Parents of a node in a Bayesian network.

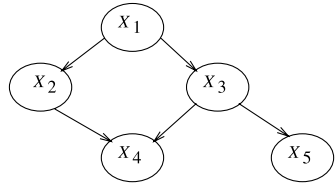

Fig. 2. Example of a Bayesian network.

(marginal) probability distribution is $P(x_1)$, where the domain of x_1 is the set of values that X_1 takes on with non-zero probability, and is called a *prior probability*. This probability can be used to represent previous knowledge of the modeled domain. Due to the independencies declared in Fig. 2, the joint probability distribution can be computed as $P(x_1, x_2, x_3, x_4, x_5) = P(x_1) \times P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_3)$.

A key advantage of Bayesian networks is their synthesized representation of probabilistic relationships. In fact, it is necessary to consider only the known independencies among the variables in a domain, rather than specifying a complete joint probability distribution. The independencies declared at modeling time are then used to infer *beliefs* for all variables in the network. The inference mechanism, though exponential in the worst case, is efficient in many practical situations, particularly in those that arise in the IR arena.

3. Bayesian networks for IR

Bayesian network models were first introduced in IR by Turtle and Croft in [17]. In their model, index terms, documents and user queries are seen as events and are represented as nodes in a Bayesian network. The model takes the viewpoint that the observation of a document induces belief on its set of index terms, and that specification of such terms induces belief in a user query or information need. This model was shown to perform better than traditional probabilistic models for the task of document ranking.

Later, a second model was proposed by Ribeiro-Neto and Muntz in [12], where the elements of an IR system are formally defined as concepts in a sample space. Their work not only provides a probabilistic justification for the model, but also demonstrates that the combination of evidence from past queries with the vector space ranking yields better results than the use of a vector space ranking alone.

More recently, Acid et al. [1] presented a third model whose network topology is defined in such way that an exact propagation algorithm, proposed in their work, can be used to efficiently compute the relevance probabilities of the documents. When compared to Turtle and Croft's work for the task of document ranking, this model shows better performance in four out of five reference collections.

Bayesian networks have also been applied to other IR problems besides ranking as, for example, relevance feedback [9], automatic construction of hypertext [15], query expansion [6], information filtering [5], assigning structure to database queries [3], and document clustering and classification [8].

In this paper, we adopt the Bayesian framework proposed by Ribeiro-Neto and Muntz in [12] for modeling distinct IR problems. Before proceeding, let us review this framework.

4. A Bayesian network for document ranking

In this section, we describe the Bayesian network model proposed in [12]. This model takes an epistemological view of the IR problem. [5] However, contrary to [17], the model is derived from a probabilistic argument based on a clearly defined sample space.

In a traditional information retrieval system, documents are indexed by keywords. We interpret the set of all keywords as *the universe of discourse* **U**, which we take as our *sample space*. Let t be the total number of keywords in a collection. Then, $\mathbf{U} = \{k_1, k_2, \ldots, k_t\}$, where each keyword k_i is interpreted as an elementary concept in the space **U**. Further, each subset u of **U**, composed of elementary concepts, is interpreted as a non-elementary concept, or simply a concept. [6]

Associated with each keyword k_i, we define a random variable, also denoted by k_i. This variable is 1 to indicate that the keyword was observed (i.e., is on the state *on*). A document d_j is modeled as a set composed of selected keywords that occur in its text. If all the variables associated with the keywords in the document are in the *on* state, we say that the document has been observed. A query q is modeled analogously. To allow referring to the state of each variable k_i, we use an indicator function $I_u(k_i)$ that returns the value of the variable k_i according to the concept u. The function $I_u(k_i)$ is 1 if $k_i \in u$ and 0 otherwise.

Let P be a joint probability distribution defined over the sample space **U**. As in [18], the probability $P(c)$, associated with a generic concept c in the space **U**, is defined as follows:

$$P(c) = \sum_{u \in \mathbf{U}} P(c|u)P(u) \tag{1}$$

$P(u)$ is a prior probability associated with each concept u as discussed below. $P(c|u)$ defines an intersection between the concepts c and u in the space **U**. Thus, $P(c)$ can be interpreted as a *degree of coverage* of the space **U** by c. Such degree of coverage is computed by contrasting each concept u, $u \subseteq \mathbf{U}$, with c (this explains the weighted sum).

Given the sample space **U**, we can have 2^t concepts (subsets of **U**). If all concepts are considered to be equally likely a priori, each prior probability $P(u)$ is set to $P(u) = (1/2)^t$. The sum in Eq. (1) is extended over all 2^t concepts.

[5] Probabilities are interpreted as degrees of *belief* that can be specified independently of experimentation.

[6] This set-theoretic vision of concepts allow us to reason with the logical notions of conjunction, disjunction, negation, and implication as operations of intersection, union, complementation, and inclusion.

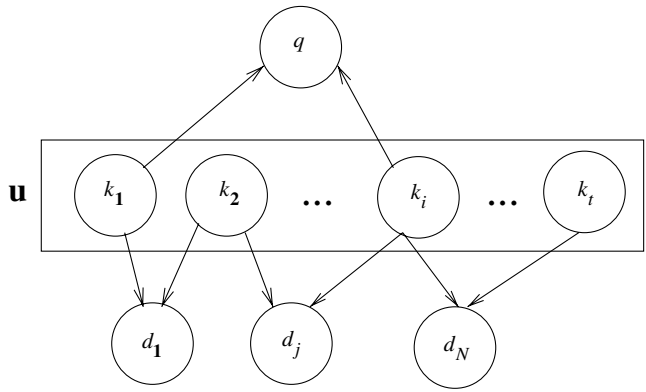

Fig. 3. Bayesian network for a query q given by the keywords k_1 and k_i.

However, information retrieval systems usually require considering the influence of a few concepts only.

Given the concept space **U**, it is natural to model queries and documents as concepts in **U**. As a result, queries and documents are treated analogously. This symmetry induces the Bayesian network of Fig. 3.

In this network, 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 set **u** is used to refer to any of the 2^t possible states of the k_i root nodes. Instantiation of the root nodes *separates* the document nodes from the query node, making them mutually independent. Thus, in the belief network of Fig. 3, we say that the query is on the *query side* of the network, while the documents are on the *document side* of the network.

With the node q is associated a binary random variable which is also denoted by q. This variable is 1 (also said to be *on*) to indicate that the concept of **U** associated with query q was observed. A document d_j is modeled analogously, i.e., there is also a binary random variable associated with d_j. This variable is 1 (also said to be *on*) to indicate that the concept of **U** associated with the document d_j was observed.

In the network of Fig. 3, the ranking computation is based on interpreting the similarity between a document d_j and the query q as an intersection between the concepts d_j and q. To quantify the degree of intersection of the concept d_j, given the concept q, we use the probability $P(d_j|q)$. Thus, to compute a ranking, we use Bayes' law and the rule of total probabilities, as follows. Let $\eta = 1/P(q)$ be a normalizing constant, as used in [11]:

$$P(d_j|q) = \eta \sum_{\mathbf{u}} P(d_j|\mathbf{u})P(q|\mathbf{u})P(\mathbf{u}) \tag{2}$$

which is the generic expression for the rank of a document d_j with regard to a query q, in the belief network model.

This expression can be used to represent the rankings generated by any of the classic models. This is important because it allows combining features of distinct models into the same representational scheme. For instance, a Bayesian network can be used to compute the vector space model ranking, as we explain in the following.

In the vector space model, queries and documents are represented as weighted vectors in a t-dimensional space. Let w_{ij} be the weight associated with the keyword k_i in the document d_j and w_{iq} be the weight associated with the keyword k_i in the user query q. Then, $\vec{d}_j = (w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots, w_{tj})$ and $\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{iq}, \ldots, w_{tq})$ are the weighted vectors used to represent the document d_j and the query q. The weights for $\vec{d}_j$ are computed using classic *tf-idf* schemes (see [2,14] for details). The weights for $\vec{q}$ are 1 to if the term is in the query and 0 otherwise. The ranking of the document d_j with regard to the query q is computed by the *cosine similarity formula*, that is, the cosine of the angle between the two corresponding vectors:

$$\mathrm{sim}(\vec{d}_j, \vec{q}) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{3}$$

To calculate this ranking, we have to make Eq. (2) equivalent to Eq. (3). This is accomplished through proper specification of probabilities $P(u)$, $P(q|u)$, and $P(d_j|u)$, as follows:

$$P(u) = (1/2)^t \tag{4}$$

$$P(q|u) = \begin{cases} 1 & \text{if } \forall k_i, \quad I_q(k_i) = I_u(k_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$P(d_j|u) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iu}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iu}^2}} \tag{6}$$

That is, we use $P(q|u)$ to select the concept u that matches the query keywords, and $P(d_j|u)$ to compute the cosine similarity measure. This specification is valid because $P(d_j|u)$ is the cosine of the angle between two vectors, a number between 0 and 1. Applying Eqs. (4)–(6) to Eq. (2), we get a ranking equivalent to the one obtained by Eq. (3). Thus, Bayesian networks can be used to easily model the traditional vectorial ranking.

We observe that the belief network is used here as a modeling framework and not as an inference engine. While more complex designs are possible, our simple representation is powerful enough to allow modeling important relationships between documents, queries, and user needs.

5. Extended Bayesian network models for evidence combination

One of the main strengths of Bayesian networks is that they can be naturally extended with additional pieces of evidence obtained from alternative sources. In this section, we present some examples of such extended models for IR.

5.1. Using evidence from past queries

Here, we review the use of past queries to improve the ranking of the current query, as proposed in [12]. Consider that a set of past queries was saved in a log. Assume also that some of the documents returned by each query were evaluated by a user, who labeled each of them as *relevant* or *non-relevant*. This knowledge about the relevance of documents to past queries can be used to improve the retrieval performance for the current query.

Let $c_1, c_2, \ldots, c_p$ be references to the past queries and c_0 be a reference to the current query. To model the relationship between each past query c_l, $1 \leqslant l \leqslant p$ and c_0, we compare them directly as follows. Each past query c_l is seen as a concept in our sample space **U**. Thus, to each past query c_l is associated a node in the network and a binary random variable. These query nodes are represented as root nodes as shown in Fig. 4. This representation allows considering the individual influence of each past query on the current query.

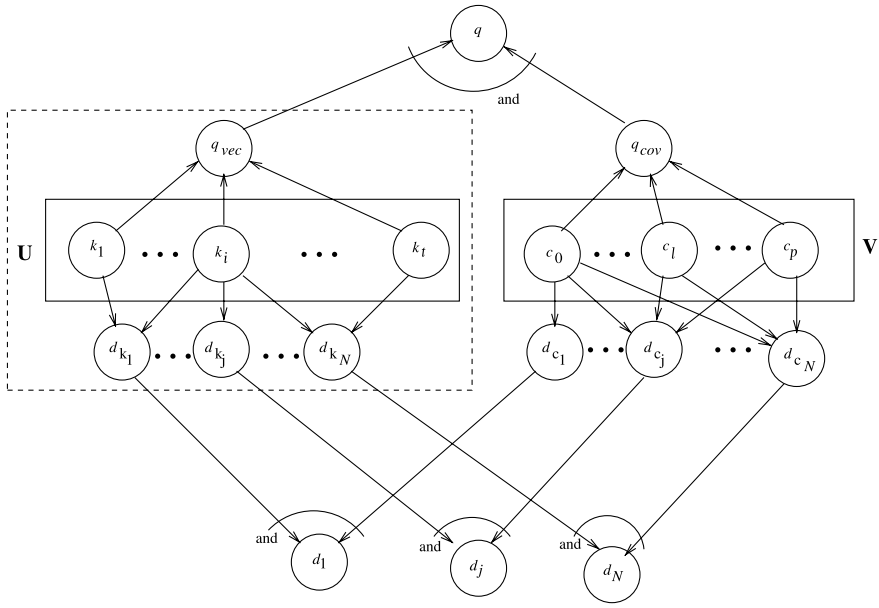

Fig. 4. Bayesian network expanded with evidence from past queries.

In Fig. 4, the left hand side of the network represents the original network of Fig. 3. The right hand side of the network models the evidence from past queries. Let $\mathbf{V} = \{c_0, c_1, \ldots, c_p\}$ be the set of all queries, where c_0 represents the current query. The document nodes, denoted by d_{c_j}, model the documents retrieved by: (1) the current query c_0 (d_{c_j} is related to c_0 if it has at least one keyword in common with q); and (2) a past query c_l (a document d_{c_j} is related to c_l if it has been identified as relevant to the query c_l).

A query c_l, $1 \leqslant l \leqslant p$, is considered related to the current query c_0 if the set of keywords in the query c_l covers, at least partially, the set of keywords in the query q, i.e., if c_l has at least one keyword in common with c_0. This partial coverage relationship can be quantified using, for instance, the cosine similarity formula. The node q_{cov} is used to model this coverage relationship. Since the probability that q_{cov} is *on* is non-zero when only one of the c_l nodes is *on*, we say that the influence of each query c_l on the node q_{cov} is considered separately, one at a time.

This influence is then combined with the belief in the original query q_{vec} through a conjunctive operator [7] (the node q). On the document side of the network, the evidence collected by the document node d_{c_j} is combined with the current keyword-based rank for the document node d_{k_j}, both representing the document d_j on distinct contexts, through a conjunctive operator to yield a final rank for the document node d_j.

Let u represent the state of the set of the k_i root nodes, and let v represent the state of the set of c_l root nodes. For the c_l root nodes, we define states v_l such that $v = v_l \iff I_v(c_l) = 1 \wedge I_v(c_j) = 0$, $\forall_{j \neq l}$, where the function $I_v(c_l)$ is as defined in Section 4. Thus, we consider only those states in which there is a single query c_l active at a time.

In Fig. 4, the rank $P(d_j|q)$ associated with a document d_j is computed through basic conditioning on the root nodes and application of Bayes' rule.

$$\begin{aligned} P(d_j|q) &= \eta \sum_{u,v} P(d_j|u,v)P(q|u,v)P(u)P(v) \\ &= \eta \sum_{u,v} P(d_{k_j}|u)P(d_{c_j}|v)P(q_{\text{vec}}|u)P(q_{\text{cov}}|v)P(u)P(v) \end{aligned} \tag{7}$$

where η is a normalizing constant. The probabilities $P(u)$, $P(q_{\text{vec}}|u)$, and $P(d_{k_j}|u)$ are given by Eqs. (4)–(6) respectively. For the probability $P(q_{\text{cov}}|v)$, we write

[7] By a conjunctive operator, we mean that a node is *on* if, and only if, all of its parent nodes are also *on*.

$$P(q_{\text{cov}}|v) = \begin{cases} \dfrac{\vec{q}_0 \cdot \vec{q}_l}{|\vec{q}_0| \times |\vec{q}_l|} & \text{if } v = v_l \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $\vec{q}_0$ and $\vec{q}_l$ are the vectors associated with the queries c_0 and c_l. Eq. (8) shows that we consider the influence of a single query node c_l at a time. For each query node c_l, Eq. (8) quantifies the similarity between the past query c_l and the current query through the cosine similarity formula.

For the probability $P(d_{c_j}|v)$, we define:

$$P(d_{c_j}|v) = \begin{cases} 1 & \text{if } v = v_l \text{ and } c_l \text{ is a parent node of } d_{c_j} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Finally, for the prior probability $P(v)$, we define:

$$P(v) = \begin{cases} P(c_0) & \text{if } v = v_0 \\ \dfrac{1 - P(c_0)}{p} & \text{if } v = v_l,\ 1 \leqslant l \leqslant p \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $P(c_0)$ is a prior probability associated with the current query. The prior probability $P(c_0)$ is an input parameter that can be used to moderate the influence of past queries in the ranking computation.

Eq. (7) sums up the individual impact of each past query on the current ranking. The current ranking itself is taken into account when c_0 is *on*. Further, the vector ranking accumulated in each node d_{k_j} is weighted by the prior probability $P(c_0)$.

Ribeiro-Neto et al. [13], using this model, reported precision improvements of at least 59% over the vector space model for the collections CFC, CISI, CACM and TREC-8, clearly suggesting that it can be used to effectively take advantage of past queries to improve retrieval performance.

5.2. Using evidence from link analysis

We now review the use of the link structure of the Web to improve retrieval results, as proposed in [4,16]. One of the richest sources of information in a hyperlinked environment, like the Web, is the knowledge about 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 [10] uses this information to measure the importance of a document based on two metrics: a degree of *authority* (how good are the document contents) and a degree of *hubness* (how good are the documents it links to). 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*. We note that all documents have an hubness and an authority degree, although in some the hubness degree will be predominant, i.e., be much greater than the authority degree, whereas in others the authority degree will predominate.

Now, we show how to extend the Bayesian network model in Fig. 3 to include evidential information extracted from the link structure of the environment, as shown in Fig. 5. As before, in this model, the left hand side represents the original network for the keyword-based evidence. The right hand side models the link-based sources of evidence.

To represent link-based evidential knowledge in the network, we associate two new nodes d_{a_j} and d_{h_j} with each document d_j in the answer set for query q. We associate a binary random variable d_{h_j} with the node d_{h_j} to model evidence associated with the document d_j as a hub. Hub values are represented in our network as the conditional probability of d_{h_j} being observed given the keywords in the query q and given an implicit knowledge of the surrounding link structure. Analogously, we associate a binary random variable d_{a_j} with the node d_{a_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

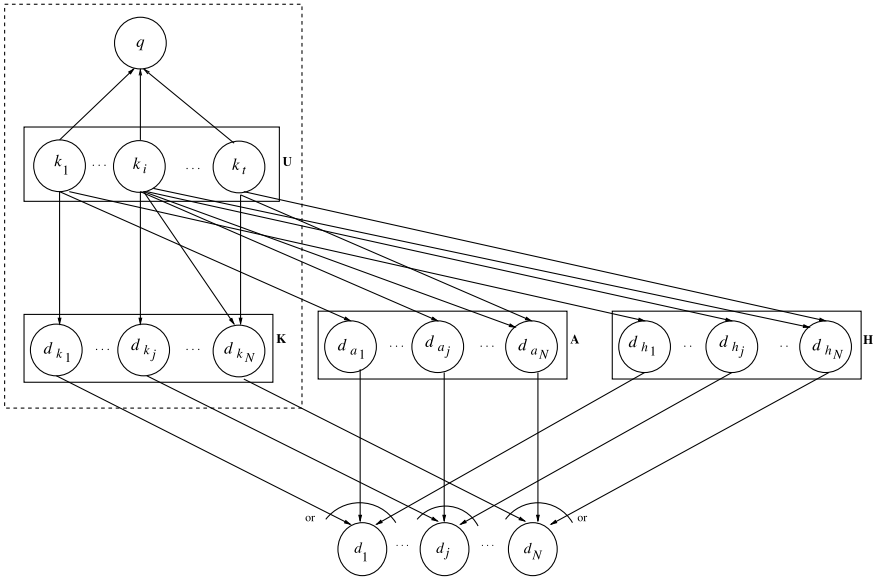

Fig. 5. Bayesian network extended with link-based evidence.

evidence, and the set **K**, composed of nodes representing keyword-based evidence. The state of the associated random variables is given by h, a, and k, respectively.

The set of nodes **U** 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 **K**, **H**, and **A**. The propagation of these beliefs in the network is done according to the conditional probabilities governing the relationships between the set **U** and each of the sets **K**, **H**, and **A**.

The binary random variable d_{h_j} associated with each node d_{h_j} of **H** is 1 if document d_j was retrieved by query q, and thus the hub evidence associated with the document is to be considered in the ranking computation. Similarly, the binary random variable d_{a_j} associated with each node d_{a_j} of **A** is 1 if document d_j was retrieved by query q, and thus the authority evidence associated with the document is to be considered in the ranking computation. The node d_j represents the combination of keyword-based and link-based evidential knowledge from the left and right hand sides of the network.

In Fig. 5, the rank $P(d_j|q)$ associated with a document d_j can be computed using Eq. (2). However, the conditional probability $P(d_j|u)$ now depends on link-based and keyword-based pieces of evidence, combined through a disjunctive [8] operator. This induces the following equation:

$$P(d_j|u) = 1 - (1 - P(d_{k_j}|u)) \times (1 - P(d_{h_j}|u)) \times (1 - P(d_{a_j}|u)) \tag{11}$$

By using a disjunctive operator, we are saying that, for a document d_j to be considered in the final ranking, it is enough that one source of evidence is available, either its similarity to the query, its hub degree, or its authority degree. We note that other combination operators could have been used, although this modeling decision as shown good results in practice.

Substituting Eq. (11) into Eq. (2), we can write:

$$P(d_j|q) = \eta \sum_u [1 - (1 - P(d_{k_j}|u)) \times (1 - P(d_{h_j}|u)) \times (1 - P(d_{a_j}|u))] \times P(q|u) \times P(u) \tag{12}$$

where $P(u)$ and $P(q|u)$ are defined as in Eqs. (4) and (5) respectively.

The computation of the probability $P(d_j|q)$ depends on the states of the nodes d_{k_j}, d_{h_j}, and d_{a_j} and can be computed through the proper specification of the conditional probabilities, establishing interesting alternatives for computing the rank of a document d_j with regard to a query q.

[8] By a disjunctive operator, we mean that a node is *on* if, and only if, at least one of its parent nodes is also *on*.

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 Eq. (6). Further, let H_{jq} and A_{jq} be the hub and authority values, respectively, associated with document d_j, computed by the HITS algorithm. If we want to represent 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 $P(d_{k_j}|u) = R_{jq}$, $P(d_{h_j}|u) = 0$, and $P(d_{a_j}|u) = 0$. Applying these probabilities and Eqs. (4) and (5) into Eq. (12), we obtain:

$$P(d_j|q) = \eta \times R_{jq} \tag{13}$$

Therefore, the general network of Fig. 5 naturally subsumes a ranking dictated by the vector space model. Similarly, we can combine any of the available evidence, thus obtaining six possible ranking functions:

1. Vector: $\eta \times R_{jq}$;
2. Hub: $\eta \times Hjq$;
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: $\eta \times [1 - (1 - R_{jq})(1 - H_{jq}) \times (1 - A_{jq})]$.

Silva et al. [16], using this model, showed that combining keyword-based and link-based sources of evidence yields better retrieval results than using any of them separately. Further, in Calado et al. [4], experiments with a Web collection suggest that the use of local link information, which is extracted from the documents related to the user query, is better than global link information, which is extracted from the whole collection. A gain in precision over the vector space model of 74% with local link information and of 35% with global link information was reported. Global link information shows better results when only the first 10 documents are considered, justifying its use by Web search engines. Combining link evidence with content based evidence is, therefore, effective to improve ranking accuracy in the Web. Bayesian networks provide a flexible, intuitive, and formally sound framework to model such evidence combination.

5.3. Using evidence from a juridical thesaurus

In this section, we review the utilization of the concepts of a juridical thesaurus and their relationships to improve the ranking for the user query, as reported in [7]. The standard IR approach does not take advantage of specialized knowledge when applied to specific domains. Thus, an alternative approach could be to combine specific domain information with that generated by a standard searching mechanism. Using this idea, we now show how to

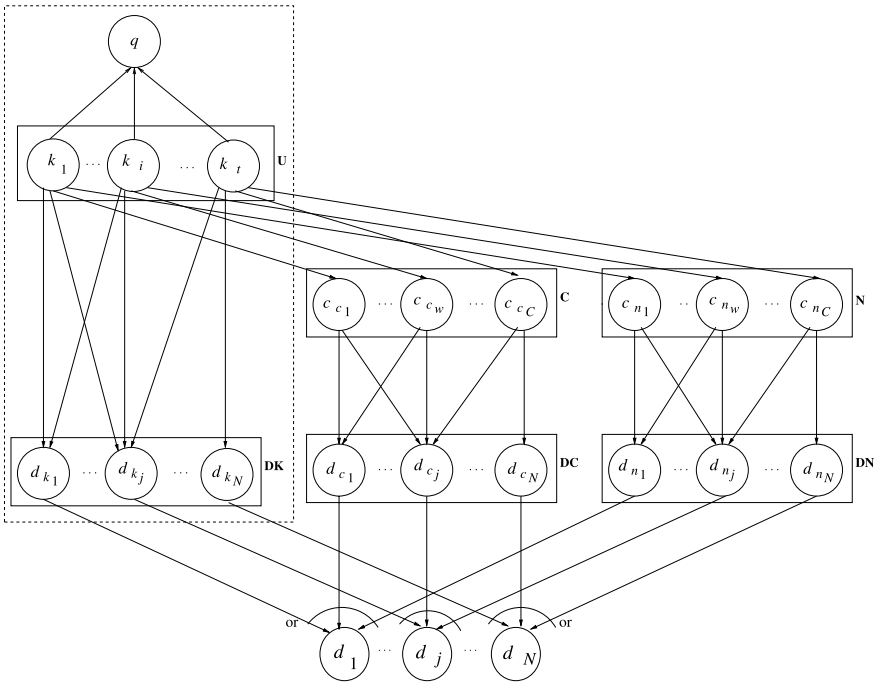

Fig. 6. Extended Bayesian network for a juridical digital library.

extend the Bayesian network model to allow combining distinct sources of evidential information obtained from a juridical thesaurus, [9] for information retrieval of juridical documents. The new network, presented in Fig. 6, can be adapted to other collections whenever there is a thesaurus representing knowledge in its domain.

In Fig. 6, **U** represents the set of keywords of the collection and **C** represents the set of concepts [10] obtained by directly mapping the query q into the concepts of the CJF Thesaurus. From these concepts, the thesaurus allows inferring related narrow terms, represented by the set **N**. Although possible to represent broad terms, related terms, and synonyms, we limit the discussion here to representing narrow terms, to simplify the model and facilitate comprehension.

[9] A thesaurus is a source of information specific to a knowledge domain that consists on a controlled list of concepts and its relationships. In this work, the CJF Thesaurus [7] for the Brazilian juridical domain was used.

[10] In this context, a concept is a word, or a set of words, that represent an entity in a knowledge domain.

Nodes d_{k_j}, d_{c_j}, and d_{n_j} represent the document d_j in distinct contexts. Node d_{k_j} is used to represent the document d_j when it appears as an answer to a keyword-based retrieval process. Nodes d_{c_j} and d_{n_j} are used to represent the document d_j when it appears as an answer to a query composed of concepts and narrower concepts, respectively, associated with the original user query. We model d_{k_j}, d_{c_j}, and d_{n_j} separately in the network to allow evaluating the impact of concept-based retrieval versus keyword-based retrieval. Evidence provided by d_{k_j}, d_{c_j}, and d_{n_j} is combined through a disjunctive operator.

The documents are ranked according to the standard vector model, meaning that the cosine formula is applied to keywords, for documents in the set **DK**, and to concepts, for documents in the sets **DC** and **DN**. These sets of ranked documents represent additional evidence that can be accumulated to yield a better ranking.

As in Eq. (2), in the extended Bayesian network, the rank of a document d_j is computed as $P(d_j|q) = \eta \sum_u P(d_j|u)P(q|u)P(u)$. Assuming that the only keywords of interest are the query keywords, as in Eq. (5), we can rewrite

$$P(d_j|q) = \eta P(d_j|u) \tag{14}$$

where u is a state of the keywords in **U** in which the only active nodes are those present in the query q.

In Fig. 6, we observe that $P(d_j|u)$ depends on the evidence obtained from the thesaurus. This evidence is used to enrich the network with distinct representations of the original query. For each representation, a ranking of documents is generated. These rankings are viewed as distinct sources of evidence on the final relevance of the documents and to combine them we use a disjunctive operator. Thus, the document node d_j accumulates all the ranking evidence through a disjunction of the beliefs associated with the nodes d_{k_j}, d_{c_j}, and d_{n_j}. This allows rewriting Eq. (14) as:

$$P(d_j|q) = \eta[1 - (1 - P(d_{k_j}|u)) \times (1 - P(d_{c_j}|u)) \times (1 - P(d_{n_j}|u))] \tag{15}$$

Evaluating each term of this equation in isolation, for instance, the term $P(d_{c_j}|u)$, we obtain:

$$P(d_{c_j}|u) = \sum_{\forall c} P(d_{c_j}|c) \times P(c|u) \tag{16}$$

Assuming that the unique concepts of interest are mapped from the query keywords, we define $P(c|u) = 1$ if the only active concepts are those mapped from the keywords in q and $P(c|u) = 0$, otherwise. As a result, we have $P(d_{c_j}|u) = P(d_{c_j}|c)$, where c is the state of the variables in **C** and only the concepts related to query q are active. The probability $P(d_{c_j}|c)$ can now be defined as the cosine similarity between the document and the active concepts.

The same reasoning can be applied to the other terms of Eq. (16), which yields:

$$P(d_j|q) = \eta[1 - (1 - P(d_{k_j}|u)) \times (1 - P(d_{c_j}|c)) \times (1 - P(d_{n_j}|n))] \quad (17)$$

where n represents the state of the set of random variables **N** where the only active nodes are those associated with the narrow terms associated to the concepts mapped from the original user query.

Eq. (17) is the ranking formula of the Bayesian model for a juridical digital library. This ranking formula allows combining evidence in several ways. For instance, consider that we are interested only in the results yielded by the vector model. To obtain this effect, we define $P(d_{c_j}|c) = 0$ and $P(d_{n_j}|n) = 0$. As a result, the ranking $P(d_j|q)$ becomes:

$$P(d_j|q) = \eta P(d_{k_j}|u) \quad (18)$$

which computes a vector ranking. To consider the combination of keyword-based and concept-based retrieval, we define $P(d_{n_j}|n) = 0$. As a result,

$$P(d_j|q) = \eta[1 - (1 - P(d_{k_j}|u)) \times (1 - P(d_{c_j}|c))] \quad (19)$$

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.

Silveira et al. [7], using this model, report a gain in average precision of about 32% over the vector model in a juridical collection composed of legal decisions taken by the high court in Brazil. Also, they showed that this vertical searching strategy is preferable to automatic query expansion on the same collection.

6. Conclusions

In this paper, we showed how Bayesian networks can be applied to several IR problems. We first discussed the general application of Bayesian networks to IR, emphasizing the representation of classic IR models. Then, we showed how new models can be created from a basic Bayesian network model, by extending it to allow combining distinct pieces of evidence. Experimental results reported in literature, obtained using the described models, confirm that Bayesian networks provide a general and powerful framework for dealing with IR problems. Further, this framework is very flexible and allows easily incorporating new pieces of information on the relevance of documents to the user query. This permits the computation of more sophisticated rankings, leading to improved retrieval results.

In general, the results here presented suggest that Bayesian belief networks provide a powerful modeling framework for ranking-related research in IR.

References

[1] S. Acid, L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, An information retrieval model based on simple Bayesian networks, International Journal of Intelligent Systems 18 (2) (2003) 251–265.
[2] R.A. Baeza-Yates, B.A. Ribeiro-Neto, Modern Information Retrieval, ACM Press/Addison-Wesley, 1999.
[3] P. Calado, A.S. da Silva, R.C. Vieira, A.H.F. Laender, B.A. Ribeiro-Neto, Searching web databases by structuring keyword-based queries, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM Press, 2002, pp. 26–33.
[4] P. Calado, B. Ribeiro-Neto, N. Ziviani, E. Moura, I. Silva, Local versus global link information in the web, ACM Transactions on Information Systems 21 (1) (2003) 42–63.
[5] J.P. Callan, Document filtering with inference networks, in: Research and Development in Information Retrieval, Zurich, Switzerland, August 1996, pp. 262–269.
[6] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, Query expansion in information retrieval systems using a Bayesian network-based thesaurus, in: Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98), Morgan Kaufmann Publishers, San Francisco, CA, 1998, pp. 53–60.
[7] M. de Lourdes da Silveira, B.A.N. Ribeiro, R. de Freitas Vale, R.T. Assumpção, Vertical searching in juridical digital libraries, in: Proceedings of the 25th Annual European Conference on Information Retrieval Research, Pisa, Italy, April 2003, pp. 491–501.
[8] S. Dumais, J. Platt, D. Heckerman, M. Sahami, Inductive learning algorithms and representations for text categorization, in: Proceedings of the Seventh International Conference on Information and Knowledge Management, ACM Press, 1998, pp. 148–155.
[9] D. Haines, W.B. Croft, Relevance feedback and inference networks, in: Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, 1993, pp. 2–11.
[10] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.
[11] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, 1988.
[12] B. Ribeiro-Neto, R. Muntz, A belief network model for IR, in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1996, pp. 253–260.
[13] B. Ribeiro-Neto, I. Silva, R. Muntz, Bayesian network models for IR, in: F. Crestani, G. Pasi (Eds.), Soft Computing in Information Retrieval Techniques and Applications, Springer-Verlag, 2000.
[14] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management: an International Journal 24 (5) (1988) 513–523.
[15] D. Shin, S. Nam, M. Kim, Hypertext construction using statistical and semantic similarity, in: Proceedings of the Second ACM International Conference on Digital Libraries, ACM Press, 1997, pp. 57–63.
[16] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, N. Ziviani, 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, ACM Press, 2000, pp. 96–103.
[17] H.R. Turtle, W.B. Croft, Inference networks for document retrieval, in: J.-L. Vidick (Ed.), SIGIR'90, 13th International Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5–7 September1990, Proceedings, ACM Press, 1990, pp. 1–24.
[18] S.K.M. Wong, Y.Y. Yao, On modeling information retrieval with probabilistic inference, ACM Transactions on Information Systems (TOIS) 13 (1) (1995) 38–68.