# Chapter 7.30
# Retrieving Medical Records Using Bayesian Networks

**Luis M. de Campos**
*Universidad de Granada, Spain*

**Juan M. Fernández Luna**
*Universidad de Granada, Spain*

**Juan F. Huete**
*Universidad de Granada, Spain*

## INTRODUCTION

Bayesian networks (Jensen, 2001) are powerful tools for dealing with uncertainty. They have been successfully applied in a wide range of domains where this property is an important feature, as in the case of information retrieval (IR) (Turtle & Croft, 1991). This field (Baeza-Yates & Ribeiro-Neto, 1999) is concerned with the representation, storage, organization, and accessing of information items (the textual representation of any kind of object). Uncertainty is also present in this field, and, consequently, several approaches based on these probabilistic graphical models have been designed in an attempt to represent documents and their contents (expressed by means of indexed terms), and the relationships between them, so as to retrieve as many relevant documents as possible, given a query submitted by a user.

Classic IR has evolved from flat documents (i.e., texts that do not have any kind of structure relating their contents) with all the indexing terms directly assigned to the document itself toward structured information retrieval (SIR) (Chiaramella, 2001), where the structure or the hierarchy of contents of a document is taken into account. For instance, a book can be divided into chapters, each chapter into sections, each section into paragraphs, and so on. Terms could be assigned to any of the parts where they occur. New standards, such as SGML or XML, have been developed to represent this type of document. Bayesian network models also have been extended to deal with this new kind of document.

In this article, a structured information retrieval application in the domain of a pathological anatomy service is presented. All the medical records that this service stores are represented

in XML, and our contribution involves retrieving records that are relevant for a given query that could be formulated by a Boolean expression on some fields, as well as using a text-free query on other different fields. The search engine that answers this second type of query is based on Bayesian networks.

## BACKGROUND

Probabilistic retrieval models (Crestani et al., 1998) were designed in the early stages of this discipline to retrieve those documents relevant to a given query, computing the probability of relevance. The development of Bayesian networks and their successful application to real problems has caused several researchers in the field of IR to focus their attention on them as an evolution of probabilistic models. They realized that this kind of network model could be suitable for use in IR, specially designed to perform extremely well in environments where uncertainty is a very important feature, as is the case of IR, and also because they can properly represent the relationships between variables.

Bayesian networks are graphical models that are capable of representing and efficiently manipulating n-dimensional probability distributions. They use two components to codify qualitative and quantitative knowledge, respectively: first, a directed acyclic graph (DAG), G=(V,E), where the nodes in V represent the random variables from the problem we want to solve, and set E contains the arcs that join the nodes. The topology of the graph (the arcs in E) encodes conditional (in)dependence relationships between the variables (by means of the presence or absence of direct connections between pairs of variables); and second, a set of conditional distributions drawn from the graph structure. For each variable $X_i \in V$, we therefore have a family of conditional probability distributions $P(X_i \mid pa(X_i))$, where $pa(X_i)$ represents any combination of the values of the variables in $Pa(X_i)$, and $Pa(X_i)$ is the parent set of $X_i$ in G. From these conditional distributions, we can recover the joint distribution over V.

This decomposition of the joint distribution gives rise to important savings in storage requirements. In many cases, it also enables probabilistic inference (propagation) to be performed efficiently (i.e., to compute the posterior probability for any variable, given some evidence about the values of other variables in the graph).

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i))$$

The first complete IR model based on Bayesian networks was the Inference Network Model (Turtle & Croft, 1991). Subsequently, two new models were developed: the Belief Network Model (Calado et al., 2001; Reis, 2000) and the Bayesian Network Retrieval Model (de Campos et al., 2003, 2003b, 2003c, 2003d). Of course, not only have complete models been developed in the IR context, but also solutions to specific problems (Dumais, et al., 1998; Tsikrika & Lalmas, 2002; Wong & Butz, 2000).

Structural document representation requires IR to design and implement new models and tools to index, retrieve, and present documents according to the given document structure. Models such as the previously mentioned Bayesian Network Retrieval Model have been adapted to cope with this new context (Crestani et al., 2003, 2003b), and others have been developed from scratch (Graves & Lalmas, 2002; Ludovic & Gallinari, 2003; Myaeng et al., 1998).

## MAIN THRUST

The main purpose of this article is to present the guidelines for construction and use of a Bayesian-network-based information retrieval system. The source document collection is a set of medical records about patients and their medical tests stored

in an XML database from a pathological anatomy service. By using XML tags, the information can be organized around a well-defined structure. Our hypothesis is that by using this structure, we will obtain retrieval results that better match the physicians' needs. Focusing on the structure of the documents, data are distributed between two different types of tags: on the one hand, we could consider fixed domain tags (i.e., those attributes from the medical record with a set of well-defined values, such as sex, birthdate, address, etc.); and on the other hand, free text passages are used by the physicians to write comments and descriptions about their particular perceptions of the tests that have been performed on the patients, as well as any conclusions that can be drawn from the results. In this case, there is no restriction on the information that can be stored. Three different free-text passages are considered, representing a description of the microscopic analysis, the macroscopic analysis, and the final diagnostic, respectively.

Physicians must be able to use queries that combine both fixed and free-text elements. For example, they might be interested in all documents concerning males who are suspected of having a malignant tumor. In order to tackle this problem, we propose a two-step process. First, a Boolean retrieval task is carried out in order to identify those records in the dataset, mapping the requirements of the fixed domain elements. The query is formulated by means of the XPath language. These records are then the inputs of a Bayesian retrieval process in the second stage, where they are sorted in decreasing order of their posterior probability of relevance to the query as the final output of the process.

## The Bayesian Network Model

Since, for those attributes related to fixed domains, it is sufficient to consider a Boolean retrieval, the Bayesian model will be used to represent both the structural and the content information related to free-text passages. In order to specify the topology of the model (a directed acyclic graph, representing dependence relationships), we need to determine which information components (variables) will be considered as relevant. In our case, we can distinguish between two different types of variables: the set that contains those terms used to index the free-text passages, $T = \{T_1,...,T_M\}$, with M being the total number of index terms used; and set $D$, representing the documents (medical records) in the collection. In this case, we consider as relevant variables the whole document $D_k$ and also the three subordinate documents that comprise it: macroscopic description, $D_{mk}$; microscopic description, $D_{\mu k}$; and final diagnostic, $D_{fk}$ (generically, any of these will be represented by $D_{\bullet k}$). Therefore,

$$D = \{D_1, D_{m1}, D_{\mu 1}, D_{f1},...,D_N, D_{mN}, D_{\mu N}, D_{fN}\},$$

with N being the number of documents that comprise the collection[1].

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. The domain of each document variable, $D_j$, is the set $\{\bar{d}_j, d_j\}$, where, in this case, $\bar{d}_j$ and $d_j$ mean the document $D_j$ is not relevant for a given query, and the document $D_j$ is relevant for the given query, respectively. A similar reasoning can be stated for any subordinate document, $D_{\bullet j}$.

In order to specify completely the model topology, we need to include those links representing the dependence relationships between variables. We can distinguish two types of nodes. The first type links between each term node $T_i \in T$ and each subordinate document node $D_{\bullet j} \in D$, whenever $T_i$ belongs to $D_{\bullet j}$. These links reflect the dependence between the (ir)relevance values of this document and the terms used to index it and will be directed from terms to documents. Therefore, the parent set of a document node $D_{\bullet j}$ is the set of term nodes that belong $D_{\bullet j}$ to (i.e., $Pa(D_{\bullet j}) = \{T_i \in T \mid T_i \in D_{\bullet j}\}$).

The second type links by connecting each subordinate document $D_{\bullet j}$ with the node document $D_j$ to which it belongs, reflecting the fact that the relevance of a document to a query will depend only on the relevance values of its subordinate documents. These links will be directed from subordinate to document nodes.

It should be noted that we do not use links between terms and documents, because we consider these to be independent, given that we know the relevance value of the subordinate documents. Consequently, we have designed a layered topology for our model that also represents the structural information of the medical records. Figure 1 displays the graph associated with the Bayesian network model.

## Probability Distributions

The following step to complete the design of the model is the estimation of the quantitative components of the Bayesian network (i.e., the probability distributions stored in each node). For term nodes, and taking into account that all terms are root nodes, marginal distributions need to be stored. The following estimator is used for every term $T_i$: $p(t_i) = (1/M)$ and $p(\bar{t}_i) = (M-1)/M$. Therefore, the prior probability of relevance of any term is very small and inversely proportional to the size of the index.

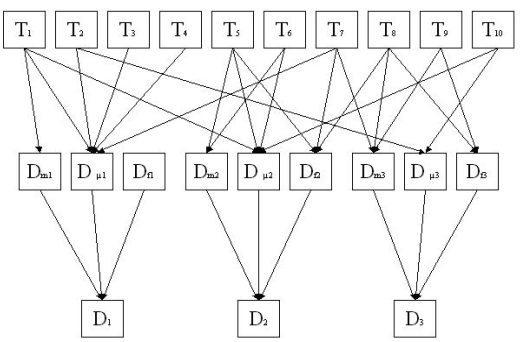
Considering now document and subordinate document nodes, it is necessary to assess a set of conditional probability distributions, the size of which grows exponentially with the number of parents. We therefore propose using a canonical model that represents the particular influence of each parent in the relevance of the node. In particular, given a variable $X_j$ (representing a document or a subordinate document node), the probability of relevance given a particular configuration of the parent set $pa(X_j)$ is computed by means of

$$p(x_j \mid pa(X_j)) = \sum_{i \in pa(X_j)} w_{ij},$$

where the expression $i \in pa(X_j)$ means that only those weights where the value assigned to the $i^{th}$ parent of $X_j$ in the configuration $pa(X_j)$ is relevant will be included in the sum. Therefore, the greater the number of relevant variables in $pa(X_j)$, the greater the probability of relevance of $X_j$.

The particular values of the weights $w_{ij}$ are first, for a subordinate document, $D_{\bullet j}$, $w_{ij} = (tf_{ij} idf_i^2)/(\sum_{T_k \in Pa(D_{\bullet j})} tf_{kj} idf_k^2)$, with $tf_{ij}$ being the frequency of the term $i^{th}$ in the subordinate document and $idf_i$ the inverse document frequency of the term $T_i$ in the whole collection; and second, for a document node, $D_j$, we use three factors $\alpha = w_{mj,j}$, $\beta = w_{\mu j,j}$ and $\delta = w_{fj,j}$, representing the influence of the macroscopic description, microscopic

*Figure 1. Topology of the Bayesian information retrieval model*



description, and final diagnosis, respectively. These values can be assigned by the physicians with the restriction that the sum $\alpha + \beta + \delta$ must be 1. This means, for example, that we can choose $\alpha = \beta = \delta = 1/3$, so we decide that every subordinate document has the same influence when calculating the probability of relevance for a document in general. Another example is to choose $\alpha = \beta = 1/4$ and $\delta = 1/2$, if we want the final diagnosis to obtain a higher influence by the calculation of the probability of relevance for a document in general.

## Inference and Retrieval

Given a query Q submitted to our system, the retrieval process starts by placing the evidences in the term *subnetwork*—the state of each term $T_{iQ}$ belonging to Q is fixed to $t_{iQ}$ (relevant). The inference process is then run, obtaining, for each document $D_j$, its probability of relevance, given that the terms in the query are also relevant, $p(d_j \mid Q)$. Finally, documents are sorted in decreasing order of probability and returned to the user.

We should mention the fact that the Bayesian network contains thousand of nodes, many of which have a great number of parents. In addition, although the network topology is relatively simple, it contains cycles. Therefore, general-purpose propagation algorithms cannot be applied for reasons of efficiency. We therefore propose the use of a specific inference process (de Campos et al., 2003), which is designed to take advantage of both the topology of the network and the kind of probability function used at document nodes, but ensuring that the results are the same as those obtained using exact propagation in the entire network. The final probability of relevance for a document, therefore, is computed using the following equations:

$$p(d_k \mid Q) = \alpha \cdot p(d_{mk} \mid Q) + \beta \cdot p(d_{\mu k} \mid Q) + \delta \cdot p(d_{fk} \mid Q)$$

where $p(d_{\bullet k} \mid Q)$ can be computed as follows:

$$p(d_{\bullet j} \mid Q) = \sum_{T_i \in D_{\bullet j} \cap Q} w_{ij} + (1/M) \sum_{T_i \in D_{\bullet j} \setminus Q} w_{ij} = (1/\mathrm{M}) \sum_{T_i \in D_{\bullet j}} w_{ij} + (M-1)/M \sum_{T_i \in D_{\bullet j} \cap Q} w_{ij}$$

## FUTURE TRENDS

Because of the excellent features offered by Bayesian networks for representing relationships between variables and their strengths, as well as their efficient inference mechanisms, these probabilistic models will be used in many different areas of IR.

Following the subject of this article (i.e., dealing with structured documents), one interesting line of research would be the introduction of decisions in the inference process. Instead of returning a ranking of documents, it might be very useful to give the user only those parts of the document that might be relevant, instead of the whole document. A first attempt has been made by de Campos, et al. (2004) using influence diagrams. This field is relatively new and is an open and promising research line. On the grounds of the basic methodology proposed in this article, an intuitive step in this line of work would be to open the field of research to the area of recommendation systems, where Bayesian networks also can perform well.

The Web is also a challenging context. As well as the large number of existing Web pages, we must consider the hyperlinks between these. A good treatment of these links by means of a suitable representation through arcs in a Bayesian network and by means of the conditioned probability distributions, which should include the positive or negative influences regarding the relevance of the Web page that is pointed to, should help improve retrieval effectiveness. Finally, another interesting point would be not to consider index terms independent among them, but to take into account relationships, captured by means of data mining techniques with Bayesian networks.

## CONCLUSION

In this article, we have presented a retrieval model to deal with medical records from a pathological anatomy service represented in XML. The retrieval model operates in two stages, given a query: the first one employs an XPath query to retrieve XML documents, and the second, using Bayesian networks, computes a probability of relevance using IR techniques on the free text tags from records obtained in the previous step. This model ensures not only an accurate representation of the structure of the record collection, but also a fast mechanism to retrieve relevant records given a query.

## ACKNOWLEDGMENT

(a) This work has been jointly supported by the Spanish Fondo de Investigación Sanitaria and Consejería de Salud de la Junta de Andalucía, under Projects PI021147 and 177/02, respectively; (b) we would like to thank Armin Stoll for his collaboration with the development of the software implementing the model presented in this article.

## REFERENCES

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval.* Addison-Wesley.

Calado, P., Ribeiro, B., Ziviani, N., Moura, E., & Silva, I. (2001). Local versus global link information in the Web. *ACM Transactions on Information Systems, 21*(1), 42-63.

Chiaramella, Y. (2001). Information retrieval and structured documents. *Lectures Notes in Computer Science, 1980*, 291-314.

Crestani, F., de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003a). A multi-layered Bayesian network model for structured document retrieval. *Lecture Notes in Artificial Intelligence, 2711*, 74-86.

Crestani, F., de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003b). Ranking structured documents using utility theory in the Bayesian network retrieval model. *Lecture Notes in Computer Science, 2857*, 168-182.

Crestani, F., Lalmas, M., van Rijsbergen, C.J., & Campbell, L. (1998). Is this document relevant? Probably: A survey of probabilistic models in information retrieval. *Computing Survey, 30*(4). 528-552.

de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003a). An information retrieval model based on simple Bayesian networks. *International Journal of Intelligent Systems, 18*, 251-265.

de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003b). Implementing relevance feedback in the Bayesian network retrieval model. *Journal of the American Society for Information Science and Technology, 54*(4), 302-313.

de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003c). The BNR model: Foundations and performance of a Bayesian network-based retrieval model. *International Journal of Approximate Reasoning, 34*, 265-285.

de Campos, L.M., Fernández-Luna, J., & Huete, J.F. (2003d). Improving the efficiency of the Bayesian network retrieval model by reducing relationships between terms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 11*, 101-116.

Dumais, S.T., Platt, J., Hecherman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Graves, A., & Lalmas, M. (2002). Video Retrieval using an MPEG-7 based inference network. *Proceedings of the 25th ACM-SIGIR Conference*.

Jensen, F.V. (2001). *Bayesian networks and decision graphs*. Springer Verlag.

Ludovic, D., & Gallinari, P. (2003). A belief network-based generative model for structured documents. An application to the XML categorization. *Lecture Notes in Computer Science, 2734*, 328-342.

Myaeng, S.H., Jang, D.H., Kim, M.S., & Zhoo, Z.C. (1998). A flexible model for retrieval of SGML documents. *Proceedings of the 21th ACM—SIGIR Conference.*

Piwowarski, B., & Gallinari P. (2002). A Bayesian network model for page retrieval in a hierarchically structured collection. *Proceedings of the XML Workshop—25th ACM-SIGIR Conference*.

Reis, I. (2000). *Bayesian networks for information retrieval* [doctoral thesis]. Universidad Federal de Minas Gerais.

Tsikrika, T., & Lalmas, M. (2002). Combining Web document representations in a Bayesian inference network model using link and content-based evidence. *Proceedings of the 24th European Colloquium on Information Retrieval Research*.

Turtle, H.R., & Croft, W.B. (1991). Evaluation of an inference network-based retrieval model. *Information Systems, 9*(3), 187-222.

Turtle, H.R., & Croft, W.B. (1997). Uncertainty in information systems. In *Uncertainty management in information system: From needs to solutions* (pp. 189-224). Kluver.

Wong, S.K.M., & Butz, C.J. (2000). A Bayesian approach to user profiling in information retrieval. *Technology Letters, 4*(1), 50-56.

## KEY TERMS

**Bayesian Network:** A directed acyclic graph where the nodes represent random variables and arcs represent the relationships between them. Their strength is represented by means of conditional probability distributions stored in the nodes.

**Information Retrieval:** A research field that deals with the representation, storage, organization, and accessing of information items.

**Probability Distribution:** A function that assigns a probability to each value that a random variable can take, fulfilling the Kolmogorov's axioms.

**Recommendation System:** Software that, given preferences expresses by a user, select those choices, from a range of them, that better satisfy these user's preferences.

**Structured Document:** A textual representation of any object, whose content could be organized around a well-defined structure.

**XML:** Acronym for Extensible Markup Language. A meta-language directly derived from SGML but designed for Web documents. It allows the structuring of information and transmission between applications and between organizations.

**XPath:** A language designed to access the different elements of an XML document.

## ENDNOTE

[1] The notation $T_i$ ($D_j$, respectively) refers to both the term (document, respectively) and its associated variable and node.

*This work was previously published in Encyclopedia of Data Warehousing and Mining, edited by J. Wang, pp. 960-964, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).*