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 T 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