An Information Retrieval Model Based on Simple Bayesian Networks

Silvia Acid,[1,*] Luis M. de Campos,[1,†] Juan M. Fernández-Luna,[2,‡]
Juan F. Huete[1,§]
[1]*Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I. Informática, Universidad de Granada, 18071 Granada, Spain*
[2]*Departmento Informática, E.P.S., Universidad de Jaén, 23071 Jaén, Spain*

In this article a new probabilistic information retrieval (IR) model, based on Bayesian networks (BNs), is proposed. We first consider a basic model, which represents only direct relationships between the documents in the collection and the terms or keywords used to index them. Next, we study two versions of an extended model, which also represents direct relationships between documents. In either case the BNs are used to compute efficiently, by means of a new and exact propagation algorithm, the posterior probabilities of relevance of the documents in the collection given a query. The performance of the proposed retrieval models is tested through a series of experiments with several standard document collections. © 2003 Wiley Periodicals, Inc.

1. INTRODUCTION

Information retrieval (IR) is a subfield of computer science that deals with the automated storage and retrieval of documents.[1] An IR system is a computer program that matches user queries (formal statements of information needs) to documents stored in a database (the document collection). In our case, the documents will always be the textual representations of any data objects. An IR model is a specification about how to represent documents and queries and how to compare them. Many IR models (as well as the IR systems implementing them) such as the vector space model[2] or probabilistic models[3–5] do not use the documents themselves but use a kind of document surrogate, usually in the form of vectors of terms or keywords, which try to characterize the document's information content.[a] Queries are also represented in the same way.

[*] e-mail: acid@decsai.ugr.es.
[†] Author to whom all correspondence should be addressed: e-mail: lci@decsai.ugr.es.
[‡] e-mail: jmfluna@ujaen.es.
[§] e-mail: jhg@decsai.ugr.es.

[a] In the rest of the article we will use the word *document* to denote both documents and document surrogates.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 18, 251–265 (2003)
© 2003 Wiley Periodicals, Inc. Published online in Wiley InterScience
(www.interscience.wiley.com). • DOI 10.1002/int.10088

Probabilistic IR models use probability theory to deal with the intrinsic uncertainty with which IR is pervaded.[6] Also founded primarily on probabilistic methods, Bayesian networks[7] (BNs) have been proven to be a good model to manage uncertainty, even in the IR environment, where they already have been successfully applied as an extension/modification of probabilistic IR models.[8,9]

In this study we introduce new IR models based on BNs. The networks are used to compute the posterior probabilities of relevance of the documents in the collection given a query and then the documents are shown to the user in decreasing order of probability (according to the probability ranking principle[10]). The first model is composed of a simple BN with two layers of nodes representing the documents and the terms in the document collection, as well as the relationships among each other. The second model extends the first one by automatically including a third document layer, which tries to capture the strongest relationships between documents, with the aim of improving retrieval effectiveness.

The remainder of this article is structured as follows. In Section 2 we briefly describe some concepts and methods about BNs and IR that are necessary for the rest of the study. Section 3 introduces the basic model and the assumptions that determine the simple network topology being considered, as well as details about the estimation of the probability distributions to be stored in the network and the specific exact probabilities propagation algorithm, designed to allow efficient inference and retrieval. In Section 4 we study the extended model and two different ways to build the relationships between documents. The modified inference process necessary to deal efficiently with the new network topology is also described. Section 5 shows the experimental results obtained with our models using several standard document collections and including an empirical comparison with two well-known IR models. Finally, Section 6 contains the concluding remarks.

2. PRELIMINARIES

The representation of documents and queries in an IR system usually is based on term-weight vectors (with a zero weight meaning that the corresponding term is not included in a given document/query). These term weights could reflect different measures. The most common weighting schemes try to highlight the importance of each term, either within a given document it belongs to,[b] or within the entire collection.[c] A representative of the first scheme is *term frequency* (within-document frequency), tf_{ij}, i.e., the number of times that the ith term appears in the jth document. The second weighting scheme is based on the *inverse document frequency*, idf_i, of the ith term in the collection, which is defined as $\mathrm{idf}_i = \lg(N/n_i) + 1$, where N is the number of documents in the collection and n_i is the number of documents that contain the ith term. The combination of both weights, $\mathrm{tf}_{ij} \cdot \mathrm{idf}_i$, is also a common weighting scheme.

To evaluate IR systems, in terms of retrieval effectiveness, several measures have been proposed. The most commonly used are *recall* (R) (the proportion of

[b] Terms appearing frequently in a document are more important.

[c] More weight is given to terms that are rare (scarce) within a collection.

relevant documents retrieved), and *precision* (P) (the proportion of retrieved documents that are relevant for a given query). For test collections, the relevance or irrelevance of a document is based on the *relevance judgments* expressed by experts for a fixed set of queries.[2] By computing the precision for a number of values of recall we obtain a recall-precision plot. If a single measure of performance is desired, the average precision for all the recall values considered may be used. Finally, if we are processing together a set of queries, the usual approach is to report mean values of the selected performance measure(s).

BNs are graphic models able to represent and manipulate efficiently n-dimensional probability distributions.[7] A BN uses two components to codify qualitative and quantitative knowledge: (a) A directed acyclic graph [(DAG); $G = (V, E)$], where the nodes in $V = \{X_1, \ldots, X_n\}$ represent the random variables from the problem we want to solve, and the topology of the graph (the arcs in E) encodes conditional (in)dependence relationships among the variables (by means of the presence or absence of direct connections between pairs of variables); (b) a set of conditional probability distributions drawn from the graph structure, where for each variable $X_i \in V$ we have a family of conditional probability distributions $P(X_i|pa(X_i))$, where $pa(X_i)$ is any combination of the values of the variables in $Pa(X_i)$ (the parent set of X_i in G).

3. THE SIMPLE BN MODEL

The set of variables V_B in our basic BN G_B is composed of two different sets of variables $V_B = \mathcal{T} \cup \mathcal{D}$: The set $\mathcal{T} = \{T_1, \ldots, T_M\}$ of the M terms in the glossary (index) from a given collection and the set $\mathcal{D} = \{D_1, \ldots, D_N\}$ of the N documents that compose the collection.[d] Each term variable T_i is a binary random variable taking values in the set $\{\bar{t}_i, t_i\}$, where $\bar{t}_i$ stands for "the term T_i is not relevant," and t_i represents "the term T_i is relevant." Similarly, the domain of each document variable D_j is the set $\{\bar{d}_j, d_j\}$, where in this case, $\bar{d}_j$ and d_j, respectively, mean "the document D_j is not relevant for a given query" and "the document D_j is relevant for a given query."

3.1. Network Topology

To determine the topology of the basic BN, we have taken into account the following guidelines.[11]

- There is a link joining each term node $T_i \in \mathcal{T}$ and each document node $D_j \in \mathcal{D}$ whenever T_i belongs to D_j. This simply reflects the dependence between the (ir)relevance values of a document and the terms used to index it.
- There are no links joining any document nodes D_j and D_k. In other words, the dependence relationships between documents are not direct, they always depend on the terms included in these documents.
- Any document D_j is conditionally independent on any other document D_k when we

[d] We will use the notation T_i (D_j, respectively) to refer to both the term (document, respectively) and its associated variable and node.

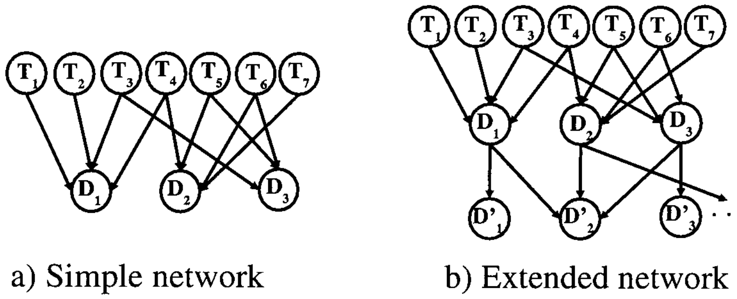

Figure 1. Two IR models based on BNs.

know for sure the (ir)relevance values for all the terms indexing D_j. This means that the degree of relevance of a document D_j for a given query can be determined completely by knowing the relevance status of all the terms that compose it. In the absence of this information, knowledge about the relevance or irrelevance of some other document D_k for the same query could have an influence on D_j.

These three assumptions determine the network structure in part; the links joining term and document nodes have to be directed from terms to documents; moreover, the parent set of a document node D_j is the set of term nodes that belong to D_j, i.e., $Pa(D_j) = \{T_i \in \mathcal{T} | T_i \in D_j\}$. The following is an additional assumption to determine completely the network topology: terms are marginally independent among each other. This forces the absence of links between term nodes (all of them are root nodes). Figure 1(a) displays the structure of our simple BN model. Note that we have obtained a bipartite graph, which is a particular case of the so-called simple graphs.[12,13] This specific topology allows us the use of a very fast propagation algorithm, as we shall see later.

3.2. Probability Distributions

The estimation of the probability distributions stored in term and document nodes is carried out in the following ways:

- Term nodes. Because all the terms are root nodes, we need to store marginal distributions. We use the following estimator, for every term T_i: $p(t_i) = 1/M$ and $p(\bar{t}_i) = (M - 1)/M$, where M is the number of terms in the collection. So, the prior probability of relevance of any term is very small, inversely proportional to the size of the index.
- Document nodes. The estimation of the conditional probabilities of relevance of a document D_j, $p(d_j|pa(D_j))$, is not an easy problem. The reason is that the number of conditional probabilities that we need to estimate and store for each D_j grows exponentially with the number of parents of D_j. For example, if D_j has been indexed with 40 terms (and this may be a quite common situation), we need 2^{40} probability distributions; hence, we cannot use a standard approach. Instead of explicitly computing and storing these probabilities, we use a probability function (also called a canonical model of multicausal interaction[7]). Each time that a given conditional probability is required during the inference process, the probability function will compute and return the appropriate value. We have developed a new general canonical model: for any

configuration $pa(D_j)$ of $\mathrm{Pa}(D_j)$ (i.e., any assignment of values to all the term variables in D_j), we define the conditional probability of relevance of D_j as follows:

$$p(d_j|pa(D_j)) = \sum_{T_i \in D_j,\, t_i \in pa(D_j)} w_{ij} \tag{1}$$

The weights w_{ij} verify that $0 \leqslant w_{ij}\ \forall i, j$ and $\Sigma_{T_i \in D_j} w_{ij} \leqslant 1\ \forall j$. The expression $t_i \in pa(D_j)$ in Equation 1 means that we only include in the sum those weights w_{ij} such that the value assigned to the corresponding term T_i in the configuration $pa(D_j)$ is t_i. So, the more terms that are relevant in $pa(D_j)$ the greater is the probability of relevance of D_j. The specific weights for this general canonical model that we use in this study will be described in Section 5.

3.3. Inference and Retrieval

Given a query Q, the retrieval process starts placing the evidences in the term subnetwork: the state of each term T_{iQ} belonging to Q is fixed to t_{iQ} (relevant). Then, the inference process is run, obtaining for each document D_j its probability of relevance given that the terms in the query are also relevant $p(d_j|Q)$. Finally, the documents are sorted in decreasing order of probability to carry out the evaluation process.

However, we have to remark that even for small document collections, the BN will contain thousands of nodes and many nodes having a great number of parents. Moreover, although the network topology is relatively simple, it contains cycles, referred to the undirected graph associated with the structure of the BN.[e] Therefore, general purpose propagation algorithms cannot be applied because of efficiency considerations. For that reason, we have designed a specific inference process that takes advantage of both the topology of the network and the kind of probability function used for document nodes in Equation 1; the propagation process is replaced by a single evaluation for each document node, but ensuring that the results are the same as the ones obtained using exact propagation in the entire network[f]:

$$p(d_j|Q) = \sum_{T_i \in D_j} w_{ij}\, p(t_i|Q) \tag{2}$$

Moreover, as term nodes are marginally independent, $p(t_i|Q) = 1$ if $T_i \in Q$ and $p(t_i|Q) = 1/M$ if $T_i \notin Q$. Therefore, the computation of $p(d_j|Q)$ can be carried out as follows:

$$p(d_j|Q) = \sum_{T_i \in D_j \cap Q} w_{ij} + \frac{1}{M} \sum_{T_i \in D_j \setminus Q} w_{ij} = \frac{1}{M} \sum_{T_i \in D_j} w_{ij} + \frac{M-1}{M} \sum_{T_i \in D_j \cap Q} w_{ij} \tag{3}$$

Note that the prior probability of relevance is $p(d_j) = 1/M\ \Sigma_{T_i \in D_j} w_{ij}$, so that $p(d_j|Q)$ is always $\geqslant p(d_j)$. A simple modification of this model takes into account

[e] Hence, we cannot use propagation algorithms for singly connected networks, such as trees or polytrees.[7]

[f] Because of space limitations, we do not include the proof of this assertion; it can be found in Refs. 11 and 14.

the information about the frequency of the terms in the query Q, qf_i, to give more importance to the terms used more frequently (as is usual in other IR models). This can be done by duplicating qf_i times in the network for each term T_i appearing in the query. Then, Equation 3 is transformed into

$$p(d_j|Q) = \sum_{T_i \in D_j \cap Q} w_{ij}\,\mathrm{qf}_i + \frac{1}{M} \sum_{T_i \in D_j \setminus Q} w_{ij} \tag{4}$$

4. THE EXTENDED BN MODEL

In the simple BN (SBN) model, if a document D_j does not contain any of the terms used to formulate the query, $D_j \cap Q = \emptyset$, almost surely, it will not be retrieved, even in the case in which these terms are related (in some way) to the ones indexing the document [its probability will remain unchanged, $p(d_j|Q) = p(d_j) = 1/M\ \Sigma_{T_i \in D_j} w_{ij}$]. The reason is that document nodes are related only through terms in common. One approach to deal with this situation could be to include arcs modeling direct relationships between terms.[15] Using these relationships, the instantiation of the query terms would increase the probability of relevance of other terms, which in turn would increase the probability of relevance of some documents containing them. However, these new arcs would make the term subnetwork more complex, thus rendering the inference process more difficult. In this study we use a different approach; we include in the model direct relationships between documents, which will play a role similar to the clustering techniques used in other IR models.[2,16] These relationships will be based on measuring (asymmetric) similarities between documents by estimating conditional probabilities of relevance of every document given that another document is considered relevant. The SBN with two layers described previously will be used to compute these probabilities.

Therefore, let $e(D_i)$ be an event representing some type of evidence about the relevance of a document D_i. If, for a given document D_j, we compute the probabilities $p(d_j|e(D_i))\ \forall D_i \in \mathcal{D}$, then those documents D_i producing the greatest values of $p(d_j|e(D_i))$ are the ones that are more related with D_j (because D_j has a high probability of being relevant when we obtain some evidence about the relevance of D_i). After ranking, for each D_j, all the documents in $\mathcal{D}$ in decreasing ordering of $p(d_j|e(D_i))$, let $R_c(D_j)$ be the set of the top c documents. We are going to include in the network explicit dependence relationships between D_j and each document $D_i \in R_c(D_j)$.

However, instead of using a document subnetwork with one layer, including arcs from the documents in $R_c(D_j)$ to document D_j, we use two layers; we duplicate each document node D_k in the original layer to obtain another document node D'_k, thus forming a new document layer, and the arcs connecting the two layers go from $D_i \in R_c(D_j)$ to D'_j. So, in the extended BN (EBN), G_E, the set of variables is $V_E = \mathcal{T} \cup \mathcal{D} \cup \mathcal{D}'$. The parent set of any duplicate document node $D'_j \in \mathcal{D}'$ is defined as $Pa(D'_j) = R_c(D_j)$. We use this topology for two reasons:

(1) The basic BN G_B remains as a subgraph of G_E induced by the subset of nodes $\mathcal{T} \cup \mathcal{D}$; therefore, we do not have to redefine the conditional probabilities associated with the document nodes in $\mathcal{D}$ (Equation 1); (2) the new topology contains three simple layers [See Figure 1(b)], without connections between the nodes in the same layer, and this fact will be essential for the efficiency of the inference process.

To complete the specification of the EBN, we have to define the conditional probabilities $p(d_j'|pa(D_j'))$ for the documents in the second document layer $\mathcal{D}'$. We use a probability function belonging to the general canonical model defined in Equation 1, where the weights w_{ij} are defined as $w_{ij} = p(d_j|e(D_i))/S_j$, i.e;

$$p(d_j'|pa(D_j')) = \frac{1}{S_j} \sum_{\substack{D_i \in Pa(D_j') \\ d_i \in pa(D_j')}} p(d_j|e(D_i)) \tag{5}$$

where $S_j = \Sigma_{D_k \in Pa(D_j')} p(d_j|e(D_k))$, and the values $p(d_j|e(D_i))$ will be obtained, using the SBN, during the building process of the EBN.

Now, the final probability values used in the retrieval process will be $p(d_j'|Q)$ instead of $p(d_j|Q)$. To compute $p(d_j'|Q)$, we can again take advantage of the layered topology and Equation 5 to replace the propagation process in the whole network by the following evaluation[1,4]:

$$p(d_j'|Q) = \frac{1}{S_j} \sum_{D_i \in Pa(D_j')} p(d_j|e(D_i))p(d_i|Q) \tag{6}$$

where the probabilities $p(d_i|Q)$ are computed according to Equations 3 or 4. Note that Equation 6 is the counterpart for the new document layer of Equation 2. The values $p(d_j'|Q)$ measure the relevance of the documents by combining both the contribution of the query itself and the document relationships.

The following are the remaining tasks: first, to give more precise semantics to the events $e(D_i)$, which so far only refer vaguely to the relevance of documents D_i; second, to explain how to calculate efficiently the values $p(d_j|e(D_i))$. We have considered two different ways to give meaning to the events $e(D_i)$. On the one hand, to think about $e(D_i)$ as the event $[T_k = t_k, \forall T_k \in D_i]$, i.e., document D_i is relevant as long as its index terms are all relevant; and then, computing $p(d_j|e(D_i))$ is equivalent to introduce a query containing only all the terms indexing document D_i and use the SBN inference formula (Equation 3) to compute $p(d_j|Q_i)$. In this case we obtain

$$p(d_j|e(D_i)) = \frac{1}{M} \sum_{T_k \in D_j} w_{kj} + \frac{M-1}{M} \sum_{T_k \in D_j \cap D_i} w_{kj} \tag{7}$$

On the other hand, we can also consider that $e(D_i)$ means exactly that the document D_i is relevant, which corresponds to the event $[D_i = d_i]$. In this case, it can be proved (See Ref. 14) that $p(d_j|e(D_i)) = p(d_j|d_i)$ can be calculated (without explicit propagation) by simply evaluating the following expression:

$$p(d_j|e(D_i)) = \frac{1}{M}\sum_{T_k\in D_j} w_{kj} + \frac{M-1}{M}\left(\frac{\sum_{T_k\in D_j\cap D_i} w_{kj}w_{ki}}{\sum_{T_h\in D_i} w_{hi}}\right) \quad \forall i \neq j, \quad p(d_j|e(D_j)) = 1 \tag{8}$$

The first sum in Equations 7 and 8 does not depend on D_i. Therefore, given a document D_j, in order to select the c parents of its copy D'_j in the second document layer, in either case we only have to select the c documents D_i with the greatest values of

$$\text{(a)} \quad \sum_{T_k\in D_j\cap D_i} w_{kj} \quad \text{or} \quad \text{(b)}\,\frac{\sum_{T_k\in D_j\cap D_i} w_{kj}w_{ki}}{\sum_{T_h\in D_i} w_{hi}} \tag{9}$$

for the model based on Equations 7 and 8, respectively. Equation 9 means that the more terms the documents D_j and D_i have in common and the more important these terms are, the more closely related is D_i to D_j. So, the probabilistic similarity between documents derived from our models closely matches standard definitions of association or similarity measures.[1] However, it should be noted that our similarity is asymmetrical. The main difference between the two models is the use of a normalization factor in Equation 9(b), to penalize those documents D_i having many terms, and the lack of normalization in Equation 9(a).

A detailed comparison of our SBN and EBN models with the two main existing retrieval models also based on BNs [the inference network[9] (IN) and the belief network[8] models], highlighting the differences in network structure, conditional probability distributions and inference method being used, can be found in Ref. 14.

5. EXPERIMENTAL RESULTS

To test the performance of the two new BN-based retrieval models, we have used five well-known document collections: ADI, CACM, CISI, CRANFIELD, and MEDLARS. The main characteristics of these collections with respect to number of documents, terms, and queries are (in this ordering) ADI (82, 828, and 35, respectively), CACM (3,204; 7,562; and 52, respectively), CISI (1,460; 4,985; and 76, respectively), CRANFIELD (1,398; 3,857; and 225, respectively), and MEDLARS (1,033; 7,170; and 30, respectively). The results obtained by our models will be compared with the ones obtained by two well-known IR systems: SMART[2,g] (SM) and the IN model.[h] The performance measure considered is the average precision for the 11 standard values of recall (denoted AP-11).

In our experiments the specific weights w_{ij}, used by our models (See Equation 1), for each document $D_j \in \mathcal{D}$ and each term $T_i \in D_j$, are

[g] We used the implementation of this IR system available at the Computer Science Department of Cornell University, using the *ntc* weighting scheme.

[h] In this case, we have built our own implementation, and we used the configuration parameters proposed by Turtle in Ref. 9: $p(t_i|d_j = \text{true}) = 0.4 + 0.6 * \text{tf} * \text{idf}$ and $p(t_i|\text{all parents false}) = 0.3$.

Table I. Average precision values for SM and IN.

	ADI	CISI	CRAN	MED	CACM
SMART	0.4706	0.2459	0.4294	0.5446	0.3768
Inf. Network	0.4612	0.2498	0.4367	0.5534	0.3974

$$w_{ij} = \alpha^{-1} \frac{\mathrm{tf}_{ij} \cdot \mathrm{idf}_i^2}{\sqrt{\sum_{T_k \in D_j} \mathrm{tf}_{kj} \cdot \mathrm{idf}_k^2}} \tag{10}$$

where α is a normalizing constant (to assure that $\sum_{T_i \in D_j} w_{ij} \leq 1 \ \forall D_j \in \mathcal{D}$). Obviously, many other weighting schemes are possible. The weights in Equation 10 have been chosen to resemble the well-known cosine measure.[2]

Table I displays the AP-11 values obtained by SM and the IN for all the test collections. The results for the experiments with the SBN are shown in Table II. The columns corresponding to the experiments that use Equation 3 are labeled with "1" and the ones that use Equation 4 are labeled with "qf." The rows labeled with %SM and %IN show the percentage of change of the performance measure obtained by our methods with respect to SM and the IN, respectively.

The results in Tables I and II show that SBN can compete with SM and IN: the AP-11 values are quite similar (very low percentages of change), except in the case of CISI and CACM. For CISI, SBN performs remarkably better than SM and IN, whereas the opposite is true for CACM. With respect to the use of Equation 4, i.e., the frequency qf of the terms in the query, instead of Equation 3, none of the two methods is clearly preferable to the other; for three collections (CRANFIELD, MEDLARS, and CACM), the best results are obtained without using qf, whereas for the other two collections (ADI and CISI), the use of qf improves the results. We conjecture that this behavior is caused by the specific characteristics of each collection. Anyway, the differences between the two methods are rather small, except in the case of CISI, where the results are remarkably better using qf (perhaps the explanation may be that the qf values for CISI are considerably larger than for the other collections).

With respect to the EBN, we have carried out experiments with the two schemes for measuring the document relationships: the model using Equation 7 (EBNa) and the model using Equation 8 (EBNb). In both cases, three different values for the number c of document nodes in the first document layer that are parents of the nodes in the second document layer have been used: $c = 5, 10, 15$.

Table II. Experiments with SBN.

	ADI		CISI		CRAN		MED		CACM	
	SBN 1	SBN qf	SBN 1	SBN qf	SBN 1	SBN qf	SBN 1	SBN qf	SBN 1	SBN qf
AP-11	0.4707	0.4709	0.2206	0.2642	0.4323	0.4309	0.5552	0.5458	0.3582	0.3435
%SM	+0.02	+0.06	−10.29	+7.44	+0.68	+0.35	+1.95	+0.22	−4.94	−8.84
%IN	+2.06	+2.10	−11.69	+5.76	−1.01	−1.33	+0.33	−1.37	−9.86	−13.56

Table III displays the results for EBNb. If we compare these values with those in Table II, we can see that the extended network EBNb systematically improves the results of SBN, showing that to take into account document interrelationships may be a good idea. It can also be observed that, except in the case of ADI, the AP-11 values obtained by EBNb increase as the parameter c increases. However, the differences between SBN and EBNb are so small (with the exception of CACM using qf), that it could question the usefulness of increasing the complexity of the BN retrieval model by including the new document layer (which implies the necessity of precomputing the probabilities $p(d_j|e(D_i)) = p(d_j|d_i)$). For each document D_j, after analyzing the values $p(d_j|d_i)$, we realized that even the greatest values of $p(d_j|d_i)\ \forall i \neq j$ are extremely low compared with $p(d_j|d_j) = 1$ (typically $p(d_j|d_i) \approx 0.0025$). This fact could be the cause of the minimum improvement obtained by EBNb with respect to SBN, because the value $p(d_j|Q) = p(d_j|d_j)p(d_j|Q)$ dominates completely the other components in Equation 6, $\sum_{D_i \in Pa(D_j'), D_i \neq D_j} p(d_j|d_i)p(d_i|Q)$, and therefore the ranking of documents obtained by using Equation 6 would be almost the same as the one obtained by the SBN model [which only uses $p(d_j|Q)$].

To verify the truthfulness of this conjecture and, in that case, trying to overcome the problem, we have modified the probability function defined in Equation 5 (only for EBNb) to reduce the importance of the term $p(d_j|Q)$ in the computation of $p(d_j'|Q)$. The new probability function $p(d_j'|pa(D_j'))$, which is also of the canonical type defined in Equation 1, is the following:

$$p(d_j'|pa(D_j')) = \begin{cases} \dfrac{1-\beta}{S_j - 1} \displaystyle\sum_{\substack{D_i \in Pa(D_j') \\ d_i \in pa(D_j') \\ D_i \neq D_j}} p(d_j|d_i) & \text{if } d_j \notin pa(D_j') \\ \dfrac{1-\beta}{S_{j-1}} \displaystyle\sum_{\substack{D_i \in Pa(D_j') \\ d_i \in pa(D_j') \\ D_i \neq D_j}} p(d_j|d_i) + \beta & \text{if } d_j \in pa(D_j') \end{cases} \tag{11}$$

where the parameter β controls the importance of the contribution of the document relationships being considered for document D_j to its final degree of relevance. Again, taking advantage of the layered topology, we can compute $p(d_j'|Q)$ as follows:

$$p(d_j'|Q) = \frac{1-\beta}{S_j - 1} \sum_{\substack{D_i \in Pa(D_j') \\ D_i \neq D_j}} p(d_j|d_i)p(d_i|Q) + \beta p(d_j|Q) \tag{12}$$

The results obtained by EBNb using Equation 12 instead of Equation 6, with a value $\beta = 0.5$, are displayed in Table IV.

The results obtained in Table IV clearly represent a remarkable improvement with respect to the ones in Table III, thus showing that the use of the document relationships is quite useful, provided that the weights measuring the strength of these relationships are set appropriately. The exception is CACM, where the new results are worse. We have also carried out some other experiments with different values for the parameter β and, generally, the results are quite similar to the ones

Table III. Experiments with the EBN using Equation 8 (EBNb).

	ADI		CISI		CRAN		MED		CACM	
	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf
c = 5										
AP-11	0.4724	0.4728	0.2211	0.2639	0.4331	0.4318	0.5651	0.5551	0.3630	0.3823
%SM	+0.38	+0.47	−10.09	+7.32	+0.86	+0.56	+3.95	+1.93	−3.66	+1.46
%IN	+2.43	+2.52	−11.49	+5.64	−0.82	−1.12	+2.29	+0.55	−8.66	−3.80
c = 10										
AP-11	0.4717	0.4719	0.2221	0.2650	0.4333	0.4321	0.5687	0.5580	0.3636	0.3827
%SM	+0.23	+0.28	−9.68	+7.77	+0.91	+0.63	+4.43	+2.46	−3.50	+1.57
%IN	+2.28	+2.32	−11.09	+6.08	−0.78	−1.05	+2.76	+0.83	−8.51	−3.70
c = 15										
AP-11	0.4715	0.4716	0.2223	0.2651	0.4332	0.4323	0.5708	0.5598	0.3629	0.3838
%SM	+0.19	+0.21	−9.60	+7.81	+0.89	+0.68	+4.81	+2.79	−3.69	+1.86
%IN	+2.23	+2.26	−11.01	+6.12	−0.80	−1.01	+3.14	+1.16	−8.68	−3.42

Table IV. Experiments with the EBNb model using Equation 12 and $\beta = 0.5$.

	ADI		CISI		CRAN		MED		CACM	
	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf	EBNb 1	EBNb qf
$c = 5$										
AP-11	0.4871	0.4892	0.2352	0.2715	0.4581	0.4550	0.6556	0.6394	0.3345	0.3560
%SM	+3.51	+3.95	−4.35	+10.41	+6.68	+5.96	+20.38	+17.41	−11.23	−5.52
%IN	+5.62	+6.07	−5.84	+8.69	+4.90	+4.19	+18.47	+15.54	−15.83	−10.42
$c = 10$										
AP-11	0.4788	0.4831	0.2393	0.2780	0.4612	0.4577	0.6620	0.6488	0.3486	0.3772
%SM	+1.74	+2.66	−2.68	+13.05	+7.41	+6.59	+21.56	+19.13	−7.48	+0.11
%IN	+3.82	+4.75	−4.20	+11.29	+5.61	+4.81	+19.62	+17.24	−12.28	−5.08
$c = 15$										
AP-11	0.4745	0.4780	0.2390	0.2786	0.4591	0.4561	0.6652	0.6502	0.3564	0.3772
%SM	+0.83	+1.57	−2.81	+13.30	+6.92	+6.22	+22.14	+19.39	−5.41	+0.11
%IN	+2.88	+3.64	−4.32	+11.53	+5.13	+4.44	+20.20	+17.49	−10.32	−5.08

Table V. Experiments with the EBN using Equation 7 (EBNa).

	ADI		CISI		CRAN		MED		CACM	
	EBNa 1	EBNa qf	EBNa 1	EBNa qf	EBNa 1	EBNa qf	EBNa 1	EBNa qf	EBNa 1	EBNa qf
$c = 5$										
AP-11	0.4873	0.4885	0.2313	0.2738	0.4767	0.4760	0.6563	0.6422	0.3715	0.3917
%SM	+3.55	+3.80	−5.94	+11.35	+11.02	+10.85	+20.51	+17.92	−1.41	+3.95
%IN	+5.66	+5.92	−7.41	+9.61	+9.16	+9.00	+18.59	+16.05	−6.52	−1.43
$c = 10$										
AP-11	0.4879	0.4863	0.2407	0.2827	0.4854	0.4802	0.6764	0.6664	0.3770	0.3826
%SM	+3.68	+3.34	−2.11	+14.97	+13.04	+11.83	+24.20	+22.37	+0.05	+1.54
%IN	+5.79	+5.44	−3.64	+13.17	+11.15	+9.96	+22.23	+20.42	−5.13	−3.72
$c = 15$										
AP-11	0.4796	0.4832	0.2404	0.2832	0.4734	0.4701	0.6900	0.6784	0.3777	0.3887
%SM	+1.91	+2.68	−2.24	+15.17	+10.25	+9.48	+26.70	+24.57	+0.24	+3.16
%IN	+3.99	+4.77	−3.76	+13.37	+8.40	+7.65	+24.68	+22.59	−4.96	−2.19

displayed in Table IV, for values $\beta \leqslant 0.5$, whereas the performance decreases for higher values of β (again, except for CACM, where values of β higher than 0.5 improve the results).

Finally, the results obtained using the other version of the EBN (EBNa) are displayed in Table V. We can observe that this approach produces excellent results: it improves systematically the results of EBNb and it is also simpler than EBNb (the computation of Equation 7 is more efficient than Equation 8). Moreover, in this case it is not necessary to modify Equation 6. Except in the case of CACM, EBNa improves considerably the results obtained by SM and the IN.

6. CONCLUSIONS

In this article we have presented two new IR models based on BNs. The first model SBN is composed of a layer of term nodes and a layer of document nodes, joining each term node to the document nodes representing the documents indexed by this term. This model has been endowed with an inference mechanism that allows us to perform exact propagation in the whole network efficiently. The experimental results obtained with five collections show that this model is competitive with respect to SM and the IN model.

This initial model has been enriched, establishing the most important relationships among documents and thus increasing the expressiveness of SBN and giving rise to an extended model with three layers (EBN). In this second approach, we have shown two mechanisms by which the document relationships are captured. The new inference method, also exact, is composed of two stages: a propagation in the original network SBN, and the combination of this information with that one stored in the second document layer, updating the probability of relevance of each document given a query with the strength of the relationships among the documents. The results show an improvement in the performance of the EBN model, revealing the suitability of the document layer extension.

As future works, we plan to study new probability functions in the second document layer, to more accurately combine the information about the relevance of the documents given the query and the strength of the document relationships. The method used to select the most important document relationships could also be modified: instead of always selecting a fixed number c of parents for each document in the second document layer, we could design a more flexible method to determine the number of parents for each document. Moreover, we want to use the weighting schemes as an alternative to the one considered in Equation 10 to define the conditional probabilities of the first document layer. We are also planning to extend our model to cope with boolean queries.

On the other hand, we have tested our models with some standard test collections, in which the sizes are smaller than actual collections. Our objective has been just to determine the validity of the proposed models for IR, focusing our attention only in modeling aspects. Experimentation with Text REtrieval Conference (TREC) collections[17] will be one of the most important points in which we are going to center our future research. The basic objective will be to determine the efficiency and effectiveness of our models with these collections. This task could suggest some modifications or refinements in our models related to the propagation and construction of the second document layer.

Acknowledgments

This work was supported by the Spanish Comisión Interministerial de Ciencia y Tecnología (CICYT) under Project TIC2000-1351.

References

1. Frakes WB, Baeza-Yates R (editors). Information retrieval. Data structures and algorithms. Upper Saddle River, NJ: Prentice Hall; 1992.
2. Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw-Hill; 1983.
3. Maron ME, Kuhns JL. On relevance, probabilistic indexing, and information retrieval. J ACM 1960;7:216–244.
4. Robertson SE, Sparck Jones K. Relevance weighting of search terms. J Am Soc Inform Sci 1976;27:129–146.
5. Sparck Jones K, Walker S, Robertson SE. A probabilistic model of information retrieval: Development and comparative experiments Part 1. Inform Process Manage 2000;36:779–808.
6. Fung R, Favero BD. Applying Bayesian networks to information retrieval. Commun ACM 1995;38(2):42–57.
7. Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible inference. San Mateo, CA: Morgan and Kaufmann; 1988.
8. Reis Silva I. Bayesian networks for information retrieval systems, PhD thesis, Universidad Federal de Minas Gerais, 2000.
9. Turtle HR. Inference networks for document retrieval, PhD thesis, Computer and Information Science Dept, University of Massachusetts, 1990.
10. Robertson SE. The probability ranking principle in IR. J Document 1977;33:294–304.
11. Fernández-Luna JM. Modelos de Recuperación de Información Basados en Redes de Creencia (in Spanish) PhD thesis, Universidad de Granada, 2001.
12. de Campos LM, Huete JF. On the use of independence relationships for learning simplified belief networks. Int J Intell Syst 1997;12:495–522.
13. Geiger D, Paz A, Pearl J. Learning simple causal structures. Int J Intell Syst 1993;8:231–247.
14. Acid S, de Campos LM, Fernández-Luna JM, Huete JF. An information retrieval model based on simple Bayesian networks. DECSAI Technical Report n. 020201, 2002.
15. de Campos LM, Fernández-Luna JM, Huete JF. Building Bayesian network-based information retrieval systems. In: 2nd Workshop on Logical and Uncertainty Models for Information Systems (LUMIS), London, 2000. pp 543–552.
16. van Rijsbergen CJ. Information retrieval, 2nd Ed. London: Butter Worths; 1979.
17. Voorhees EM, Harman D. Overview of the 9th Text REtrieval Conference (TREC-9). In: Voorhees EM, Harman D, editors. Proc 9th Text REtrieval Conf, NIST Special Publication 500-249; 2000. pp 1–13.