A Document Retrieval System Using the Maximum Entropy Principle and Fuzzy Requests

Kenji Saito, Hiroyuki Shioya and Tsutomu Da-te.

Faculty of Engineering, Hokkaido University,
Kita 13 Nishi 8 Kita-ku Sapporo-shi 0608628 Japan.

kenji@main.eng.hokudai.ac.jp, hiro@main.eng.hokudai.ac.jp,
date@main.eng.hokudai.ac.jp

Abstract

It becomes more important to construct an information retrieval system that can give flexible responses to variable user's requests in accordance with the evolution of the internet. We proposed a document retrieval system based on the maximum entropy principle on a Bayesian network (Saito, et al. 1998).

We can claim two advantages for this document retrieval system. The first is that this system can produce several candidates of keywords which will help a user in retrieving a useful document, even if he can not recall a group of adequate keywords. The second is that this system can be customized corresponding to an individual user by tuning the parameters of a Bayesian network.

But there is a problem for this system that it does not produce a retrieval result for some combinations of keywords. We have solved this problem, in this paper, by introducing fuzzy requests.

Keywords: Document retrieval, The maximum entropy principle, Fuzzy requests, Probabilistic ranking principle, Bayesian network.

1 Introduction

Over the century, the technology of printing and broadcasting has been developed and used for releasing information to many and unspecified persons. In addition, we have come to use the techniques of communications, sound or video recording, duplication of documents or pictures in recent years. These techniques enable us to acquire, store, process and disseminate a large amount of information. Furthermore, the computer techniques have accelerated these information managements for several decades. With these advances, we can say, there has been a flood of information in human life, but the information we really need is only a part of them. Hence, we need the technique of information retrieval to obtain the information we really need.

In the field of document retrieval, computer systems, appeared in 1940s, came to be used for arranging documents in various kinds of document managements in

1950s. Then many researchers began to study document retrieval systems using computer systems.

In early days of document retrieval studies, researchers proposed various methods based on boolean algebra or another theories in logic. On the other hand, the retrieval methods using weighted keywords have been studied to express the retrieval requests more flexibly. Maron and Kuhns proposed a statistical approach by using probability as these weights in 1960. And many theoretical researches have been carried out based on this statistical approach(Fuhr 1995, Crestani and Rijsbergen 1995).

In accordance with a recent explosive increase of information with the advance of the internet system, researchers began to study some functions that make a retrieval system more effective. The functions are seen in the system accepting wide varieties of user's requirements, or in the system being customized for an individual user, or in the system digesting useful information automatically from a large quantity of data when the system collects a document data and prepares for retrieval processes.

Cooper proposed a document retrieval method based on the maximum entropy principle (Cooper 1983). In this method, weighted keywords are used as a user's request that expresses the degrees of probabilistic relations among the keywords and useful documents, and this method is proved to give a retrieval result without bias. A document retrieval system based on the Cooper's method can be customized easily by modifying parameters of system. In many studies of probabilistic document retrieval, researchers adopted this Cooper's method, and achieved good results.

A Bayesian network is a joint probability distribution model for many random variables (Jensen 1995). This model is used in many fields of engineering, because of high flexibility and usability of probabilistic inference method. A Bayesian network can learn a probability distribution from sampling data, and makes a directed graph representing a data structure contained by sampling data.

We proposed a document retrieval system based on the Cooper's method by using a Bayesian network (Saito, et al. 1998,1999). This system can provide a keyword network automatically by learning a document data. And this keyword network is used to indicate candidates of keyword. A user can select these candidates of keywords to retrieve documents in detail. This system can be customized for each user and store additional new document data by modifying a Bayesian network. But, in general, a document retrieval system based on Cooper's method has a drawback that the system can not retrieve useful documents when some combination of keyword is selected as a user's request.

In this paper, we rearrange the Cooper's method in the form of a constrained optimization problem, and we propose a new method with a fuzzy request from the viewpoint of this interpretation. The system applying this method can surely accept a user's request consisting of any combination of keywords, and a user can specify an importance of each keyword by using this fuzzy request in a retrieval process.

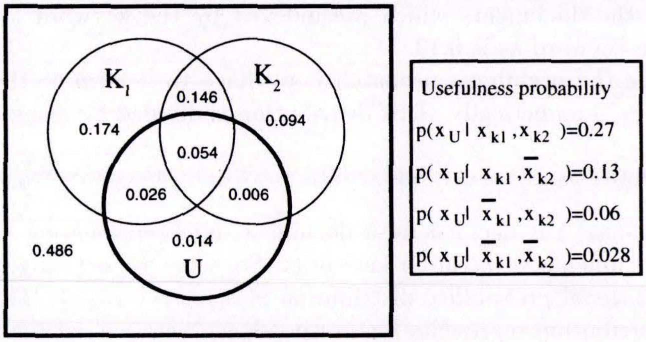

K₁
K₂
U
0.174
0.146
0.094
0.054
0.026
0.006
0.014
0.486
Usefulness probability
$p(x_U | x_{k1}, x_{k2})=0.27$
$p(x_U | x_{k1}, \bar{x}_{k2})=0.13$
$p(x_U | \bar{x}_{k1}, x_{k2})=0.06$
$p(x_U | \bar{x}_{k1}, \bar{x}_{k2})=0.028$

Fig. 1: An example for the usefulness probability.

2 A Document Retrieval System Based on the Maximum Entropy Principle

A probability ranking principle is widely used in many studies of probabilistic retrieval. This principle can be explained by the next sentence. "A most useful document retrieval system calculates usefulness probability of all documents, using all available information, then outputs the documents in order of usefulness probability." In this principle, a calculation method of the usefulness probability is very important, because it relates to a result of retrieval. Cooper proposes a method to calculate the usefulness probability based on the maximum entropy principle. This method is widely used by many researchers because it gives a retrieval result having no bias. We will explain this method below.

Let there be m keywords k_i and m events K_i $(i = 1, \cdots, m)$. The event K_i is assigned to a set of some documents indexed by a keyword k_i. And let U be a set of useful documents for each request. x_{k_i}, x_U are variables defined by

$$
\begin{aligned}
x_{k_i} &= \begin{cases} 0 : \text{A document isn't indexed by the keyword } & k_i, \\ 1 : \text{A document is indexed by the keyword } & k_i, \end{cases} \\
&\quad\; (i = 1, \cdots, m) \\
x_U &= \begin{cases} 0 : \text{A document is not useful,} \\ 1 : \text{A document is useful.} \end{cases}
\end{aligned}
\tag{1}
$$

If a distribution $p(x_U, x_{k_1}, \cdots, x_{k_m})$ is decided, a conditional probability $p(x_U|x_{k_1}, \cdots, x_{k_m})$ can be calculated. This conditional probability is used as the usefulness probability in Cooper's theory. To represent a probabilistic character of x_{k_i}, x_U, we use random variables X_{k_i}, X_U. $p(x_{k_i})$ and $p(x_U)$ are probability distributions of X_{k_i}, X_U which are able to calculate too.

Fig.1 shows a simple example of a set of documents and the usefulness probability. In this figure $p(x_U = 1|x_{k_1} = 1, x_{k_2} = 0) = 0.13$ means that the usefulness probability of the documents which are indexed by the keyword k_1 and are not indexed by the keyword k_2 is 0.13.

To calculate the usefulness probability, we have to determine the distribution $p(x_U, x_{k_1}, \cdots, x_{k_m})$ numerically. This distribution is divided by using chain rule.

$$p(x_U, x_{k_1}, \cdots, x_{k_m}) = p(x_U | x_{k_1}, \cdots, x_{k_m}) \cdot p(x_{k_1}, \cdots, x_{k_m}), \tag{2}$$

where $p(x_{k_1}, \cdots, x_{k_m})$ is decided by a document data representing the correspondence of all documents with the m keywords. So, what we actually have to determine is a conditional probability distribution $p(x_U | x_{k_1}, \cdots, x_{k_m})$. This conditional probability distribution represents a correspondence between each combinations of $\{x_U, x_{k_1}, \cdots, x_{k_m}\}$ and a real number in $[0, 1]$. In short, $p(x_U | x_{k_1}, \cdots, x_{k_m})$ can be represented by 2^{m+1} parameters, and we have to determine 2^{m+1} parameters with consideration for weighted keywords and a distribution $p(x_{k_1}, \cdots, x_{k_m})$. This weighted keyword is given as "(k_1 : 0.3)", and interpreted as $p(x_U = 1 | x_{k_1} = 1) = 0.3$. This equation shows that the usefulness of a set of documents indexed by the keyword k_1 is 0.3. There are m constraint equations corresponding to m weighted keywords.

These m constraint equations restrict the parameters of conditional distribution $p(x_U | x_{k_1}, \cdots, x_{k_m})$, but these are not enough to determine the parameters uniquely. Then we select the parameters which have the highest conditional entropy from all parameters satisfying m constraint equations. This is an application of the maximum entropy principle, which determines the parameters uniquely. The maximum entropy principle can be said that it decides the distribution which contains only given information and has no bias.

3 Bayesian Network

In this paper, we use a Bayesian network (Jensen 1995, Chin and Cooper 1989) to realize a document retrieval system based on the maximum entropy principle. The node of a Bayesian network represents a random variable and it's probability distribution responding to changing of input nodes. This relation between input nodes and a probability distribution is represented by a conditional probability distribution. Fig.2 provides an example. A probability distribution of the random variable X_4 is represented by $p(x_4 | x_2, x_3)$. x_2, x_3, x_4 are the values of random variable X_2, X_3, X_4. Ant we can get a joint probability distribution $p(x_1, x_2, \cdots, x_5)$ by

$$p(x_1, x_2, \cdots, x_5) = p(x_1) \cdot p(x_2 | x_1) \cdot p(x_3 | x_1) \cdot p(x_4 | x_2, x_3) \cdot p(x_5 | x_4). \tag{3}$$

We can take another look at a Bayesian network from a view point of a probability distribution model. In the beginning, we discuss a conditional probability distribution model corresponding to each node of Bayesian network. There are many methods for modeling a conditional probability distribution, but generally the

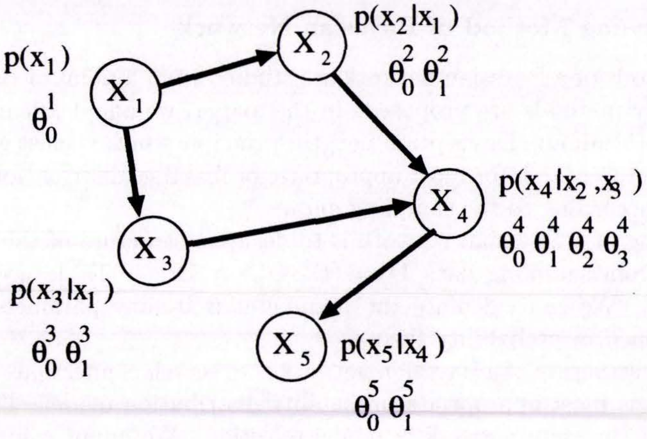


Fig. 2: A Bayesian network.

CPT(Conditional Probability Table) is adopted in many researches of a Bayesian network.

We explain this CPT model by using a concrete example of conditional probability distribution $p(x_i|x_{s^i_1}, x_{s^i_2}, \cdots, x_{s^i_{m_i}})$ which corresponds to ith node of a Bayesian network.($i = 1, \cdots, n$. m_i is a number of parents of ith node. s^i_j is a node number of jth parent of ith node.) We assume that all random variable are binary($\{0, 1\}$). In this situation, the probabilistic conditional probability is a relation which connect a binary number $(x_i, x_{s^i_1}, x_{s^i_2}, \cdots, x_{s^i_{m_i}})$ to a value in $[0, 1]$. Let $\mathcal{I}$ represent this binary number, and $\theta^i_{\mathcal{I}}$ represent a value corresponding $\mathcal{I}$. So $\boldsymbol{\theta}^i = \{\theta^i_0, \theta^i_1, \cdots, \theta^i_{\mathcal{I}}, \cdots, \theta^i_{d_i}\}$, (Where $d_i = 2^{m_i+1} - 1$) is just a conditional probability distribution, and it can be regarded as parameters of a conditional probability distribution too. This concept is adaptable in case of many-value random variable.

In binary case, a size of parameter d_i is 2^{m_i+1}. But conditional probability distribution have a relationship represented by

$$p(x_i = 0|x_{s^i_1}, x_{s^i_2}, \cdots, x_{s^i_{m_i}}) + p(x_i = 1|x_{s^i_1}, x_{s^i_2}, \cdots, x_{s^i_{m_i}}) = 1. \quad (4)$$

So a number of parameters which we need is 2^{m_i}. We use the parameters which correspond to $p(x_i = 1|x_{s^i_1}, x_{s^i_2}, \cdots, x_{s^i_{m_i}})$.

To get a joint probability distribution model corresponding to a Bayesian network, we simply multiply conditional probability distribution models corresponding to each node like eq.3. And parameters of a Bayesian network is defined by gathering parameters of all nodes. That is to say parameters of a Bayesian network is represented by $\boldsymbol{\theta} = \{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \cdots, \boldsymbol{\theta}^i, \cdots, \boldsymbol{\theta}^n\}$. And parameter size of a Bayesian network d is defined by $d = d_1 + d_2 + \cdots + d_n$.

3.1 A Learning Method of Bayesian Network

Learning methods of a Bayesian network are studied from a point of Bayesian inference, and many methods are proposed. In this paper, we adopt a learning method based on MDL(Minimum Description Length) principle which is most general. MDL principle is used to search for most appropriate probability distribution model from many models according to the sampling data.

The learning of a Bayesian network is to decide a structure of the network and parameter $\boldsymbol{\theta}$ from sampling data $\mathrm{D} = \{\mathrm{D}^1, \mathrm{D}^2, \cdots, \mathrm{D}^j, \cdots, \mathrm{D}^N\}$. (Where N is a sampling size.) We can calculate the parameter $\boldsymbol{\theta}$ by any parameter estimation method developed in probability theory.

To decide a structure of a Bayesian network is to search connections of a Bayesian network which is most appropriate probability distribution model. That is to say, a learning of a Bayesian network is model selection. We adopt a model selection based on MDL principle as a learning algorithm of our system. MDL criterion is defined by

$$MDL = -\log p^N_{\hat{\boldsymbol{\theta}}}(\mathbf{D}) + \frac{d \log N}{2}. \tag{5}$$

To adopt MDL principle is to select a model which makes this criterion minimum. The hat symbol of $\hat{\boldsymbol{\theta}}$ represents that this parameter is decided by using a parameter estimation method. We adopt the maximum likelihood method as an inference algorithm.

And this model selection can be regarded as a combination optimization problem by regarding a structure of a Bayesian network as a combination of connections among some node to another node. It is proved that this combination optimization problem is NP-heard. So we have to use a heuristic and approximate algorithm. We adopt the simulated annealing algorithm for our system.

Then we show that the MDL criterion can be divided into parts corresponding to nodes. First, we show more detailed expression of sampling data. As mentioned above, sampling data D is divided into a sample number. So it can be represented by $\mathrm{D} = \{\mathrm{D}^1, \mathrm{D}^2, \cdots, \mathrm{D}^j, \cdots, \mathrm{D}^N\}$. The sampling data numbered j is divided into parts corresponding to nodes. So it can be represented by $\mathrm{D}^j = \{D^j_1, D^j_2, \cdots, D^j_i, \cdots, D^j_n\}$. Then MDL criterion can be represented by

$$\begin{aligned} -\log p^N_{\hat{\boldsymbol{\theta}}}(\mathbf{D}) + \frac{d \log N}{2} &= \\ & \sum_{i=1}^{n} \Big\{ -\sum_{j=1}^{N} \log p_{\hat{\boldsymbol{\theta}}^i}(D^j_i | D^j_{s^i_1}, D^j_{s^i_2}, \cdots, D^j_{s^i_{m_i}}) + \frac{d_i \log N}{2} \Big\}. \end{aligned} \tag{6}$$

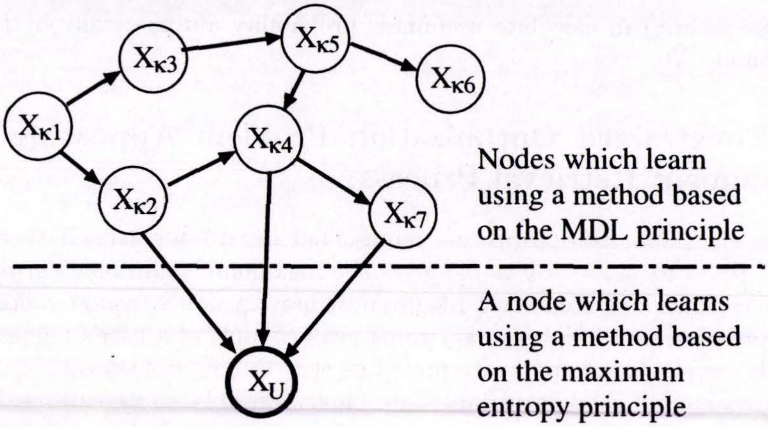

X_{K3}
X_{K5}
X_{K6}
X_{K1}
X_{K4}
X_{K2}
X_{K7}
X_U
Nodes which learn using a method based on the MDL principle
A node which learns using a method based on the maximum entropy principle

Fig. 3: A Bayesian network used for our document retrieval system.

We can calculate MDL criterion more speedy by using this divided version of MDL criterion.

4 A Bayesian Network Used for Our Document Retrieval System

Fig.3 represents a Bayesian network used for our document retrieval system. The network drawn above the broken line represents a probability distribution $p(x_{k_1}, x_{k_2}, \cdots, x_{k_m})$ and corresponds to a database of probabilistic relations among keywords. This part of network learns a structure and parameters by using document data. The document data is a data representing which keywords index a document. This learning process is executed only one time when the document retrieval system start. This learning method makes a network of probabilistic relations among keywords automatically.

The node drawn below the broken line represents a random variable corresponding to an event that a document is useful. A learning method used at this node X_U is based on the document retrieval method proposed by Cooper, and is different from the learning method used at another nodes. This learning process is executed at each user's request. First the node X_U connects to nodes which correspond to each keywords contained in a user's request. Then the node X_U learns the parameter $\boldsymbol{\theta}^U$ given by

$$\arg\max_{\boldsymbol{\theta}^U} H(p_{\boldsymbol{\theta}^U}(x_U|x_{k_1}, x_{k_2}, \cdots, x_{k_m})). \tag{7}$$

Then we have to calculate usefulness probability to retrieve document. But the parameters of conditional probability distribution are the just the usefulness probability. So we can calculate usefulness probability automatically by learning process of node X_U.

5 A Constrained Optimization Problem Appearing in A Document Retrieval Process

We discuss the maximization process represented in eq.7 further. If there is no constraint, $\boldsymbol{\theta}_1^U = \boldsymbol{\theta}_2^U =, \cdots, \boldsymbol{\theta}_2^U = 0.5$ gives the maximum conditional entropy, but we have to consider weighted keywords given by user. A user's request contains not only keywords but also weights of keywords. An example of a user's request input form is "$(k_\iota : w_\iota)$". This form is interpreted as $p(x_U = 1|x_{k_\iota} = 1) = w_\iota$.

In our system, all random variables are binary variable, so we can regard $(x_{k_1}, x_{k_2}, \cdots, x_{k_{(m-1)}}, x_{k_m})$ as a binary number . Let $\mathcal{I}$ represent this binary number. Then to calculate a summation for all combinations of $x_{k_1}, x_{k_2}, \cdots, x_{k_{(m-1)}}, x_{k_m}$ $(\sum_{x_{k_1}=0}^{1} \cdots \sum_{x_{k_m}=0}^{1})$ is equals to calculate a summation for $\mathcal{I} = 0$ to $2^m - 1$ $(\sum_{\mathcal{I}=0}^{2^m-1})$.As mentioned section3, we adopt parameters of a conditional probability distribution $p(x_U = 1|x_{k_1}, x_{k_2}, \cdots, x_{k_m})$ as set of value $\{\theta_0^U, \cdots, \theta_{\mathcal{I}}^U, \cdots, \theta_{2^m-1}^U\}$. In the same way, $\{\pi_0, \cdots, \pi_{\mathcal{I}}, \cdots, \pi_{2^m-1}\}$ represents parameters of $p(x_{k_1}, x_{k_2}, \cdots, x_{k_m})$, and $\{\rho_0,, \cdots, \rho_{\mathcal{I}}, \cdots, \rho_{2^m-1}\}$ represents parameters of $p(x_{k_1}, \cdots, x_{k_{(\iota-1)}}, x_{k_{(\iota+1)}}, \cdots, x_{k_m}|x_{k_\iota})$. $\pi_{\mathcal{I}}$ and $\rho_{\mathcal{I}}$ is a constant value in retrieval process.

Using above symbols, conditional entropy $H(X_U|X_{k_1}, X_{k_2}, \cdots, X_{k_m})$ can be expanded.

$$\begin{aligned} & H(X_U|X_{k_1}, \cdots, X_{k_m}) \\ = & -\sum_{\mathcal{I}=0}^{2^m-1} \pi_{\mathcal{I}} \{\theta_{\mathcal{I}}^U \log \theta_{\mathcal{I}}^U + (1 - \theta_{\mathcal{I}}^U) \log(1 - \theta_{\mathcal{I}}^U)\} \end{aligned} \tag{8}$$

In the same way, a conditional equation $p(x_U = 1|x_{k_\iota} = 1) = w_\iota$ obtained from a user's request can be expanded.

$$\begin{gathered} p(x_U = 1|x_{k_\iota} = 1) = w_\iota \\ \sum_{\mathcal{I}=0}^{2^m-1} \mu_\iota(\mathcal{I})\, \theta_{\mathcal{I}}^U \; \rho_{\mathcal{I}} = w_\iota \end{gathered} \tag{9}$$

$\mu_\iota(\mathcal{I})$ is defined by

$$\mu_\iota(\mathcal{I}) = \begin{cases} 0 : \text{if } \iota\text{th bit of } \mathcal{I} \text{ is } 0 \\ 1 : \text{if } \iota\text{th bit of } \mathcal{I} \text{ is } 1. \end{cases} \tag{10}$$

$\theta_{\mathcal{I}}^U$ must satisfy the following equation, because $\theta_{\mathcal{I}}^U$ represents conditional probability.

$$0 \leq \theta_{\mathcal{I}}^U \leq 1 \ (\text{for } \mathcal{I} = 0, \cdots, 2^m - 1) \tag{11}$$

The above description can be summarized in the following nonlinear programming problem,

$$\begin{array}{ll}
\text{objective function:} & \\
\quad -\sum_{\mathcal{I}=0}^{2^m-1}\pi_{\mathcal{I}}\{\theta^U_{\mathcal{I}}\log\theta^U_{\mathcal{I}}+(1-\theta^U_{\mathcal{I}})\log(1-\theta^U_{\mathcal{I}})\} & \\
\text{constraint:} & \\
\quad \sum_{\mathcal{I}=0}^{2^m-1}\mu_{\iota}(\mathcal{I})\theta^U_{\mathcal{I}}\ \rho_{\mathcal{I}}=w_{\iota} & (\text{for } \iota=1,\cdots,m) \\
\quad 0\leq\theta^U_{\mathcal{I}}\leq 1 & (\text{for } \mathcal{I}=0,\cdots,2^m-1)
\end{array} \qquad (12)$$

That is to say, learning process of node X_U is convex optimization problem consist of convex objective function and convex feasible region. So we can obtain a optimal solution by using hill-climbing method.

6 Features and Drawbacks of the Former Document Retrieval System

The most important features of a document retrieval method based on the maximum entropy principle is in a strict definition of the usefulness probability and existence of a mathematical proof that a result of retrieval is valid. Cooper asserts the other features like an easiness of input form, a delicate controllability of retrieval constraints and an advantage of using user information (Cooper, 1983).

And our document retrieval system on a Bayesian network can make a database of probabilistic relations among keywords automatically, and can represent these relations by a directed graph. These abilities bring an objectivity of a system database and an useful function which selects the candidates of keywords for next detailed retrieval (Saito, et al. 1998).

But the former document retrieval system can not work in a case that there is no relevant answer satisfying all retrieval constraints. Sometime, all answers which satisfy m constraint equations are conditional probability distributions having a minus probability or a probability which is larger than 1.0.

7 Fuzzification of Document Retrieval System Based on the Maximum Entropy Principle

We use Fuzzy theory to make the document retrieval system stable against the case that there is no relevant conditional probability distribution which satisfies all retrieval constraints. Our system searches a distribution corresponding to a retrieval result not only in the family of distributions satisfying the retrieval constraints but also in the neighborhood of these distributions.

A document retrieval system selects one conditional probability distribution as the usefulness probability from all possible conditional probability distributions. These all possible conditional probability distributions make a hypercube $[0,1]^{2^{m+1}}$.

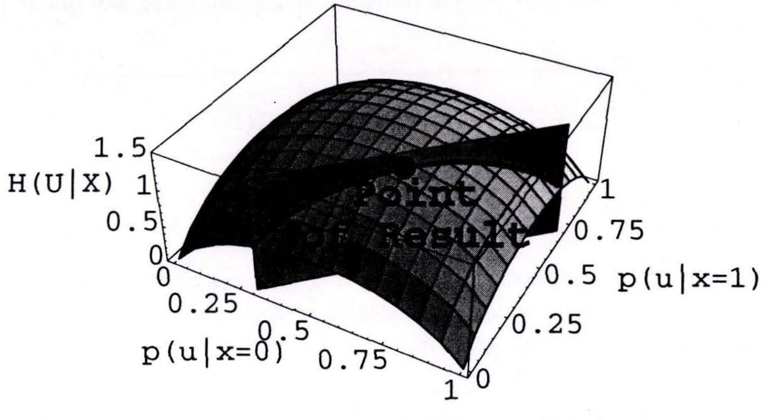


Fig. 4: A probabilistic inference of a traditional document retrieval system.

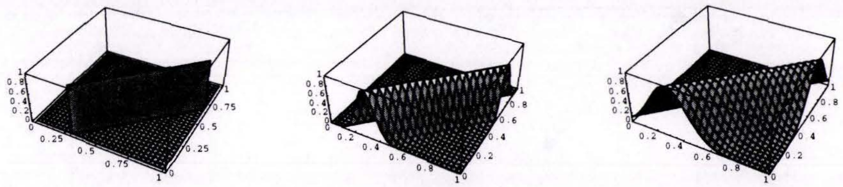
Because a conditional probability distribution $p(x_U|x_{k_1},\cdots,x_{k_m})$ is represented by 2^{m+1} variables in $[0,1]$. So we can consider that the conditional probability distribution is a vector or a point in $[0,1]^{2^{m+1}}$. And the retrieval constraints obtained from user's requests represent a hyperplane. Because the retrieval constraints consist of linear equations.

Fig.4 represents a probabilistic inference of a document retrieval system based on the maximum entropy principle. In this figure, the square made by $p(u|x=0)$ axis and $p(u|x=1)$ axis corresponds to a hypercube made by all possible conditional probability distributions. The meshed convex function represents conditional entropy of conditional probability distributions. A point which assigns high conditional entropy corresponds to a conditional probability distribution which has low bias. And the plane represents retrieval constraints. Only points on this plane satisfy retrieval constraints. The point shown in this figure has the highest conditional entropy in points which satisfy the retrieval constraints. This point corresponds to a result of document retrieval system based on the maximum entropy principle.

In this paper, we expand the retrieval constraints represented by a hyperplane on $[0,1]^{2^{m+1}}$ into a fuzzy set C on $[0,1]^{2^{m+1}}$. A membership function of this fuzzy set is defined by,

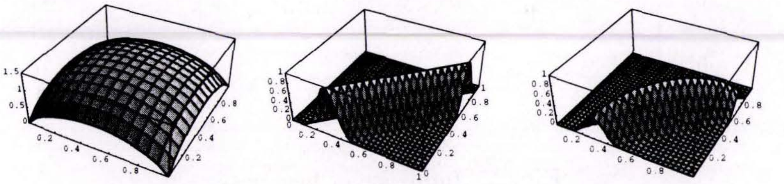
$$\mu_C(\mathrm{p}) = \prod_l \exp\left(-d_l^2(\mathrm{p})\cdot v_l^2\right). \tag{13}$$

In this equation, p is a point in hypercube $[0,1]^{2^{m+1}}$. l is number of a linear equation representing the retrieval constraint. $d_l(\mathrm{p})$ is Euclidean distance between p and the retrieval constraint numbered l. Because the constraint originally corresponds to a hyperplane, Euclidean distance between p and the retrieval constraint can be calculated simply. v_l is fuzziness of the retrieval constraint numbered l. If v_l is



(a)A crisp constraint (b)A fuzzy constraint ($v_l = 10$) (c)A fuzzy constraint ($v_l = 5$)

Fig. 5: Illustrations of the retrieval constraints.



(a)A fuzzy set H (b)A fuzzy set C (c)A fuzzy set B

Fig. 6: Illustrations of an algebraic product.

large, fuzziness is small. If v_l is small, fuzziness is large. Fig.5(a) represents a crisp retrieval constraint. Fig.5(b) and (c) represent fuzzy retrieval constraints. Fig.5(c) is vaguer than Fig.5(b).

And we assume a fuzzy set H. A membership function of this fuzzy set H is defined by,

$$\mu_H(\mathrm{p}) = \frac{H\mathrm{p}(U|X_{k_1}, \cdots, X_{k_m})}{\max H\mathrm{p}(U|X_{k_1}, \cdots, X_{k_m})}. \tag{14}$$

This function is normalized conditional entropy. A fuzzy set H corresponds to a set of conditional probability distributions having a low bias.

Then we assume a fuzzy set B. B is an algebraic product of C and H in fuzzy theory(i.e. $B = CH$). This equation means that a fuzzy set B corresponds to a set of conditional probability distributions which satisfy the retrieval constraints to some degree, and have a small bias. Fig.6 represents this algebraic product of fuzzy sets. Fig.6(a),(b),(c) correspond to H, C, B, respectively.

Next, we have to choose a point from a fuzzy set B to obtain a result of a document retrieval request. Generally a center of gravity method is used in many applications of fuzzy theory. But we choose a point which has the maximum fuzzy grade. Because this method is a natural expansion of a document retrieval method based on the maximum entropy principle. This method converges at a document retrieval method based on the maximum entropy principle when $v_l \to \infty$.

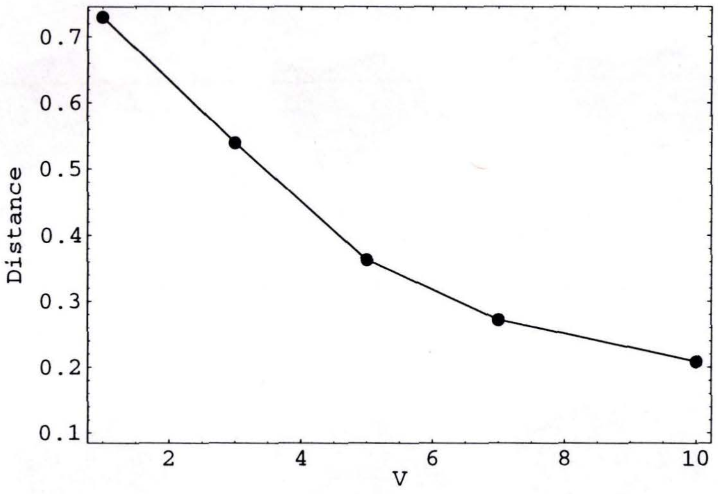


Fig. 7: Influence of fuzziness.

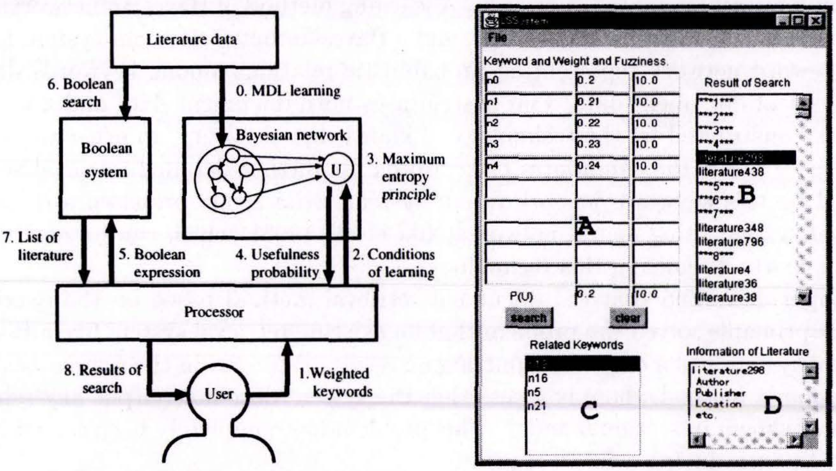
We made an experiment to observe an influence of fuzziness of user's requests. We use artificial document data made from a Bayesian network which has random connections and random values of parameter to estimate experiments.

Fig.7 represents an influence of fuzziness of user's request. The V-axis of this graph corresponds to v_l. The distance-axis represents a distance between a result of a traditional system and a result of our system. This result of document retrieval system corresponds to a conditional probability distribution, and a conditional probability distribution corresponds to a vector on $[0,1]^{m+1}$. So we can calculate Euclidean distance between two results. If v_l is larger than 10, there is almost no difference between a traditional system and our system in actual ranking of document as a retrieval result.

8 An Overview of Our System

We will show an overview of our document retrieval system based on above theory. Fig.8(a) represents components of our system and a flow of retrieval process. Numbers in Fig.8(a) represent order of retrieval process. The process numbered 0 is preparation of database. And processes numbered 1 to 8 are actual retrieval process.

The system executes a learning process of a Bayesian network representing relations among keywords by using a document data(0). This learning process is executed only one time as a preparation of database. A user inputs weighted keywords as a retrieval request(1). The processor translates weighted keywords into constraints of an optimization problem. And it sends constraints to a Bayesian



n0	0.2	10.0
n1	0.21	10.0
n2	0.22	10.0
n3	0.23	10.0
n4	0.24	10.0
P(U)	0.2	10.0

1
2
3
4
literature298
literature438
5
6
7
literature348
literature796
8
literature4
literature36
literature38

n16
n5
n21

literature298
Author
Publisher
Location
etc.

(a) A flow of a retrieval process (b) Interface of our system

Fig. 8: A practical document retrieval system.

network(2). The Bayesian network learns parameters based on maximum entropy principle in consideration for constraints(3). The processor receives usefulness probability obtained by above learning(4). The processor sorts usefulness probability and selecting high useful probability documents, sends Boolean expressions of documents(5). The Boolean system gets practical documents from Boolean expressions(6). The processor receives a result obtained in above(7). The processor displays the retrieval result to a user(8).

A user interface of our system is shown in Fig.8(b), where documents and keywords are represented by integers. In the component A, weighted keywords and fuzziness are displayed. In the component B, useful documents are displayed in the order of the usefulness. In the component C, the candidates of keywords for next detailed retrieval are displayed. These candidates are selected by referring a Bayesian network representing probabilistic relations among keywords. In the component D, detailed literature information is displayed.

9 A Comparison of our document retrieval system with existing systems

In recent years, many researchers have studied the technology of knowledge discovery in databases. This technology contains the methods for discovering causal relationships among attributes, and for constructing conceptual trees or Bayesian networks. It enables us to derive information that can not be obtained by the existing retrieval methods, such as a Boolean retrieval system using keywords or a retrieval method using numerically weighted keywords. A learning method of Bayesian networks is an important example of this technology, and a Bayesian network in our system is used as a keyword network representing probabilistic relations among keywords derived from a set of document data. Our system uses both document data and a keyword network constructed by the technology of knowledge discovery. In other words, the function to select the candidates of keywords for further detailed retrieval steps is realized by this keyword network in our system. Niki et al. proposed a document retrieval system using neural networks(Niki et al. 1995), which can be regarded as another method realizing this technology.

Cooper described that the document retrieval method based on the maximum entropy principle solved the problem that an existing retrieval system like a Boolean retrieval system has a case of outputting no result. But even in the system based on the Cooper's method, there is a case that the system dose not output any retrieval results as shown in section 6 and 7. This problem has completely been solved in our system by introducing a fuzzy request.

This fuzzy request consists of keywords, weights and fuzziness coefficients. By using the fuzzy request, the retrieval system becomes more flexible than the existing system using a general weighted keyword or Boolean operations, and specifies the outputs more precisely, since the fuzzy requests are processed under the maximum entropy principle.

Many search systems at WWW servers have not yet adopted the method of knowledge discovery, but these systems use a hypertext technique that realizes a display of retrieval results with first lines of documents hyperlinking to the corresponding real documents. Our document retrieval system adopts the method of knowledge discovery, and can easily utilize this hypertext technique as a new function of user interfaces.

10 Conclusion

We developed previously a document retrieval system based on the maximum entropy principle on a Bayesian network. Because our system is based on Cooper's retrieval method, a user can input a request simply and flexibly by using weighted keywords. And user can get a retrieval result containing no bias. Because our system is made on a Bayesian network, a keyword network is made automatically from document data. This keyword network is used to display a set of keyword candidates and the user can select keywords from this set to further detailed retrievals. Our system is customized corresponding to an individual user by modifying a Bayesian network.

The user's request is vague specially when weighted keywords are adopted as a user's request, because the weight of the weighted keyword is an extremely subjective value. Then it is thought that our fuzzification of a document retrieval method based on the maximum entropy principle is relevant.

The original document retrieval system can not output any significant result in some case, because there is no intersection of a hyperplane corresponding to the retrieval constraints and a hypercube corresponding to all possible conditional probability distributions in such case. On the other hand, our document retrieval system can search the neighborhood of a hyperplane corresponding to the retrieval constraints, and output a significant result in any case, because a membership function of B is not equal to $\emptyset$. Then the user surely get none-empty result that somewhat satisfies the retrieval constraints with low bias. In many cases, a document retrieval process is repeated several times, referring the former retrieval results. Then some proper results help the user much, even though there is a small gap between the results and the retrieval constraints.

The time for learning a Bayesian network increases rapidly with the number of keywords. And it is remained to find a faster retrieval algorithm. The research on these problems is under way.

References

William S. Cooper (1983). *Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness.* Journal of the American Society for Information Science, Vol.34 No.1, pp.31-39.

Finn V. Jensen (1995). *An introduction to Bayesian networks.* UCL press.

Homer L. Chin and Gregory F. Cooper (1989). *Bayesian Belief Network Inference Using Simulation.* Uncertainty in Artificial Intelligence 3, Elsevier Science Publishers B.V, North-Holland.

Norbert Fuhr (1995). *Probabilistic Datalog - A Logic For Powerful Retrieval Methods.* SIGIR'95, Washington, USA, pp.282-290, July.

F. Crestani and C.J. Van Rijsbergen (1995). *Probability Kinematics in Information Retrieval.* SIGIR'95, Washington, USA, pp.291-299, July.

Kazuhisa Niki and Katsumi Tanaka (1995). *Information Retrieval Using Neural Networks.* Journal of Japanese Society for Artificial Intelligence. Vol.10 No.1, pp.45-51.

Kenji SAITO, Hiroyuki SHIOYA, and Tsutomu DA-TE (1998). *A Literature Search System Based on the Maximum Entropy Principle on a Bayesian Network.* IEICE, Japan, Vol.J81-D-I No.6, pp.770-778.

Kenji SAITO, Hiroyuki SHIOYA, and Tsutomu DA-TE (1999). *A Treatment of Usefulness of Keywords in Fuzzy Requests for an Information Retrieval System with Bayesian Networks.* submitting to IJUFKS.