UDK 681.5

M. Voronenko, T. Honcharova-Ilina, J. Krejci

BAYESIAN BELIEF NETWORK OF COMMODITY RELEVANCE ASSESSMENT

Annotation. *In this work we consider the creation of a search engine for relevance assessment in searching commodity orders on the Internet by means of Bayesian methods.*

Keywords: *relevance assessment, Bayesian network, network, document*

Introduction

In this work we examine the relevance assessment in searching orders for goods on the Internet. Home appliances market is dynamic, changing subject area and search for orders in the market is the task of information search, which more and more people tend to perform on the Internet. Creating a search engine requires the solution of relevant assessment task, and in our case, this problem was solved by Bayesian methods of probabilistic reasoning.

The purpose

So let us assume we have a certain set of documents (advertisements on buying goods) obtained from the Internet. Each document is characterized by certain details typical of the advertisement on buying goods, such as type of product, brief description, specifications, customer reviews, price, counterparts and others. Search engine user specified in his request keywords describing the product appropriate for him. The user also has the opportunity to specify a (possibly empty) set of criteria such as the location of the store (storage) and the desired price. We will hereinafter call the set of keywords and criteria user's request.

We will call the degree of compliance of each particular document to user's request a document relevance to request. The task of a search engine is providing the user with the most relevant results, the advertisements that best meet his request. [1]

Thus, it is necessary to build an intelligent system to determine the extent of the relevance of each available document to entered user's request. The number, according to the value of which we can carry out sorting documents by relevance is called the measure of relevance. This number must have the following properties:

1) the measure of the relevance is a nonnegative real number;

2) the higher the measure of relevance, the higher the relevance of the request;

3) the measure of relevance should be limited from the top.

The latter condition is extremely important in terms of user's convenience. Giving the user the ability to analyze the measure of relevance as a number of a certain range, we to some extent give him the opportunity to assess the absolute degree of compliance with his request document.

Let us put as the aim to demonstrate relevance to the user as a real number from 0% to 100%. For this purpose we round off and normalize our measure of relevance, implying thus that 100% relevant document is an advertisement that definitely meets the user's request in terms of our system.

One of the problems for solution of which Bayesian networks have been successfully applied is the task of classification. The so-called naive Bayes classifier, which is a simple Bayesian network is one of the most effective classifiers [2,3]. Our approach provides that the task of relevance assessment can also be seen as a problem of classification. Indeed, let us consider each document (advertisement on buying goods) as the one that belongs to one of two non-overlapping areas: C1 - relevant documents, C2 irrelevant documents.

In this case, the task of assessment of the document relevance to the request is represented as a task of attributing it to one of two classes. In this case, belonging of the document to the first class lets us indicate that this document is relevant to the request.

In our case, we implemented our own approach to determining relevance - intelligent full-text post-processing of found documents by using Bayesian belief network.

The main material

We solve this problem by using a Bayesian network, taking the concept of "document" to network peak. This top can be in two states: C1 - "relevant document" and C2 - "irrelevant document." A priori probabilities of these states are equal to 0.5, which corresponds to the concept of uncertainty in probabilistic analysis. If after performing calculations, we find that the probability of this point in "relevant document" state is equal to 0.9, it would mean that with probability of 0.9, this document belongs to C_1 class.

Further, let us assume that $F = \{F_i\}, i = 1..n$ is a set of factors that affect the relevance of the document. For example, let us consider such factors as the availability of a key word of the request in the document title. Obviously, the presence of the key word in the title increases the relevance of the document. Then we introduce the top F1 to the network, relevant to the event "key word in the title of the document." This top will have two states: f_{11} - «Key word is met in the title of the document" and f_{12} - «Key word is not met in the title of the document." If we know the conditional probabilities $P(f_{1j} \mid c_i),\ i, j = 1..2$, we have a table of conditional probabilities for the top F_1, and we can calculate the probabilities $P(c_i \mid f_{1j}),\ i, j = 1..2$.

For the assignment of the document D to relevant class in case when we know the condition of f_{1j}, an obvious rule is used: if $P(c_1 \mid f_{1j}) > P(c_2 \mid f_{1j})$ then $D \in C_1$.

Thus, to determine the relevance we must identify all the factors that make the F set, and to specify the tables of conditional probabilities for each factor. Each factor is calculated accordingly for each key word of request.

Thus, network tops in our case are the factors that affect the probability of our "main" unit responsible for the relevance of the document as a whole. Bayesian network for our task has the following form (Fig. 1).

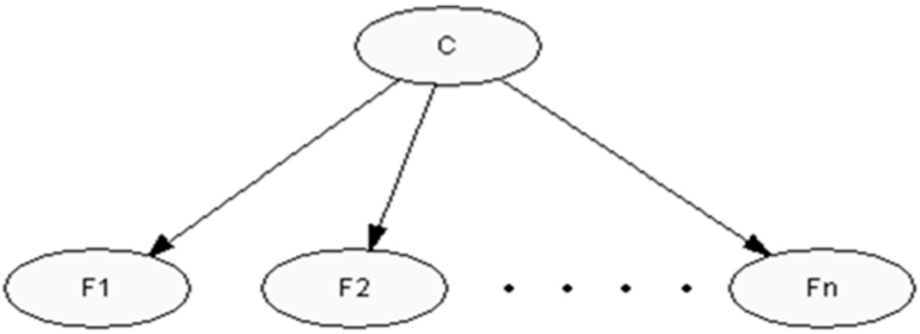

Fig.1 – Bayesian network for assessement of the document relevance to the request

C is the top of the network, which is a possibility of that the document is relevant to request and F_1, F_2 ... F_n are the factors taken into account in calculating this probability. An important point is the direction of causality in the network. Thus, the arrows come from the top C and enter the tops F_i. Here Bayesian network performs inverse logical conclusion – it determines the probability of each state of the top C under certain states F_i.

The case of F_1 factor mentioned above took one of two values. But factors can have different nature - they can take multiple values and might not be discrete at all. In general, we consider a certain range of changes in the values of each factor. Let the factor F_i be set to $x \in [x_{\min}; x_{\max}]$.Then we normalize the value of this factor in the range of [-1; 1] using the formula:

$$\tilde{x} = \frac{x - \frac{x_{\min} + x_{\max}}{2}}{x_{\max} - \frac{x_{\min} + x_{\max}}{2}} \qquad (1)$$

and take the estimated probability to the respective top equal:

$$\begin{aligned} \tilde{P}(f_i \mid c_1) &= \frac{1 - [1 - 2 \cdot P(f_i \mid c_1)] \cdot \tilde{x}}{2}, \ i = 1..n \\ \tilde{P}(f_i \mid c_2) &= \frac{1 - [1 - 2 \cdot P(f_i \mid c_2)] \cdot \tilde{x}}{2}, \ i = 1..n, \end{aligned} , \qquad (2)$$

where $P(f_i \mid c_1)$ is the element of the table of conditional probabilities for the i-th network top that shows how likely a factor F_i in relevant document takes the maximum value $x = x_{\max}$; $P(f_i \mid c_2)$ is the probability with which a factor F_i takes the maximum value $x = x_{\max}$ in irrelevant document.

Obtained estimated probabilities $\tilde{P}(f_i \mid c_1)$ i $\tilde{P}(f_i \mid c_2)$ can now be used in Bayesian formula:

$$P(c_1 \mid f_i) = \frac{P(f_i \mid c_1) \cdot P(c_1)}{P(f_i \mid c_1) \cdot P(c_1) + P(f_i \mid c_2) \cdot P(c_2)}, \ i = 1..n. \qquad (3)$$

It is noted that $\tilde{P}(f_i \mid c_1) = P(f_i \mid c_1)$ where $x = x_{\max}$, and $\tilde{P}(f_i \mid c_1) = 1 - P(f_i \mid c_1)$ as $x = x_{\min}$. For other values $x \in (x_{\min}; x_{\max})$ the estimated probability is $\tilde{P}(f_i \mid c_1) \in (1 - P(f_i \mid c_1); P(f_i \mid c_1))$. This means that the increase in value of x factor leads to a serial (linear) increase in the value of corresponding estimated probability.

Therefore, the scheme described above makes it possible to take into account both discrete and continuous values of the factors that affect the overall relevance of a document. However, if the increase in a factor value corresponds to a decrease of relevance (eg number of days elapsed from the date of publication of an advertisement), it is sufficient to repeat these steps for the case where the elements of the table of conditional probabilities $P(f_i \mid c_1)$ shows how likely the Fi factor takes a minimum value $x = x_{\min}$ in a relevant document.

In this case, the network is fairly trivial to perform estimation of probabilities by means of consistent application of Bayesian theorem. Of course, such estimation is possible only if we make a considerable assumption of conditional independence of network tops. Conditional independence of Bayesian network tops means blocking influence between these tops. Variables (sets of variables) F1 and F2 are independent at a certain state of variable A, if

$$P(F_1 \mid A) = P(F_1 \mid A, F_2)\,. \tag{4}$$

This means that if the state of top A is known, any information about F1 doesn't change the probability of F2. If case of our network it is presented by absence of any causal relationships between all the factors of set F.

In fact, this assumption is, obviously, completely unrealistic (that is why classificators of such structure are called "naive"). At the same time violation of this assumption in a real world shows no significant effect on the final result. It turns out that a consistent approach is an advantage in this case, as it dramatically reduces the computational complexity and therefore the speed of the algorithm.

Considering the question of obtaining numerical values for conditional probabilities tables, it should be noted that, conceptually, to solve this problem there exist two approaches [4,5]:

- Getting information from domain experts;
- Getting information based on data.

Tables of conditional probabilities are often generated based on the data using statistical methods. However, it should be noted that fundamentally subjective Bayesian approach does not require the "objectivity" of probability, and therefore allows the formation of tables of conditional probabilities based on subjective assessment of experts. Conditional probabilities, numerical values that we use for calculation,

 ISSN 1562-9945

are obtained under merging results of statistical studies and expert assessments. We conducted a statistical analysis of a set of relevant and irrelevant documents for various requests by different sources of information and put the values in the table of conditional probabilities network.

The main advantages of using Bayesian networks in selecting relevant product are the ability of combined consideration of qualitative and quantitative indicators, dynamic incoming data processing as well as clear relationship between advertisement semantics and the factors that affect the decision on the application for shipment of goods.

Algorithm of using BN is as follows:

1. Conducting qualitative analysis of documents (advertisements about goods, forums etc) and the degree of their impact on relevance.

2. Determining the influence of factors on each other.

3. Creating rules that describe causal relationships between factors with regard to their particularities.

4. Developing the BN, which meets the requirements of the task.

5. Setting tables of conditional probabilities tables for each of the non-leaf tops of the BN.

6. BN learning, testing BN adequacy.

Further improvement of quality of relevance determination can be achieved by learning BN on available experimental data. Learning is traditionally divided into two components – the choice of an effective network topology, including the possible addition of units that match latent variables and adjusting parameters of conditional distributions for values of variables in the units.

In implementing the system, we have identified the following factors that affect the relevance of job advertisements (Table 1):

We presented the factors in Table. 1 in the form of Bayesian network tops, each of that can take appropriate state and we set a table of conditional probabilities for these tops. When a request arrives the system estimates each factor for each keyword and performs sharing of appropriate estimated probabilities in the network. The result of work is the probability $P(C \mid F_1, F_2 \dots F_n)$ for each available document D, which is the measure of relevance of a request document.

Table 1

Factors introduced to Bayesian network as tops

Factor	States	Explanation
Match of keyword with document title	1) 0 matches 2) 1 and more matches	The presence of a keyword increases relevance of an advertisement; the absence decreases relevance of an advertisement
Matching keyword with first five lines of an advertisement	1) 0 matches 2) 1 match exactly	The presence of a keyword among 25 words of a short description increases relevance of advertisement; the absence does not affect relevance
Repeated matches of keyword with short description of a commodity	1) Less than 2 matches 2) 2 and more matches	Two and more matches increases relevance of advertisement
Number of matches of keyword with advertisement text	1) Less than 2 matches 2) From 2 to 7 matches 3) More than 7 matches	The presence of a keyword in advertisement text 2 and more times increases relevance of advertisement (non-linearly, by discrete values of factor "2», «3», «4», «5», «6», «7 and more»); one match exactly does not affect relevance of advertisement, the absence of matches decreases relevance of advertisement
Matches of bigrams (pairs of words) with advertisement title	1) 0 matches 2) 1 and more matches	The presence of a bigram that coincides with a phrase of two keywords increases relevance of advertisement; the absence of a phrase does not affect relevance of advertisement
Value of factor TF * IDF for keyword	1) 0 2) Value in the range from 0 to 4 2) Value is more than 4	Larger value of factor Більше TF * IDF [7], that takes into account the frequency of matches of a keyword (TF) and the weight of a keyword in a document (IDF), increases relevance of advertisement (non-linearly, by continuous values of the factor in the range from 0 to 4, at the value «4 and more»-maximal); value 0 does not affect relevance of advertisement
Date of publication of advertisement	Value in the range from 0 to 50 (days)	The factor presents the number of days that passed from the moment of advertisement publication and to a current date. Larger value of this factor decreases relevance of advertisement (non-linearly, by discrete values «1», «2», ... «49», «50 and more»); value 0 («today») does not affect relevance of advertisement

 ISSN 1562-9945

If $P(C = c_1 \mid F_1, F_2 \ldots F_n) > P(C = c_2 \mid F_1, F_2 \ldots F_n)$, then the document is relevant to request, i.e.

$$P(C = c_1 \mid F_1, F_2 \ldots F_n) > 0.5),\ mo\ D \in C_1. \tag{5}$$

Documents that suit the decisive rule (3), are displayed for a user with normalized measure of relevance.

$$P = \frac{P(C \mid F_1, F_2 \ldots F_n) - P_{\min}}{P_{\max} - P_{\min}} \cdot 100\% = 2 \cdot [P(C \mid F_1, F_2 \ldots F_n) - 0.5] \cdot 100\%. \tag{6}$$

To build a structure of BN relations the expert knowledge in this field ids used.

For each component the registry of indicators for evaluation was compiled, then the relations of parameters of the components and the parameters of finished products are set. To represent the relationship between variables and brief specification of joint distribution of probabilities we used Bayesian network that represents the general structure of causal processes rather than specific details. Table of conditional probabilities (Table 2) provides a decomposition of the whole into components.

Table 2

Table of probabilities of product relevance based on expert assessment

Relevance indicators	**Probability of buying product that suits us**		**Probability of buying product that does not suit us**	
	100%	***80%***	***60%***	***Less than 60%***
1. Match of a keyword with advertisement title	40%	30%	20%	10%
2. Match of a keyword with first five lines of advertisement about a product	40%	20%	20%	20%
3. Matching of bigrams (word combinations) with advertisement title	35%	30%	25%	10%
4. Value of the factor TF*IDF	25%	25%	35%	15%
5. Date of publication	100%	75%	50%	25%

First we built a graph of mutual influence of factors, then this graph was expanded by presence of visual connections between associated factors (Fig. 2).

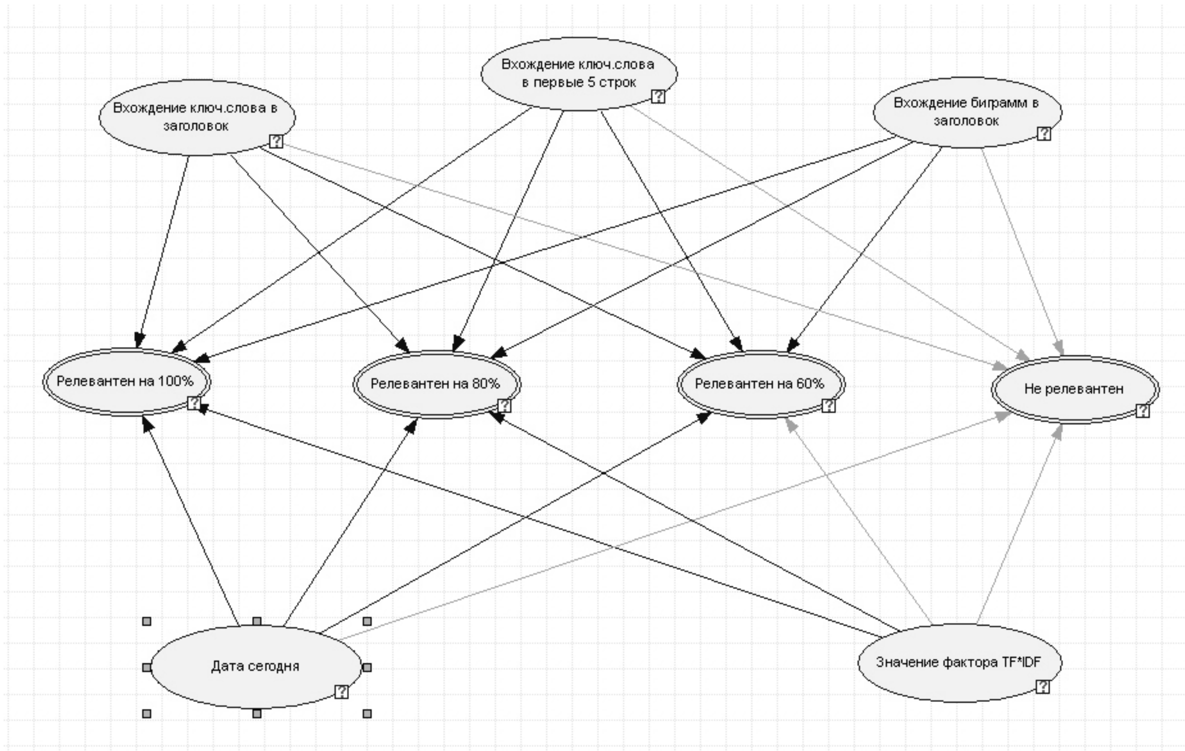

Fig. 2 – Initial state of relevance assessment

To solve this task we use assessment of trends of the most common factors. Some of these indicators are deterministic because they depend on determined variables, but most indicators are probabilistic.

In order to find the most probabilistic combination of states of all tops, you must use the distribution of maximums. After re-estimation a new distribution is obtained in display windows of network and tops. Herewith each state of the tops, having the value of 100% will belong to the most probabilistic combination of states.

The structure of the BN includes qualitative factors, terms of time, indicators of specifications, affecting the decisions of relevance.

Parameters of the Bayesian network have been obtained through learning using the data provided by specialists and experts. The resulting structure of the BN is presented in Fig. 3.

Conclusion

Thus, the essence of our approach to the analysis of relevance of product advertisement is in the use of Bayesian belief network. We adapted the mechanism of probabilistic decision-making for assessing

 ISSN 1562-9945

the relevance, presenting this task as a problem of classification of document - attributing it to the class of relevant or irrelevant. This classification is based on estimating probability of affiliation document to a particular category. The same probability serves as relevance event, allowing us to select advertisements with higher relevance, sort the set of obtained results, provide the user with the ability to select relevance threshold. BN has the possibility to use the probabilities derived empirically or those obtained from the experience of multiple usage of the system, as well as expert assessment, allowing to use in the process of creating orders for goods specialist expert knowledge expressed in the form of assumptions.

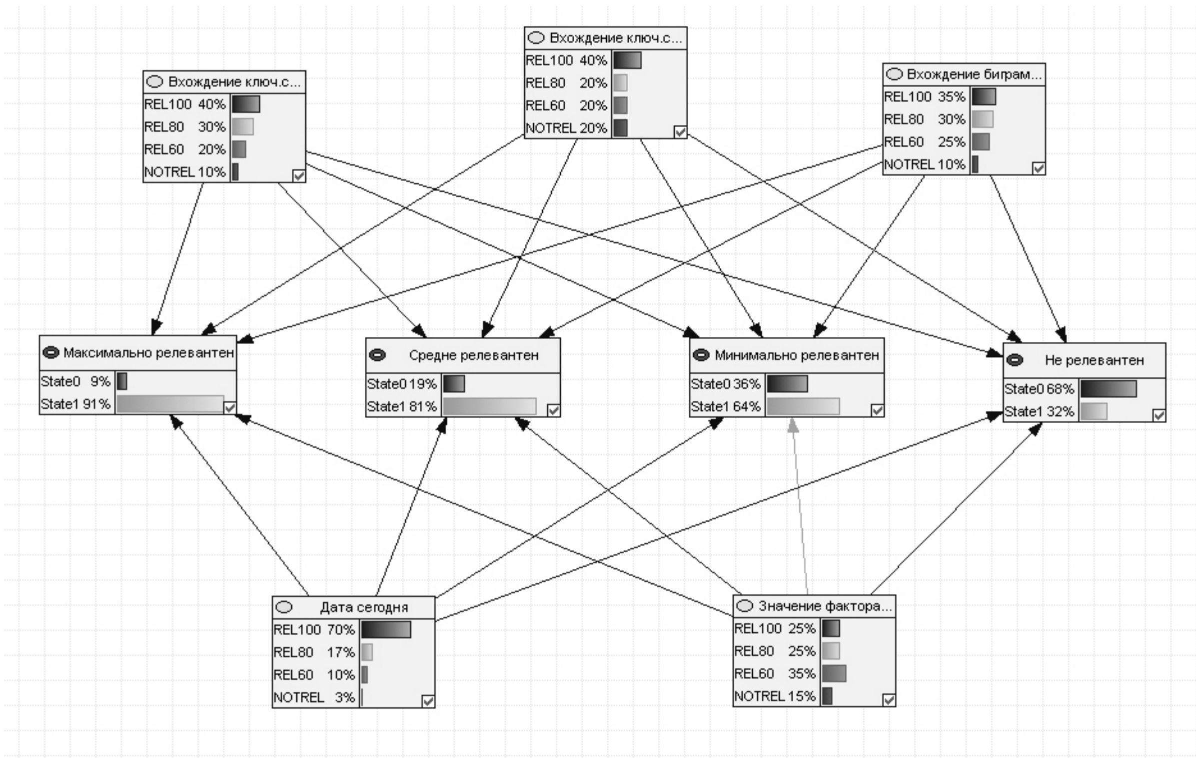

Fig.3 – Bayesian network for assessment of product relevance

The proposed approach has been successfully applied in the practical implementation of a search engine to search advertisements of products online. Further development of methods applied may be in the development of individual search agents that have mechanisms for adaptation of numerical values of tables of conditional probabilities for each particular user.

References:

1. Бідюк П.І. Аналіз ефективності функціонування мережі Байеса / П.І.Бідюк, В.І.Литвиненко,А.В.Кроптя // Автоматика.

Автоматизация. Электротехнические комплексы и системы. – 2007. - №2(20). – С.6-15.
2. Сахаров А.А. Концепции построения и реализации информационных систем, ориентированных на анализ данных// Системы управления базами данных. - 1996. - №4. - С. 5-70.
3. Терехов С.А.. Введение в байесовы сети //Школа-семинар "Совр. пробл. нейроинформатики", 29-31 января 2003. МИФИ, Москва.-V Всеросс. конф. "Нейроинформатика-2003"/Отв.ред. Ю.В. Тюменцев- Часть I: Лекции по нейроинформатике. -М.: МИФИ, 2003.- 188 с. (149-186).
4. Luger G. Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 5th edition /G. Luger. – UK: Pearson Education. – 2006. – 889 p.
5. Peter D. Turney, Michael L. Littman, Jeffrey Bigham, Victor Shnayder: Combining independent modules in lexical multiple-choice problems. RANLP 2003. – p.101-110.