Web Intelligence and Agent Systems: An international journal 5 (2007) 127–138
IOS Press

An improved Bayesian network structure learning algorithm and its application in an intelligent B2C portal

Junzhong Ji[a,*], Chunnian Liu[a], Jing Yan[a] and Ning Zhong[b]
[a]*College of Computer Science and Technology, Beijing University of Technology, Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing 100022, China*
[b]*Department of Information Engineering, Maebashi Institute of Technology, 460-1 Kamisadori-Cho, Maebashi-City, 371-0816, Japan*

Abstract. Web Intelligence (WI) is a new and active research field in current AI and IT. Intelligent B2C Portals are an important research topic in WI. In this paper, we first investigate and analyze the architecture of a B2C portal for personalized recommendation from the viewpoint of conceptual levels of WI. Aiming at knowledge-level data mining in a B2C portal, we present a new improved learning algorithm of Bayesian Networks, which consists of two major contributions, namely, reducing Conditional Independence (CI) test costs by few lower order CI tests and accelerating search process by means of sort order for candidate parent nodes. Experimental results on benchmark ALARM data sets show that the improved algorithm has high accuracy, and is more efficient in the time performance than other algorithms. Finally, we apply this algorithm to learning Customer Shopping Model (CSM) in an intelligent recommendation system. By a number of experiments on real world data, we find that the recommendation method based on the learned CSM outperforms some traditional ones in rates of coverage and precision.

Keywords: Personalized recommendation, Intelligent B2C portal, Bayesian networks

1. Introduction

Web Intelligence (WI) is a new and active research field in current AI and IT [21,22]. The WI techniques bring many revolutionary changes for scientific research and development on the Internet [8,11,20]. In particular, there are great potentials for WI to make useful contributions to e-commerce, and e-commerce intelligence becomes a major subfield in WI. For the B2C e-commerce, the shopping activities of private customers are undergoing a significant revolution. It has been a focus on how to find meaningful knowledge about customer shopping behaviors and enabling an intelligent portal.

An intelligent B2C portal is a multi-functional gateway, it provides various information and services at a single website [4]. In an e-business B2C portal, a recommendation system is used to suggest commodities to potential customers and to provide online consumers with some personalized services [16]. In order to achieve different recommendation goals, a recommendation system often adopts two ways: the targeted marketing and the personalized recommendation. The first way can predict the potential buyers when a new commodity is present; and the second way can help a marketer to decide what commodity should be recommended to a special customer. Data mining for targeted marketing has been investigated in our papers [19,24]. This paper focuses on personalized recommendation.

Generally speaking, the basic prerequisite for the success of an e-business B2C portal is to fulfil the per-

*Corresponding author. E-mail: jjz01@bjut.edu.cn.

sonalized commodity recommendation. With the personalized needs of customers rapidly increasing, the personalized recommendation has become an important issue for a customer and a business alike. For a customer, a personalized portal creates a true customized shop, which will save his/her time and make him/her to go to needs immediately. In addition, an intelligent portal can capture potential customers demands for a business, and help a business to adapt its marketing strategies, enhancing the competitiveness of the business site. Thus, many researchers focus on how to find customers' behavior patterns and accomplish the personalized recommendation in an intelligent B2C portal.

There are many recommendation systems studied over the last decade, including several techniques, e.g., the nearest neighbor algorithm, the Bayesian analysis, the clustering technique, and many others. In general, these techniques can be divided into two categories, one is called user-based technique which is based on user relations, such as the nearest neighbor algorithm and the customers cluster technique. The other is called model-based technique which is based on item relations, such as association rules, posteriors probabilities, and so on.

In [7], we proposed and experimentally evaluated a new approach in making commodities recommendation, which is based on an Bayesian Customer Shopping Model (CSM) learned from commerce transaction data. This approach formalizes commodity recommendation as knowledge representation of customer shopping information and the knowledge inference process. In order to enhance performance of the learning algorithm, we present an improved Bayesian Network learning algorithm in [8]. This paper is an extension of [8], giving more detailed algorithm description and further experimental results.

The remaining of this section first discusses about related work, and then gives an architecture of an intelligent B2C portal based on conceptual levels of WI. Section 2 introduces preliminary knowledge on Bayesian Network structure learning. In Section 3, we describe an improved learning algorithm of BNs in detail. Section 4 reports our experimental results. Finally, we conclude the paper in Section 5.

1.1. Related work

A Bayesian Network (BN) is an annotated directed acyclic graph that encodes a joint probability distribution over a finite set of random variables. By means of BNs, people can depict and discover many convincing probability dependencies. Thus, a BN has been a powerful knowledge representation and reasoning tool in the uncertainty knowledge field. Especially, the development of data mining in the last decade has significantly stimulated the interest in learning a BN structure from data.

Generally speaking, there are two basic approaches to build a BN structure. The first one poses a BN's learning as an optimization problem [5,6,15,17]. The second approach poses a BN's learning as a constraint satisfaction problem [1,10]. As each has its own weaknesses, a lot of hybrid algorithms [13,14,18] uniting these two approaches have been developed in recent years. The general idea is quite straightforward. First, conditional independence (CI) tests are performed to get an initial network that people are willing to consider, which reduces the search space. Then, a search algorithm is called to find a good network structure which has the best motivated score. Friedman et al. [3] proposed a fast iterative algorithm, which restricts the parent nodes of each variable to a small candidate subset by means of dependent relationships learned from data, then searches for a network that satisfies these constraints. In [18], the researchers introduced a novel hybrid algorithm, which also combines CI tests with the score-and-search process. Based on a similar idea, Qiang Lei [13] proposed a hybrid algorithm, the I-B&B-MDL algorithm, which restricts the search space using heuristic knowledge and enhances the search efficiency by Branch&Bound technology. Given an order of nodes, the I-B&B-MDL algorithm is an efficient one in comparison with other algorithms. However, there are two drawbacks: expensive costs for performing CI tests and no assured efficiency of pruning branches.

This paper proposes an enhanced I-B&B-MDL algorithm, which reduces constraint computation costs by performing few lower order CI tests (order-0 and partial order-1) and improves the pruning efficiency by means of sort order for candidate parent nodes (according to the mutual information between nodes). We also give an application in an intelligent B2C portal.

1.2. The Architecture of an intelligent B2C portal

An intelligent B2C portal enables a business company to create a virtual personalized marketplace on the web, where abundant commodities information and services are provided for every customer. Although specific features of different portals are different, the common functionalities of these portals for B2C e-commerce are the same, such as usability, customiza-

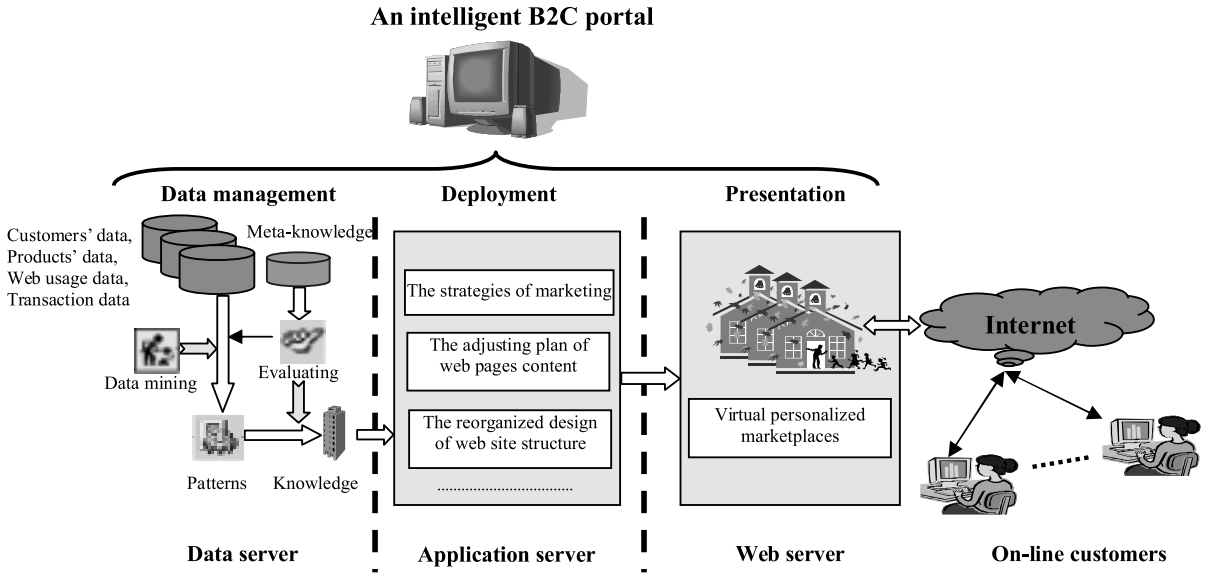

Fig. 1. A typical logic architecture of an intelligent B2C portal.

tion, openness, and transparency [4]. These common features are implemented by using WI techniques and are evolved with the development of WI techniques.

Figure 1 shows a typical logic architecture of an intelligent B2C portal. In this diagram, there are three kinds of servers from the viewpoint of logic, namely data server, application server, and web server. The data server provides the basic support for the application server, and the application server gives the instructions for the web server presentation. Moreover, the web server serves as a contact window for collecting data in the data server. Thus, a cycle of workflows is formed. Extending the four conceptual levels of WI [22,23], the architecture of this B2C portal can be shown in Fig. 2.

(1) **Internet-level communication.** A B2C portal is a computer-network system, which employs some network protocols to communicate with customers. By means of internet media, the B2C portal builds the interconnection of client-server model between the web server and customer's browsers. That is, the internet-level provides general communication infrastructure with protocol softwares.

(2) **Interface-level contact.** A B2C portal first serves as a web server. By way of a web server, a customer's browser connects to a B2C portal and requests a page, then the portal responds and sends back the requested page. That is, a customer's browser forms a connection to a web server, requests a page and receives it. It is obvious that the interface-level provides customers with a single point of contact for online access to commodity information and resource of a B2C portal.

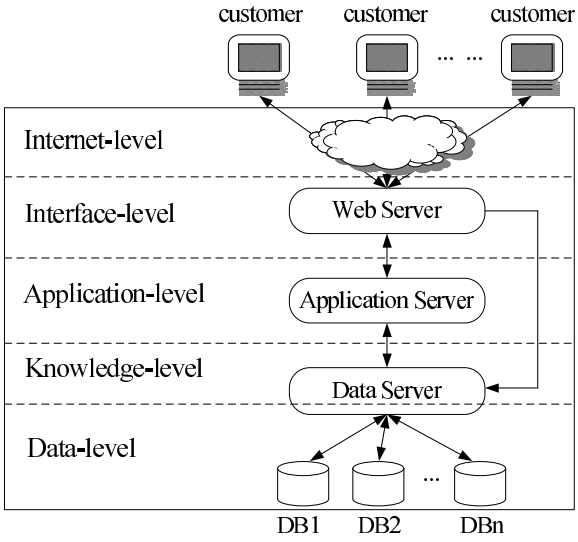

Fig. 2. The architecture of the B2C portal based on conceptual levels of WI.

(3) **Application-level service.** The application server fulfills many specific projects, such as the personalized information search, the proactive recommendation, the mutual customization, the shopping manual, the commodity order, and so on. These functional modules support the B2C portal to achieve several application services, and meet the needs of customers by means of a web server. Though almost all intelligent

functions are carried out in this level, many capabilities in these services mainly depend on knowledge discoveries in a business portal's data. Thus, the application-level acts as a link man between a web server and a data server.

(4) **Knowledge-level discovery.** With the exception of preprocessing data, the knowledge-level mainly performs many tasks of data integrating and mining, knowledge evaluating and reasoning in a B2C portal. These tasks are some prerequisite steps to all of the techniques that provide customers with intelligent services. For a high-performance B2C portal, this level is the most key element of an intelligent portal, and it employs many WI techniques, such as web mining and farming, web information management, web agent, and so on.

(5) **Data-level collection.** As mentioned above, there are various data sources in a B2C portal. Hence, the data-level mainly accomplishes collecting, storing and managing data. In such a distributed web environment, all kinds of database techniques can be employed to design proper database management systems (DBMS), which give each B2C portal the ability to store, retrieve and manage information. Whatever type of information the system wants to track, manage, or analyze, this level can build a database application to suit different needs. Thus, the level provides the most basic and immediate support for the previous levels.

Currently, most of intelligent B2C portals are used to provide automatically online customers with commodities information and services. Especially, the intelligence and personalization of B2C portals have gradually become an acknowledged measure of comparison for advanced or high-performance portals. This trend also inspires the development of WI techniques. In this paper, we stress knowledge discovery in customers' transaction databases, and discuss a new learning algorithm of BNs and its application in an intelligent B2C portal.

2. Preliminary knowledge

2.1. Bayesian networks

A Bayesian Network (BN) is represented by $BN = \langle \mathbf{X}, A, \Theta \rangle$, where $\langle \mathbf{X}, A \rangle$ depicts a directed acyclic graph, each node $X_j \in \mathbf{X} = \{X_1, \cdots, X_n\}$ represents a domain variable. Each arc $a \in A$ represents a probabilistic dependency between the associated nodes. $\Theta = \{\theta_j\}$ is a set of network parameters, where θ_j depicts a conditional probability table associated with each node X_j, which quantifies how much does a node depend on its parents. In graph way, a BN specifies a unique joint probability distribution over $\mathbf{X}$:

$$P(X_1, X_2, \cdots, X_n) = \prod_{j=1}^{n} P(X_j | \Pi(X_j)) \qquad (1)$$

where n is the number of variables in set $\mathbf{X}$, $\Pi(X_j) = \{X_i : i \in \phi(j)\}$ denotes the parent set of X_j in the graph, $\phi(j)$ is a subset of $\{1, 2, \cdots, j-1\}$.

2.2. Conditional independence test

In a BN's learning, a CI test is a typical metric that checks the independence relationship between two variables under conditional set of variables. The basic of CI is the metric of information flow in information theory, thus the mutual information of two variables X_1, X_2 is defined as

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \qquad (2)$$

and the conditional mutual information is defined as

$$I(X_i, X_j | C) = \sum_{x_i, x_j, c} P(x_i, x_j, c) \log \frac{P(x_i, x_j | c)}{P(x_i|c)P(x_j|c)} \qquad (3)$$

where C is a conditional set of nodes, P denotes the instance frequency observed from the sample database. The mutual information can show if the two variables are dependent and if so, how close their relationship is. Hence, when $I(X_i, X_j|C)$ is smaller than a certain threshold value ε, we can say that X_i is independent of X_j given the set C, or else X_i is dependent of X_j. So we can deduce if there is a connection between two variables in light of their mutual information [1]. A χ^2 test is another method to estimate a connection between two variables [10]. Given a degree of confidence σ, we can deduce if there is a connection between two variables using the p-value generated by χ^2 test. In effect, if the p-value is greater than or equal to σ, X_i is independent of X_j, this implies that there is no direct connection between these two nodes. Otherwise, if the p-value is lesser than σ, X_i is dependent of X_j, this implies that a connection between X_i and X_j can exist in the resultant network.

2.3. Previous work on MDL-based learning algorithms

In a feasible solution space, the MDL-based learning algorithm searches for an optimal structure that satisfies the condition of minimum description length (MDL). The main idea of the algorithm can be described as follows.

If the finite set of random variables for a BN is denoted by $\mathbf{X}$, where each variable X_j, $j \in J = (1, 2, \ldots, n)$, may take on values from a finite set $V^j = \{0, 1, \ldots, v^j - 1\}(v^j \geqslant 2$: some integer), the learning of a network structure virtually is the identifying of parent sets $\{\Pi^1, \Pi^2, \ldots, \Pi^n\}$, where Π^j is the set of nodes that a node X^j depends on. Given a sample set $x^{(i)}$=$\{x^1, \ldots, x^n\}$ of $\mathbf{X}$, $i \in$\{1,2,. . .,N\}, where N is the sample size, then $x^{(i)} \in V = \prod_{j \in J} V^j$, $x^N = x^{(1)}x^{(2)} \ldots x^{(N)} \in V^N$. Supposing that G is the set of possible network structures, $g \in G$, then the description length L (g, x^N) of a BN is expressed as [14]:

$$L(g, x^N) = H(g, x^N) + \frac{k(g)}{2} \log N \tag{4}$$

where empirical entropy $H(g, x^N)$ describes the fitness of each possible structure to the observed data, and $H(g, x^N) = \sum_{j \in J} H(j, g, x^N)$ (see below). $k(g)$ is the node description complexity, which stands for the number of independent conditional probabilities embedded in the structure g, and $k(g) = \sum_{j \in J} k(j, g)$ (see below).

$$H(j, g, x^N) = \sum_{s \in S(j,g)} \sum_{q \in V^j} -n[q, s, j, g] \log \frac{n[q, s, j, g]}{n[s, j, g]} \tag{5}$$

$$k(j, g) = (v^j - 1) \prod_{k \in \phi(j)} v^k \tag{6}$$

$$n[s, j, g] = \sum_{i=1}^{N} I(\pi_i^j = s) \tag{7}$$

$$n[q, s, j, g] = \sum_{i=1}^{N} I(x_i^j = q, \pi_i^j = s) \tag{8}$$

where $S(j, g)$ is an instance set of corresponding parent nodes $\Pi^j = \pi^j$ when a model is g, π_i^j is a tuple of realized values in the i-th example, $n(s, j, g)$ and $n[q, s, j, g]$ are two instance numbers, each of which denotes a case frequency occurring in the date sets, and $I(E) = 1$ when a predicate E is true and $I(E) = 0$ otherwise.

Based on the above preparation, the problem of learning a BN becomes a search problem for a structure with MDL metric. In general, an exhaustive search is recursively applied to the MDL-based search procedure. This search examines all possible local changes in each set of parent nodes, so the evaluation cost is acute for massive data sets.

In order to reduce the computational complexity for empirical entropy, Suzuki further proposed a Branch&Bound-MDL-based learning algorithm (B&B-MDL) [14], which can avoid worthless recursive calls for some search branches by estimating a MDL score with a lower cost. In other words, if the value of MDL_1 in the last step is smaller than the lower bound of MDL_2 in current step, and if the lower bound can be computed with small computation, then the further recursive calls in current step can be avoided. That is, the branch including remanent search nodes can be pruned.

The mechanism of pruning a branch is as follows: although the structure complexity of a node increases along with the number of its parent nodes increasing, the value of empirical entropy is nonnegative and descending monotonously, and the decrement of empirical entropy is at most the current empirical entropy $H(j, g, x^N)$. Thus, for a new increasing parent node q, if

$$H(j, g, x^N) \leqslant \frac{k(j, g)(v^q - 1)}{2} \log N \tag{9}$$

then $MDL_2 \geqslant MDL_1$ always holds in this step, namely, because the value of $k(j, g)$ is more increased, any recursive calls is meaningless.

The B&B-MDL-based learning algorithm improved the MDL-based learning algorithm only from the viewpoint of search. However, most of the candidates can be considered to be eliminated in advance based on statistical understanding of the domain. Aiming at this problem, Qiang presented an improved algorithm called as I-B&B-MDL [13]. The general idea is quite straightforward. By using a set of lower order independence tests (χ^2 test), the algorithm restricts the search space and enhances the search efficiency. More precisely, the algorithm uses much mutual information to construct initial network, which restricts the possible parents of each node. Thus, instead of having j-1 potential parents for a node, the algorithm only considers k ($k \ll j$-1) possible parents in each search. Since the search space is significantly restricted, the search performs faster than that of B&B-MDL.

3. The EI-B&B-MDL-based learning algorithm

For the I-B&B-MDL, there are two major problems when dealing with large numbers of nodes. First, the tuple number of each conditional set is so large that it is expensive that the cost of collecting various statistics about data and computing the mutual information, even if only performing lower order independence tests, e.g., the set number of performing order-0 CI tests is C_n^2, thus this phase is of the complexity $O(n^2)$; the set number of performing order-1 is $C_n^2 * C_{n-2}^1$ and requires $O(n^3)$ CI tests, and the set number of performing order-2 is $C_n^2 * C_{n-2}^2$ and requires $O(n^4)$ CI tests, and so on. Secondly without optimizing for search process by heuristic knowledge, there are still many worthless recursive calls. Moreover, because there is extra costs of CI tests, the algorithm cannot ensure that there are enough pruned branches to make I-B&B-MDL be more efficient than B&B-MDL.

3.1. Improving strategies in EI-B&B-MDL algorithm

In order to overcome above drawbacks, we propose an Enhanced I-B&B-MDL algorithm (EI-B&B-MDL) that includes two major improving strategies. Firstly, instead of initial CI test method, order-0 and partial order-1 independence tests are used to obtain an original network, which reduces the number of CI tests and database passes while effectively limiting the search space. In order to account distinctly for our algorithm, we first give the definition of order-1 unilateral double-connection.

Definition 1. (order-1 unilateral double-connection) Given an arc between any two nodes in a BN, if there is another directed path which is the same direction as the arc, and the path only include an extra node , we call this acyclic subgraph as *order-1 unilateral double-connection*.

In Fig. 3, $X_i \rightarrow X_j$ and $X_i \rightarrow X_k \rightarrow X_j$ are two paths connecting X_i and X_j; they are not only in the same direction, but also only include an extra node X_k, so we say that the subgraph is an order-1 unilateral double-connection.

If we only perform order-1 independence tests for the cases of order-1 unilateral double-connection, we get a set of condition tests $Z'_{ij} = \{X_k | X_i \text{ and } X_j \text{ satisfy with order-1 unilateral double-connection}, X_i \prec X_k \prec X_j, k \in J, \text{ and } k \neq i, j\}$. Apparently, $Z'_{ij} \subseteq Z_{ij} = \{X_p | p \in J, \text{ and } p \neq i, j\}$, that is, $|Z'_{ij}| \ll |Z_{ij}|$ (Z_{ij} is the condition set of order-1 according to [13]),

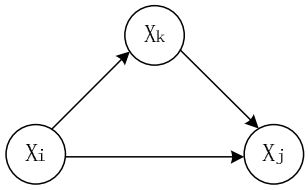

Fig. 3. The sketch map of order-1 unilateral double-connection.

i.e. the phase of order-1 CI tests for the set Z'_{ij} is of the complexity $O(C_n^2 * C_k^1) = O(kn^2)$ ($k \ll n - 2$), thus the EI-B&B-MDL algorithm complexity for lower order CI tests is $O(n^2)$, which is observably smaller than that of I-B&B-MDL (larger than $O(n^3)$). Obviously, this strategy for CI tests evidently improve the testing time of initial algorithm.

Moreover, as mentioned in the Section 2.3, the basic idea of MDL-based algorithms is to decide the parent set $\Pi(X_j)$ for each node X_j ($j \in J$) by a search process. Because the number of possible parent sets is 2^{j-1} for a node X_j, the number of the possible network structures and the number of comparisons for an exhaustive search with the MDL are respectively $2^{n(n-1)/2}$ and $2^n - n - 1$ [15]. To avoid an exhaustive search, the B&B-MDL-based algorithm reduces the number of computations and comparisons by using branch and bound technique. In order to further improve the $B\&B$ search and save computations, we employ heuristic knowledge acquired by computing the mutual information to sort ordering for candidate parent nodes. Our research discovers that different searching places of parent sets in a search tree result in the different amounts of pruning branches, so we try out several orderings of candidate parent nodes and find that the ascending order is a good ordering according to the mutual information between two nodes. This ordering can increase the cut-offs of B&B search tree, and then reduce the computational complexity of the search phase. The effect of the improving strategy is illustrated in Section 3.3.

3.2. Algorithm description

Based on above strategies, this section gives a detailed algorithm description of EI-B&B-MDL. The major steps of the learning algorithm are as follows.

Step 1. Given a node ordering, conduct order-0 CI tests for each pair variables in light of Eq. (2) and corresponding χ^2 test, then build the initial graph G_0, in which each arc passed the χ^2 test, and record the mutual information of each arc in G_0.

Step 2. For every cases of order-1 unilateral double-connection in G_0, conduct order-1 CI tests in light of Eq. (3) and corresponding χ^2 test, and remove the invalid arc that does not pass a CI test. As a result, simplify G_0 to G_1.

Step 3. For each node X_j, ascertain its candidate parents $\Pi(X_j)$ according to the structure of G_1, and produce an ordering of parent nodes by sorting each arc's mutual information in ascending order. Then adopt the enhanced B&B-MDL technique to find a $\Pi'(X_j)$ with the minimum MDL score by top-down search and confirm the local optimized structure of X_j. Let $\pi_1 = \phi$, $p_1 = \frac{v^j - 1}{2} \log N$, the main procedure of the EB&B-MDL algorithm is shown as Algorithm 1.

Algorithm 1: EB&B-MDL(π_1, p_1, MDL_1, Π_1)
/* π_1 : the initial parent set for the current node
p_1 : the initial complexity description
MDL_1 : the optimization score
Π_1 : the parent nodes set after this search */
Begin:
1. Compute the empirical entropy H_1 and
$MDL_1 \leftarrow H_1 + p_1$; $\Pi_1 \leftarrow \pi_1$;
2. if $\pi_1 = \Phi$ then $j \leftarrow 0$ else $j \leftarrow$ the last element in π_1;
3. For $j + 1 \leqslant q \leqslant k$
/* k: the cardinality of candidate parents' set */
{
$\pi_2 \leftarrow \pi_1 \cup Node(q)$;
/* attach a new node q at the end according to the ascending order of candidate parents */
$p_2 \leftarrow p_1 \times v^q$;
/* update complexity description of node */
if $H_1 > p_1 \times (v^q - 1)$ then
EB&B-MDL(π_2, p_2, MDL_2, Π_2);
/* predict the MDL of the node, if it diminishes, then call recursive search*/
if $MDL_1 > MDL_2$ then
$MDL_1 \leftarrow MDL_2$; $\Pi_1 \leftarrow \Pi_2$;
}
End.

3.3. An Example of local structure learning process

In this section, we give an example of learning local structure for a node x_k to illustrate the search effect of the Algorithm 1.

Suppose that when learning local structure of x_k, there are four candidate parent nodes to be searched after passing CI tests , namely, $\Pi(x_k) = \{x_{k1}, x_{k2}, x_{k3}, x_{k4}\}$, and the initial node order is $x_{k1} \prec x_{k2} \prec x_{k3} \prec x_{k4}$. Then, there are sixteen possible network structures in Fig. 4. The search goal is to find an optimization structure with MDL from these structures. Table 1 on the next page demonstrates search processes of different search techniques.

If we perform the MDL-based exhaustive search, the list π_1 traces as the third column in Table 1, all possible network structures are computed and searched. By comparing MDL values, we can get the search result that the parent node set is $\{x_{k1}, x_{k2}, x_{k3}\}$, namely, the resulting network structure is (l) in Fig. 4. This structure has the smallest value of minimum description length among all possible network structures.

Furthermore, if using B&B technique, we can predict that there are two search branches in the search tree which satisfy with the pruning condition in light of Eq. (9), then there are two search branches to be pruned, Fig. 5 on page 9 shows the instance of B&B-MDL search tree from top to down. In this case, the search list π_1 traces as the forth column in Table 1, six scoring evaluations are avoided. The search result is still that the parent set is $\{x_{k1}, x_{k2}, x_{k3}\}$.

In the end, if we rank the parent nodes in an ascending order of the mutual information, and get the order $x_{k2} \prec x_{k3} \prec x_{k1} \prec x_{k4}$, here there are three search branches in the search tree which satisfy with the pruning condition, therefore the three search branches can be pruned, the EB&B-MDL search tree from top to down is shown in Fig. 5 on page 10. In this case, the search list π_1 traces as the fifth column in Table 1, eight scoring evaluations are avoided. Fortunately the search result is still the same as that of two previous searches.

From this example, we can see that different search techniques have different search efficiencies, and the EB&B-MDL search is most efficient while getting same search result. In other words, the Algorithm 1 can effectively accelerate the search process comparing with the B&B-MDL search algorithm.

4. Empirical study

4.1. The performance of the EI-B&B-MDL algorithm

To evaluate the performance of the proposed algorithm, the benchmark ALARM network is used. It is a medial diagnostic system containing 37 nodes (variables) and 46 arcs. The ALARM database contains 10000 samples, and each variable has two to four possible values. The platform used for conducting following experiments is a PC with PIV 2.0 GHz CPU, 256 M memory, and running under Windows XP. The data sets were stored in an access database. When the confidence degree of χ^2 test is 99.5%, the EI-B&B-MDL algorithm has the same results as [1], which arrives at the best accuracy so far.

Table 1
Search processes of different techniques for a set of four candidate parents nodes

Sequence number	Possible parent set	Search based on MDL value	Search based on B&B-MDL	Search based on EB&B-MDL
1	ϕ	$\surd$	$\surd$	$\surd$
2	x_{k1}	$\surd$	$\surd$	$\times$
3	x_{k1}, x_{k2}	$\surd$	$\surd$	$\times$
4	x_{k1}, x_{k2}, x_{k3}	$\surd$ $(result)$	$\surd$ $(result)$	$\surd$ $(result)$
5	$x_{k1}, x_{k2}, x_{k3}, x_{k4}$	$\surd$	$\surd$	$\surd$
6	x_{k1}, x_{k2}, x_{k4}	$\surd$	$\surd$	$\times$
7	x_{k1}, x_{k3}	$\surd$	$\surd$	$\times$
8	x_{k1}, x_{k3}, x_{k4}	$\surd$	$\surd$	$\times$
9	x_{k1}, x_{k4}	$\surd$	$\surd$	$\times$
10	x_{k2}	$\surd$	$\times$	$\surd$
11	x_{k2}, x_{k3}	$\surd$	$\times$	$\surd$
12	x_{k2}, x_{k3}, x_{k4}	$\surd$	$\times$	$\surd$
13	x_{k2}, x_{k4}	$\surd$	$\times$	$\surd$
14	x_{k3}	$\surd$	$\times$	$\times$
15	x_{k3}, x_{k4}	$\surd$	$\times$	$\times$
16	x_{k4}	$\surd$	$\surd$	$\surd$

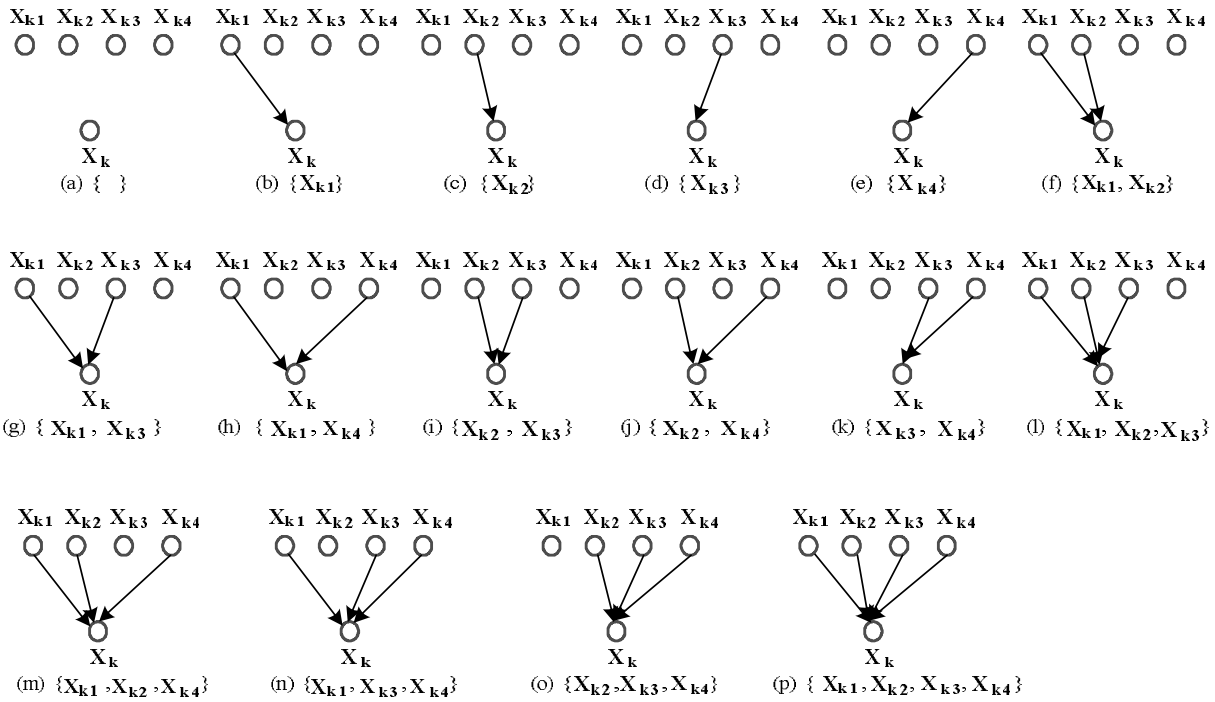

Fig. 4. Possible network structures for a node x_k with four candidate parent nodes.

4.1.1. Experiments on different strategies of independence tests

For different sample sizes, both algorithms are applied, I-B&B-MDL and EI-B&B-MDL, in order to compare their phase results. The results are shown in Table 2 on the next page. The compression effect of EI-B&B-MDL, whose number of arcs and maximal number of parent nodes are larger than those of I-B&B-MDL, is not as good as that of I-B&B-MDL. However, as the time saved in CI testing is much longer than the time increased in searching (especially when the sample size is large), hence our strategy for CI tests can improve the time performance of initial I-B&B-MDL algorithm.

4.1.2. Experiments on different methods of sort order

For a search tree with a given topological structure, there are various selecting order for candidate parent

Table 2
The phase results of both algorithms

Sample capacity	The number of arcs after order-0 test		The number of arcs after order-1 test		The maximal number of parent nodes		Test time (seconds)		Search time (seconds)	
	I-B&B -MDL	EI-B&B -MDL	I-B&B -MDL	EI-B&B -MDL	I-B&B -MDL	EI-B&B -MDL	I-B&B -MDL	EI-B&B -MDL	I-B&B -MDL	EI-B&B -MDL
5000	282	282	61	120	7	9	21.313	3.968	3.813	14.704
6000	284	284	61	119	7	9	23.906	4.625	4.485	17.406
7000	286	286	61	121	7	9	28.110	5.407	5.141	12.625
8000	287	287	64	127	7	10	33.516	6.187	8.500	15.953
9000	289	289	65	123	8	13	38.047	6.969	11.187	19.187
10000	293	293	66	127	7	13	41.593	7.797	9.532	21.312

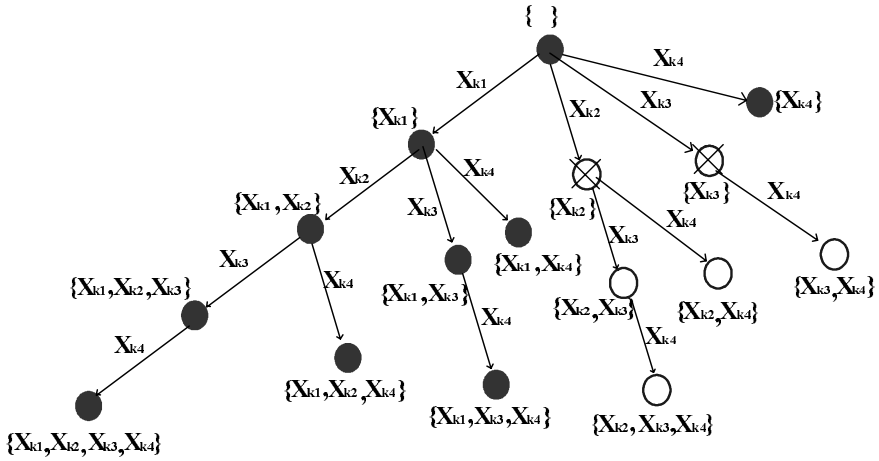

Fig. 5. An example of B&B-MDL search tree for a set of four candidate parents nodes.

nodes. Different selecting orders can lead the place of status point in the tree to change, make the predicted evaluation value in every branch different, thus bring about changing the pruning places and amount. As a result, they induce the algorithm search efficiency different.

In the worst case (when no branch is pruned), the B&B search is equal to the exhaustive search. Furthermore, as there are extra costs of predicted estimate, the time performance of B&B algorithm may go to the bad. Thus it is a key problem for a search how to find a good selecting order of parent nodes by means of heuristic knowledge. Table 3 demonstrates that different sort order methods produce results on time performance under the same conditions of CI tests.

From experimental results, we find that the time performance of the search algorithm can be significantly improved when we rank the parent nodes in an ascending order of the mutual information and then search with the MDL's scoring. That is, with the ascending order of candidate parent nodes, there are many branches to be pruned, so the algorithm reduces many blindly scorings and searchings, enhances the search efficiency. This is the reason that the sort order strategy is adopted in our EI-B&B-MDL algorithm.

4.1.3. The performance of the EI-B&B-MDL algorithm

Under the same conditions, we perform the I-MDL algorithm based on order-0&order-1 CI tests, the I-B&B-MDL algorithm based on order-0&order-1 CI tests, and the EI-B&B-MDL algorithm, respectively. Figure 7 shows the time performance of different algorithms on the ALARM data sets. The results show that the running time of our algorithm is lower than that of other algorithms over the whole scope of sample capacity. And moreover, the advantage is very obvious when the data set is large, namely, the bigger the sample size is, the more obvious the improvement is. This is because our algorithm not only can enhance the search efficiency using heuristic knowledge of the mutual information, but also can reduce the number of independence tests and database passes by means of few lower order's CI tests. In a word, the fact that the running time of our algorithm increases so slowly, suggests that our algorithm will be able to handle very large data sets, so the EI-B&B-MDL algorithm is promising.

Table 3
The comparison of time performance of different sort order methods

Sample capacity	Given node order	Reverse node order	Descending order by MI	Ascending order by MI
5000	26.046	29.156	31.438	18.672
6000	21.329	40.282	35.860	22.031
7000	25.407	46.891	48.798	18.032
8000	28.453	69.719	65.969	22.140
9000	44.939	105.907	162.751	26.156
10000	51.063	123.546	198.375	29.109

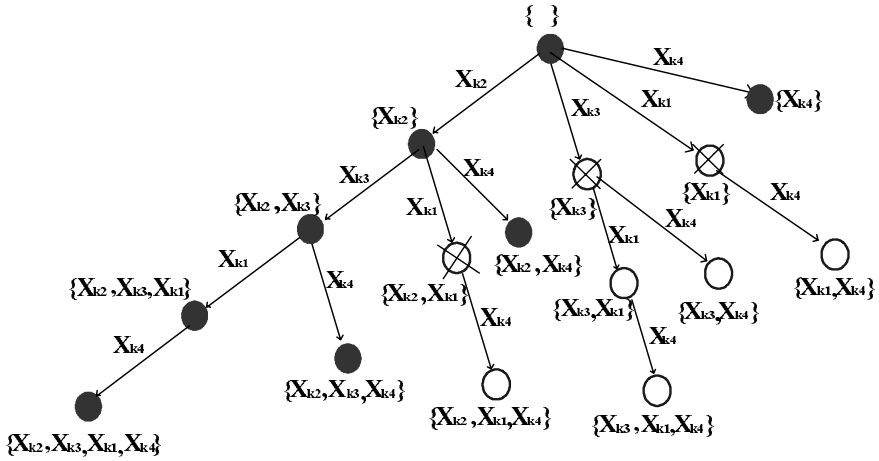

Fig. 6. An example of EB&B-MDL search tree for a set of four candidate parents nodes.

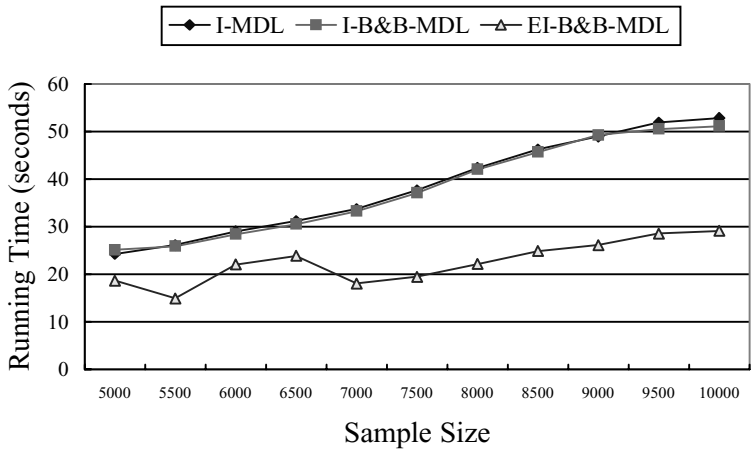

Fig. 7. The comparison of the time performance for different algorithms.

4.2. Application in the personalized recommendation of a B2C portal

We have applied the EI-B&B-MDL algorithm to learning the customer shopping model in the CSM-based (based on Customer Shopping Models) recommendation system [7], which is a kind of personalized recommendation system used to find unsold commodities to online customers, and conducted our experiments with real world data. Tests with real world data allow us to evaluate whether or not the method is potentially useful in practice.

By a recommendation engine of the probability inference, we can get a commodity recommendation set for each online customer. Then we make use of two metrics of the precision and the coverage to evaluate the

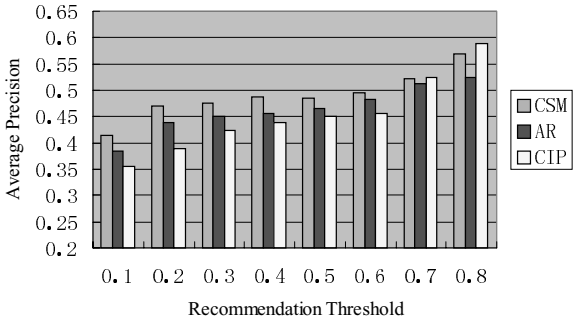

Fig. 8. The average precision of different recommendation methods.

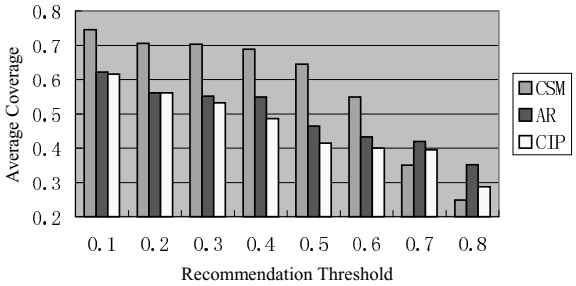

Fig. 9. The average coverage of different recommendation methods.

effectiveness of the recommendation system in a B2C portal. Because the accuracy of our learning algorithm is higher than the former algorithm, the effectiveness of the CSM-based recommendation method is improved in some sort.

The average precision and coverage of different recommendation methods are shown in Figs 8 and 9, respectively. From these results, we can see that our method is better than the AR method (based on association rules) [12] and the CIP method (based on condition independence possibility) [9]. The reason is that the CSM-based recommendation method virtually is a hybrid of the AR and CIP methods. A local structure of each node in a CSM can be seen as a kind of association relations, and parameters of each node denote the conditional probabilities between nodes of commodity category. As the knowledge representation of a BN is compact, our method can avoid the defects of the AR method while the rules are not integrated, and also overcome the shortcomings of the CIP to some extent, which uses strong constraints of the conditional independence.

5. Conclusions

In this paper, we studied the architecture of an intelligent B2C portal, and proposed a new improved algorithm, EI-B&B-MDL, for learning Bayesian networks efficiently and effectively. The algorithm first makes full use of few order-0&1 CI tests to obtain an original network graph, which reduces the number of independence tests and database passes while effectively restricting the search space. By means of the heuristic knowledge of the mutual information, the algorithm fulfills sort order of candidate parent nodes, which increases the cut-offs of B&B search tree and accelerates the search process. In experiments on the benchmark ALARM data sets, the improved algorithm is faster than some hybrid algorithms while keeping with high accuracy, and it is suggested that large data sets can be handled. Hence the improved algorithm is a powerful and efficient algorithm.

We have applied the EI-B&B-MDL algorithm to a personalized recommendation system based on CSM (Customer Shopping Model) and compared the effectiveness of the recommendation method with that of other methods. The experimental results show that the CSM-based recommendation method outperforms other methods by its overall performance. This study also shows that the EI-B&B-MDL is a promising data mining approach to personalized recommendations in a B2C portal. Our future work includes finding more rigorous bound expression and improving the algorithm of recommendation engine.

Acknowledgments

This work is supported by the NSFC major research program: Basic Theory and Core Techniques of Non-Canonical Knowledge (60496322, 60496327), the Institute of Beijing Educational Committee research program : Uncertain Knowledge Representation and Reasoning in Web Intelligence (KM 200610005020), Beijing University of Technology Youth Scientific Research Foundation and Doctor Scientific Research Foundation.

References

[1] J. Cheng, R. Greiner, J. Kelly, D. Bell and W. Liu, Learning belief networks from data: An information theory based approach, *Artificial Intelligence* **137** (2002), 43–90.

[2] C.M. Chen, Incremental Personalized Web Pages Mining Utilizing Self-organizing HCMAC Neural Network, *Web Intelligence and Agent Systems* **2**(1) (2004), 21–38, IOS Press.

[3] N. Friedman, I. Nachman and D. Peer, *Learning Bayesian Network Structures from Massive Datasets: The Sparse Candidate Algorithm*, in: Proc. the Fifteenth Conf. on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 1999, pp. 206-215.

[4] J.P. Gant and D.B. Gant, *Web Portal Functionality and State Government e-Service*, in: Proc. 35th Hawaii Inter. Conf. on System Sciences, Island of Hawaii, January, 2002.

[5] D. Heckerman, A tutorial on learning Bayesian networks, in: *Learning in Graphical Models*, M.I. Jordan, ed., Kluwer, 1996, pp. 301–354. .

[6] E. Herskovits and G. Cooper, A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* **9**(4) (1992), 309–347.

[7] J.Z. Ji, C.N. Liu, Z.Q. Sha and N. Zhong, Personalized Recommendation Based on a Multilevel Customer Model, *International Journal of Pattern Recognition and Artificial Intelligence* **19**(7) (2005), 895–916.

[8] J.Z. Ji, C.N. Liu, J. Yan and N. Zhong, *Bayesian Network Structure Learning and Its Application to Personalized Recommendation in a B2C Portal*, in: Proc. of 2004 IEEE/WIC/ACM Inter. Conf. on Web Intelligence, IEEE-CS Press, Beijing, 2004, 179–184.

[9] B. Kitts, D. Freed and M. Vrieze, *Cross-sell: A Fast Promotion Tunable Customer-item Recommendation Method Based on Conditionally Independent Probabilities*, in: Proc. the Sixth ACM SIGKDD Inter. Conference, 2000, 437–446.

[10] M. Luis, de Campos and J.F. Huete, A new approach for learning belief networks using independence criteria, *International Journal of Approximate reasoning* **24**(1) (2000), 11–37.

[11] P. Lingras, M. Hogo and M. Snorek, Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets, *Web Intelligence and Agent Systems* **2**(3) (2004), 217–230, IOS Press.

[12] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, *Effective Personalization Based on Association Rule Discovery from Web Usage Data*, in: Proc. ACM Workshop on Web Information and Data Management, 2001, 103–112.

[13] L. Qiang, T.Y. Xiao and G.X. Qiao, An Improved Bayesian Networks Learning Algorithm, *Journal of Computer Research and Development* **39**(10) (2002), 1221–1226.

[14] J. Suzuki, Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: An Efficient Algorithm Using the B&B Technique, *IEICE Transactions on Information and Systems* **E82-D**(2) (1999), 356–367.

[15] J. Suzuki, Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties, *IEICE Transactions on Fundamentals* **E82**(10) (1999), 2237–2245.

[16] J.B. Schafer, J.A. Konstan and J. Riedl, E-commerce Recommendation Applications, *Data Mining and Knowledge Discovery* **5**(1/2) (2001), 115–153.

[17] M.L. Wong, W. Lam and K.S. Leung, Using Evolutionary Programming and Minimum Description Length Principle for Data Mining of Bayesian Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(2) (1999), 174–178.

[18] M.L. Wong, S.Y. Lee and K.-S. Leung, *A Hybrid Approach to Discover Bayesian Networks from Databases Using Evolutionary Programming*, in: Proc. 2002 IEEE Inter. Conf. on Data Mining, 2002, 498–505.

[19] Y.Y. Yao, N. Zhong, J. Huang, C. Ou and C. Liu, Using Market Value Functions for Targeted Marketing Data Mining, *International Journal of Pattern Recognition and Artificial Intelligence* **16**(8) (2002), 1117–1131.

[20] Y.H. Tian, T.J. Huang and W. Gao, Two-phase Web site classification based on Hidden Markov Tree models, *Web Intelligence and Agent Systems* **2**(4) (2004), 249–264, IOS Press.

[21] N. Zhong, J. Liu, Y.Y. Yao and S. Ohsuga, *Web Intelligence* (*WI*), in: Proc. 24th IEEE Computer Society International Computer Software and Applications Conference, 2000, 469–470.

[22] N. Zhong, J. Liu and Y.Y. Yao, In Search of the Wisdom Web, *IEEE Computer* **35**(11) (2002), 27–31.

[23] N. Zhong, J. Liu and Y.Y. Yao, Web Intelligence (WI): A New Paradigm for Developing the Wisdom Web and Social Network Intelligence, in: *Web Intelligence*, N. Zhong, ed., Springer, 2003, pp. 1–16.

[24] N. Zhong, Y.Y. Yao, C. Liu, J. Huang and C. Ou, Data Mining for Targeted Marketing, in: *Intelligent Technologies for Information Analysis*, N. Zhong and J. Liu, eds, Springer, 2004, pp. 109–131.