Information Retrieval Using Bayesian Networks

Lukasz Neuman, Jakub Kozlowski, and Aleksander Zgrzywa

Department of Information Systems, Wroclaw University of Technology, Poland
{neuman,zgrzywa}@pwr.wroc.pl, topic@oporow.net

Abstract. Information retrieval (IR) systems are used for finding those documents, which satisfy user information need. By such a great increase of documents in the Internet, income of information in databases, precise and quick retrieval of relevant documents is of great significance. Artificial intelligence methods can be essential for achieving this goal. The article describes one of such methods – a model of IR based on Bayesian networks. Usage of the network and an experiment aiming in showing that using this method improves information retrieval is presented. An emphasis was made on the benefits of using the Bayesian networks and the way of adapting such a network to information retrieval system is presented.

1 Introduction

In the late 90s, World Wide Web has caused an explosion of the amount of information available for user. The number of Internet web sites has increased and is still rapidly increasing. This fact causes many problems connected with locating relevant information. Among other, implementation of ordering, classification and filtering of accessible information is needed. To do that one uses many methods helping information retrieval process such as: indexing, classification, query formulation, comparison of documents, feedback.

The main aim of web-based adaptive systems is to determine a set of documents which are relevant to given information need. The problems are well-known, but using information retrieval systems may be insufficient that's why using the mechanisms of artificial intelligence in browsers, which support the users with technology of processing the text of natural queries, classify the documents, group them and estimate their relevance, is needed.

Information retrieval gives many Internet users many benefits, so testing on improving mechanism, models and tools are still lasting. An example of such model is the Bayesian network, described in detail in the following part of paper.

2 Bayesian Networks

A Bayesian network [9] is a representation of a joint probability distribution. It consists of two components. The first component is a directed acyclic graph G (DAG), whose vertices correspond to the random variables $X_1,\ldots,X_n$. The second

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3038, pp. 521–528, 2004.

component describes a conditional distribution for each variable, given its parents in *G*. Together, these two components specify a unique distribution on $X_1,\ldots,X_n$.

The graph *G* represents conditional independence assumptions that allow the joint distribution to be decomposed on the number of parameters. The graph *G* encodes Markov assumption: **each variable X_i is independent of its non descendants, given its parents in *G***.

By applying the chain rule of probabilities and properties of conditional independencies, any joint distribution that satisfies Markov's assumption can be decomposed into product form:

$$P(\mathrm{X}_1,\ldots,\mathrm{X}_n)=\prod_{i=1}^{n}P\left(X_i\left|Pa^G(X_i)\right.\right) \tag{1}$$

where $Pa^G(X_i)$ is the set of parents of X_i in *G*. Below (Fig. 1) there is an example of Bayesian network.

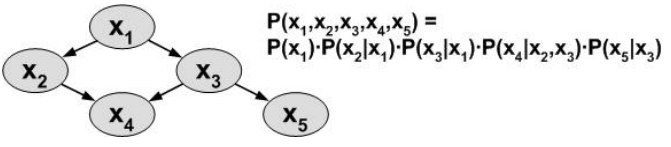

Fig. 1. A simple Bayesian network consisting of 5 nodes

Bayesian structure can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. The Bayesian network has both a causal and probabilistic semantics; it is an ideal representation for combining prior knowledge and data. In conjunction with statistical methods they offer an efficient and principled approach for avoiding the overfitting of data. All these advantages of this structure make the Bayesian network one of the most important research area in Artificial Intelligence especially in information systems [8].

2.1 Application of Bayesian Networks to Information Systems and Related Works

Probabilistic networks [5], [9] have become an established framework for representing and reasoning with uncertain knowledge. They consist of a dependency structure coupled with a corresponding set of probability tables. Distinguish oneself two types of probabilistic networks, namely Bayesian and Markov.

In this chapter we focus on Bayesian networks and their application to web-based systems, which are becoming an increasingly important area for research and application in the entire field of Artificial Intelligence. They model stochastic processes such as medical systems, military scenarios, academic advising, information retrieval [1], [2] system troubleshooting [3], [4], language understanding, business and in many more [8].

Nowadays, many expert systems taking the advantage of Bayesian approach, work very effectively. It is caused, because Bayesian nets readily permit qualitative inferences without the computational inefficiencies of traditional decision making

systems, value of information and sensitivity analysis. Despite offering assistance in the searching process, they support a form of automated learning.

The most important aspect of Bayesian networks is that they are not reasoning processes but they are the direct representations of the world. The arrows in the diagram represent real causal connections and not the flow of information during reasoning [10]. Reasoning processes can operate on Bayesian networks by propagating information in any direction. This kind of graphical representation is easy to construct and interpret. It has formal probabilistic semantics making it suitable for statistical manipulation.

Also the use of Bayesian networks in Internet search browsers [7] is very important because of making connections between documents exceeding out of presence the keywords. The Bayesian networks are used documents classification, thesauri construction [6] and keywords extraction. The vocabulary problem, especially the discrepancies between terms used for describing documents and the terms used by the users to describe their information need, is the one of the key problems of modern Information Retrieval (IR) systems. We can deal with vocabulary problems by using thesaurus, because it shows us the relationships between terms, it is mostly a semantic relationship. We can see three ways of creating thesaurus: first, statistical co-occurrence analyses or the concept space approach, and finally, Bayesian networks.

Example 1. Center for Intelligent Information Retrieval in the University of Massachusetts, has developed INQUERY [1], the most interesting application of Bayesian networks to information systems, which is based on some kind of probabilistic retrieval model, later called document retrieval inference network [14], [15]. It can represent many approaches to information retrieval and combine them into a single framework [6]. It contains two kinds of networks. First net is a set of different representation techniques presenting varying levels of abstraction, and the second represents a need for information (Fig. 2). Nodes in the both nets are either true or false and are representing by values in binary system and interpreted as belief.

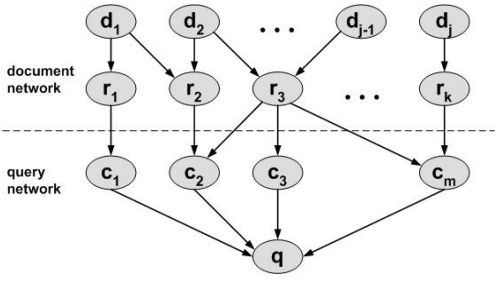

Fig. 2. Simple document retrieval inference network [1]

INQUERY is based on a probabilistic retrieval model and provides support for sophisticated indexing and complex query formulation. It has been used with database containing about 400.000 documents. Recent experiments showed that this system improves information retrieval using Bayesian approach.

3 Experiment

The aim of our experiment was to examine the use of Bayes' methods in the process of information retrieval. What seemed interesting was checking the possibilities of this structure to learn relevance feedback by the user, and on the base of that – increasing the search effectiveness.

This experiment was conducted using a collection of 3204 records from the journal "Communications of the Association for Computing Machinery" (CACM). This collection contains also 64 questions, of which 52 have an additional information about relevant documents. We followed with using the Porter algorithm [11] to remove the commoner morphological and inflexional endings from words. Below we present a table containing 8 most frequently stems in the CACM collection and a diagram fragment of keywords occurring, sorted according to occurring frequency.

Table 1. Ten most frequently stems in the CACM collection with the number of occurrence

Order	Stem	Number of occurrence
1	program	2111
2	algorithm	2015
3	system	1946
4	comput	1945
5	language	1071
6	method	1013
7	data	950
8	time	860

On the basis of common relevant documents for every pair of questions, we have defined a factor of similarity *S* by means of the formula:

$$S(a,b)=\frac{|A\cap B|}{\frac{1}{2}\left(|A|+|B|\right)} \tag{2}$$

where *A* and *B* mean the set of documents relevant adequately to question *a* and *b*. For *S >= 0.25*, we have created a graph of question similarity (Fig. 3); every node of the graph has an annotated average arithmetical value of similarity to the questions represented by nodes joined with it by the edge.

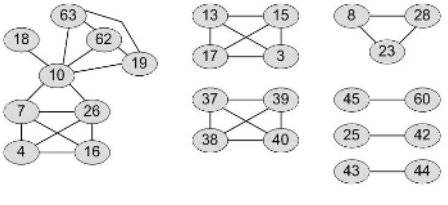

Fig. 3. Graph of question similarity; numbers indicates question number in CACM collection.

Examining the influence of learning feedback on the effectiveness of the search consists in learning the model of answering to the questions with highest similarity factor, and than on checking how this process has reflected on effectiveness of searching questions directly connected with it. Following, the desired behavior of the the intelligent IR system would consist of increasing the effectiveness of answering to particular questions.

3.1 The Implementation of the Bayesian Model

Let I be the information need and D_j denotes the nodes configuration *(d)* representing the documents in Bayes network, so that node j was observed ($d_j=1$), and the others were not, meaning.

$$D_j = \langle d_1 = 0, \ldots, d_{j-1} = 0, d_j = 1, d_{j+1} = 0, \ldots, d_N = 0 \rangle \tag{3}$$

Than the ranking of documents $rank(I,d_j)$ is created counting $P(I=1|D_j)$ for $j=1....N$ as follows:

$$rank(I, d_j) = P(I = 1|D_j) = \frac{P(I = 1, D_j)}{P(I = 1, D_j) + P(I = 0, D_j)} = \frac{P(I = 1, D_j)}{P(I = 1, D_j) + 1 - P(I = 1, D_j)} = P(I = 1, D_j) \tag{4}$$

The information need I is expressed by the sequence of different keywords, which probability of occurring in document j is q_{ij}, $i=1...n$. Above expression can be presented in a simplified form:

$$P(I = 1, D_j) = \frac{1}{n} \sum_{i=1}^{n} q_{ij} \tag{5}$$

Initially q_{ij} was initialized in accordance with TF-IDF (term-frequency, inverse document frequency) method [12]:

$$q_{ij} = d_{ij} = (1 + \log df_{ij}) \log \frac{idf_i}{N} \tag{6}$$

This method was presented as one of the possibilities defined by Turtle [13]. In this case, working of the model with Bayesian's network, before the process of learning, is identical as for the vector model.

Learning of the feedback was realized by means of the MAP (maximum a posteriori) algorithm. It was based on changing values q_{ij}, which were changeable for those documents which were considered. The values q_{ij} were not changed for the documents regarded as irrelevant. The modification was given by the formula:

$$q'_{ij} = \frac{\beta q_{ij} + 1}{\beta + 1} \tag{7}$$

where β is an arbitrary chosen factor of belief, $\beta=10$.

From the test questions of the CACM collection, we have chosen three questions of the highest level of probability. They have performed the following condition: there is no connection between those questions. That means, we have chosen questions with highly similar adequates, but not conjoined between each other by the similarity relation. Above conditions were fulfilled by questions no 17, 39 and 63.

Table 2. Question chosen to the experiment

Question number	Similarity factor	The numbers of neighbor's question in the graph of the probability
17	0,50	3,13,15
39	0,53	37,38,40
63	0,62	10,19,62

Recall	Model's precision	
	Perfect	The worst
0,1	1,0000	0,0109
0,2	1,0000	0,0217
0,3	1,0000	0,0322
0,4	1,0000	0,0425
0,5	1,0000	0,0526
0,6	1,0000	0,0625
0,7	1,0000	0,0721
0,8	1,0000	0,0816
0,9	1,0000	0,0909
1,0	1,0000	0,1000

Fig. 4. Hypothetic values of recall and precision of answers of perfect and the worst model

3.2 Evaluation of Results

We have used the basic criteria of valuation in the information retrieval systems, precisely – recall and precision. The properties of the chart are important because the foundation was using of the models of searching not only to separate the documents for relevant and irrelevant sets, but to create an order according to level of relevance list of documents. Because of that for the given question the recall will always be 1, (*r=1*), while the precision will always equal *p=Rel/N*, where *Rel* is the number of relevant documents and *N* – the size of the collection.

A good model places the relevant documents at the beginning of the list. A wrong model will place at the beginning documents which are irrelevant. A question arises – how do the best and the worst diagrams look like *p(r)* (Fig. 4). One can follow it by an example of two system's answers: both give 100 documents from which 10 are relevant. One of the systems places all ten at the very beginning, other – at the end.

The curve of an ideal model is constant, while the curve of the worst is increasing. It results from the dependence:

$$\forall_{\substack{a>b \\ a,b\in N}} \frac{a+1}{b+1} > \frac{a}{b} \tag{8}$$

This example presents that the better the model, the higher the curve *p(r)* lays on the diagram. It gives an idea of using additional criteria of effectiveness, meaning the area under the curve *p(r)*. The higher is the curve, the higher the value of the area, which implies the better effectiveness of the model. In counting for the data achieved during the experiment, the value of integral $\int_0^1 p(r)dr$ was estimated by the sum:

$$\sum_{i=1}^{|R|}(r_i - r_{i-1})p(r_i) \tag{9}$$

where R means the set of the points of recall, extended with an extra point $r_0=0$. The sum takes values of the range $(0,1]$. Because usually the relevant documents number for the question is much smaller than the total collection size, one can say that for the worst model the value of above sum is close to zero. The perfect model gives value 1:

- the worst $\sum_{i=1}^{|R|}(r_i - r_{i-1})p(r_i) \approx 0$
- perfect $\sum_{i=1}^{|R|}(r_i - r_{i-1})p(r_i) = 1$ (10)

The diagrams presented below show the results of learning done the Bayesian network and the results of non–thought network. 5 possibilities of teaching were estimated, they differ by the number of iterations of teaching steps: 1 step, 5, 10, 20, 50 steps. Because the tables of probability of terms in document occurrences were initialized by the same values as weights in the vector model, effectiveness of the non–thought network are exactly the same as the effectiveness of the vector model. Because of that, the curves $p(r)$ of the vector model were not placed in the diagram.

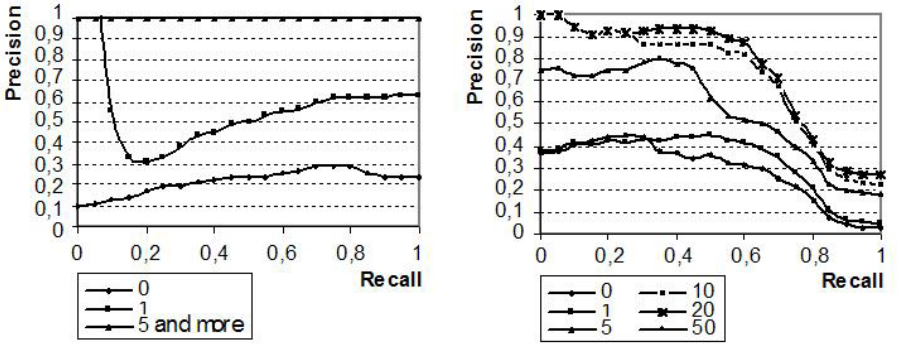

Fig. 5. The effectiveness of Bayesian networks for question 63 of the CACM collection and joined effectiveness of the web for questions 10, 19, 62 for curtain amount of iteration number of learning question 63.

The diagrams at above figure (Fig. 5) present that teaching Bayesian network of answers for specific questions gives very good results. Following iterations significantly improve the results making the net answers for thought questions almost without mistakes. What is important is the fact that the correction does not only concern specific question but also the questions relative to it. The experiment proved that the biggest improvement exists in from 0 to 10 iterations of thought algorithm.

4 Summary

To fulfill the growing needs of the information retrieval system users a lot of methods and artificial intelligence mechanisms are used. Those methods aim is helping the users and fulfilling their information requirements. This paper describes one of the information retrieval methods based on the Bayesian networks. Some of the usage of Bayesian structures applied in information retrieval were presented.

Our experiment presents an example of using the structures of the network in process of information retrieval. Adequately used structures significantly improve the effectiveness and quality. It is very important in the information retrieval.

Summarizing we claim that the Bayesian network is a simple structure, as for initiating apriori values of the distribution of probability tables, allows to be thought and propagate in time *O(n)*. Although it is happening by significant simplification.

References

1. Callan, J.P., Croft, W.B., Harding, S.M.: The INQUERY Retrieval System. Proceedings of the 3rd International Conference on Database and Expert Systems Applications (1992) 78-83
2. Fung, R., Favero, B.D.: Applying Bayesian networks to information retrieval. Communication of the ACM, **38**(3) (1995).
3. Heckerman, D., Breese, J.S., Rommelse, K.: Troubleshooting under uncertainty. Technical Report MSR-TR-94-07, Microsoft Research, Redmond, WA (1994).
4. Horvitz, E.: Lumiere Project: Bayesian Reasoning for Automated Assistance. Decision Theory & Adaptive Systems Group, Microsoft Research, MS Corp. Redmond, WA (1998).
5. Jensen, F.: An Introduction to Bayesian Networks. UCL Press Ltd, London (1996).
6. Jing, Y., Croft, W.B.: An association thesaurus for information retrieval. In RIAO'94 Conference Proceedings, New York (1994) 146-160.
7. Klopotek, M.: Inteligentne wyszukiwarki internetowe. Exit, Warszawa (2001), (in Polish).
8. Neuman, L., Zgrzywa A.: Application the Bayesian Networks to Information Systems. IIAS-Transactions on Systems Research and Cybernetics **II**(1) (2002) 19-23.
9. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers (1988).
10. Pearl, J., Russell, S.: Bayesian Networks. TR R-277, University of California (2000).
11. Porter, M.F.: An algorithm for suffix stripping. Program, **14**(3) (1980) 130-137.
12. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. New York, NY: McGraw-Hill (1983).
13. Turtle, H. R.: Inference networks for document retrieval. Ph.D. dissertation. Computer and Information Science Department, University of Massachusetts. COINS TR **90–92** (1990).
14. Turtle, H., Croft, W.B.: Efficient probabilistic inference for text retrieval. In RIAO'91 Conference Proceedings, Barcelona, Spain (1991) 644-661.
15. Turtle, H., Croft, W.B., Evaluation of an Inference Network-Based Retrieval Model. ACM Transactions on Information Systems **9**(3) (1991) 187-222.