Bayesian Networks Structure Learning and Its Application to Personalized Recommendation in a B2C Portal

Junzhong Ji, Chunnian Liu, Jing Yan
College of Computer Science and Technology, Beijing University of Technology
Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology
Beijing 100022, China
jjz01@bjut.edu.cn

Ning Zhong
Department of Information Engineering, Maebashi Institute of Technology,
460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan
zhong@maebashi-it.ac.jp

Abstract

Web Intelligence (WI) is a new and active research field in current AI and IT. Personalized recommendation in an intelligent B2C portal is an important research topic in WI. In this paper, we first investigate the architecture of a B2C portal from the viewpoint of conceptual levels of WI. Aiming at data mining of knowledge-level in a B2C portal, we present a new improved learning algorithm of Bayesian Networks, which consists of two major contributions, namely, making the best of lower order Conditional Independence (CI) tests and accelerating search process by means of sort order for parent nodes. By a number of experiments on ALARM datasets, we find that the proposed algorithm is both more efficient and effective than others. We have applied this algorithm to a commodity recommendation system in a B2C portal. Our experimental results demonstrate that the recommendation method based on a Customer Shopping Model (CSM) produced by the new algorithm outperforms some traditional ones in rates of coverage and precision.

1 Introduction

Web Intelligence (WI) is a new and active research field in current AI and IT [1, 2]. WI technologies bring many revolutionary changes for scientific research and development on the Internet. In particular, there are great potentials for WI to make useful contributions to e-commerce, and e-commerce intelligence becomes one of major subfields in WI. For B2C e-commerce, the shopping activity that involves private customers is undergoing a significant revolution. It is a focus on how to find meaningful knowledge about customer shopping behaviors and enable a portal intelligence.

An intelligent B2C portal is a multi-functional gateway, which provides various information and services available at a single website [3]. The basic prerequisite for the success of a B2C portal is to fulfill the personalized recommendation for commodities. As the personalized needs of customers are rapidly increasing, the personalized recommendation for customers has become an important issue for customers and business alike. For a customer, a personalized portal creates a true customized shop for each one, which will save every customer's time and make every customer go to needs nonstop. On the other hand, an intelligent portal can capture potential customers' demands and adapt its marketing strategies, which increases many chances of cross-selling and enhances the competitiveness of the business site. Thus, many researchers focus on how to find customers' behavior patterns and realize the personalized recommendation in a B2C portal.

Many recommendation systems have been studied over the last decade. Existing techniques include nearest neighbor algorithm, Bayesian analysis, clustering technique, and many others. Generally speaking, we can divide these techniques into two categories, one is called user-based technique which is based on user relations, such as the nearest neighbor algorithm and 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 [4], we proposed and experimentally evaluated a new approach in making commodities recommendation, which is based on a Customer Shopping Model (CSM) of Bayesian networks by learning from commerce transaction data. This approach formalizes commodity recommendation as knowledge representation of customer shopping information and knowledge inference process. In order to enhance performance of the learning algorithm, we present an improved Bayesian Networks (BN) learning algorithm in this paper. The modified algorithm is more efficient than the original one which is used in the personalized recommendation system. Moreover, as the modified algorithm has higher accuracy, it can preserve precision and coverage of the recommendation system.

The rest of the paper is organized as follows. Section 2 glances at the architecture of a B2C portal from the viewpoint of conceptual levels of WI. In Section 3, we review the personalized recommendation method based on BN. In Section 4, we describe the improved learning algorithm of BN in detail. Section 5 reports our experimental results. Finally, we conclude the paper in Section 6.

2 The Architecture of a B2C Portal

A B2C portal enables a business company to create a virtual marketplace on the Web where abundant information and services of commodities are provided for customers. Although specific features of different portals are different, the common functionalities of the portals for B2C e-commerce are the same, such as usability, customization, openness, and transparency [3]. These common features are implemented by using WI technologies and are evolved with the development of WI technologies. In light of four conceptual levels of WI [2, 5], the architecture of a B2C portal can be shown in Figure 1.

(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. The internet-level provides general communication infrastructure with protocol softwares.

(2) **Interface-level contact.** A B2C portal first serves as a Web server. Through the server, millions of Web customers can access the vast of information, services, and applications available on the business site. In the meantime, the server collects the shopping information of customers, and traces customers surfing behaviors. It is obvious that interface-level provides customers with a single point of contact for online access to commodity information and resource.

(3) **Application-level service.** The application server fulfills many specific projects, such as personalized information search, proactive recommendation, mutual customization, shopping manual, commodity order, and so on. These functional modules support the B2C portal to achieve several application services, and meet the needs of customers by way of Web server. Though almost all intelligent functions are carried out in the level, many capabilities in these services depend on knowledge discovery in customers' data. Thus, the application-level acts as a link man between Web server and data integration and mining.

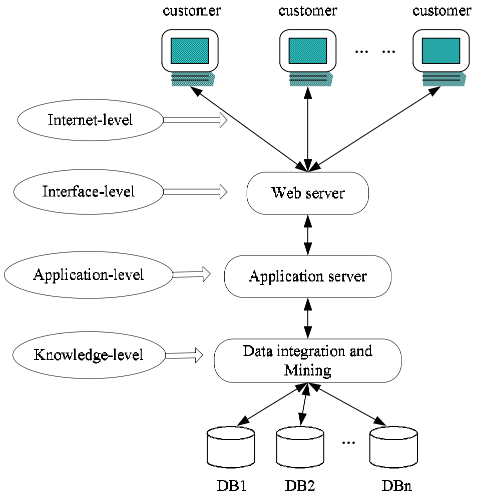

Figure 1. The architecture of a B2C portal based on conceptual levels of WI

(4) **Knowledge-level discovery.** There are various data sources in a B2C portal, which include Web server log, customer's basic information, shopping business data, and Web metadata, and so on. Hence, the data integration and mining is a prerequisite step to all of the techniques that provide customers with personalized service. With the exception of processing and managing data, the knowledge-level performs tasks of data integration and mining in a B2C portal. For a high functioning B2C portal, this level is the key element of an intelligent portal.

Aiming at personalized recommendation in a B2C portal, we stress knowledge discovery in customers' transaction databases in this paper, and discuss a new learning algorithm of Bayesian networks.

3 The Personalized Recommendation Based on CSM

Bayesian Networks (BN) offer a powerful framework that can handle uncertainty and tolerate noise, and the BN model has been widely applied to data mining systems. If we take each attribute of commodity category in transaction databases T as a network variable, we can use BN to represent the customers shopping model (CSM). In light of customer shopping behavior in T, the CSM can capture mutual relationships between commodity categories. Let $\{C_1, C_2, \ldots, C_n\}$ be the variables of the model and $Pa(C_i)$ be the parent set of variable C_i in the network. The joint probability distribution, P, can be obtained by means of the following expression:

$$P(c_1, c_2, \cdots, c_n) = \prod_{i=1}^{n} P(c_i|pa(c_i)) \tag{1}$$

where c_i denotes a value in the domain of variable C_i, $pa(c_i)$ denotes any assignment of compounding values to the variables in the set $Pa(c_i)$. In the CSM, the variable C_i is a boolean variable, a true value means that a customer has purchased a commodity in this category, and a false value quite means the contrary.

In [4], we discussed in detail the personalized recommendation method that consists of data preprocessing, model learning, and a recommendation engine. After data preprocessing, the system acquires the CSM by means of a learning algorithm of BN, which can reflect the individual shopping behaviors and habits of the customer. Then the recommendation engine based on probability inference takes the effect to generate a personalized recommendation set of commodity categories for each customer. Because the algorithm adopted in the recommendation system needs to perform large numbers of conditional independence (CI) tests and some test results may be inaccurate, it is difficult to adapt to the real dynamic business environment. In order to overcome the drawbacks, we propose, in this paper, a modified hybrid algorithm of BN called EI-B&B-MDL.

4 EI-B&B-MDL

A BN is an annotated directed acyclic graph that encodes a joint probability distribution over a finite set of random variables. Generally speaking, there are two approaches for finding structure [6]. The first approach poses learning as a constraint satisfaction problem. By estimating properties of conditional independence (CI) among the attributes in the data, people build a network that exhibits the learned relationships of dependence and independence. The second approach poses learning as an optimization problem. By means of a statistically motivated score, people search for an optimal structure that is optimum to the observed data. As each has its own weakness, there has grown a lot of hybrid algorithms in the last decade [6, 7, 8].

4.1 The MDL-based Learning Algorithm

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 basic flow of the algorithm can be described as follows.

If the finite set of random variables for a BN is denoted by $\chi = \{X_1, \cdots, X_n\}$, where each variable X_j, $j \in J = (1, 2, \cdots, n)$, may take on values from a finite set $A^j = \{0, 1, \cdots, a^j - 1\}$($a^j \geq 2$: some integer), the learning of the network structure virtually is the identifying of parent sets $\{\Pi^1, \Pi^2, \cdots, \Pi^n\}$, where Π^j is the set of nodes that a node X^j depends on. Given a sample set $x^{(i)}$=$\{x^1, \cdots, x^n\}$ of χ, $i \in${1,2,$\cdots$,N}, where N is the sample capacity, then $x^{(i)} \in A = \prod_{j \in J} A^j$, $x^N = x^{(1)} x^{(2)} \cdots x^{(N)} \in A^N$. While G is the set of possible network structures, $g \in G$, then the description length L (g, x^N) of BN is expressed as [9]:

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

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

$$H(j, g, x^N) = \sum_{s \in S(j,g)} \sum_{q \in A^j} -n[q, s, j, g] \log \frac{n[q, s, j, g]}{n[s, j, g]} \tag{3}$$

$$k(j, g) = (a^j - 1) \prod_{k \in \phi(j)} a^k \tag{4}$$

$$n[s, j, g] = \sum_{i=1}^{N} I(\pi_i^j = s) \tag{5}$$

$$n[q, s, j, g] = \sum_{i=1}^{N} I(x_i^j = q, \pi_i^j = s) \tag{6}$$

where $\phi(j) = \{1, 2, \cdots, j - 1\}$ is the sequence number set of parent nodes, $S(j, g)$ is the set of the corresponding realization.

Based on the above preparation, the problem of learning BN becomes a search problem for a structure with MDL metric. Generally speaking, an exhaustive recursive search can be applied to the MDL-based search procedure. However, the cost of those evaluations is acute for massive datasets since this search examines all possible local changes in the set of parent nodes.

4.2 The B&B-MDL-based Learning Algorithm

In order to reduce the computational complexity for empirical entropy, Suzuki proposed a Branch&Bound-MDL-based learning algorithm called (B&B-MDL) [9], which can avoid worthless recursive calls for some branches of search tree by estimating the 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 a lower cost, then the further recursive calls in current step can be avoided.

Although the conditional probability $k(j,g)$ increases along with the number of parent nodes increasing, the value of empirical entroy 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

$$H(j,g,x^N) \leq \frac{k(j,g)(a^q-1)}{2}\log N, \tag{7}$$

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

4.3 The I-B&B-MDL-based Learning Algorithm

Although the B&B-MDL-based learning algorithm can improve the MDL-based learning algorithm only from the viewpoint of search, there still are most of the candidates considered during a search process 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 [8]. 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 the 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, we only consider k possible parents in each search. Since the search space is significantly restricted, the search performs faster than B&B-MDL.

Unfortunately, when the number of nodes is large, there are two major problems for I-B&B-MDL. Since the number of tuples of each conditional set is too large, it is expensive that the cost of collecting various statistics about data and computing mutual information even if only performing lower order independence tests. Moreover, the algorithm cannot ensure that there are enough pruned subtrees to make I-B&B-MDL more efficient than B&B-MDL, because there is extra cost of CI tests.

4.4 The EI-B&B-MDL-based Learning Algorithm

In order to overcome the above drawbacks, we propose an improved algorithm called enhanced I-B&B-MDL (EI-B&B-MDL). There are two major contributions in this algorithm. Firstly, order-0 and partial order-1 independence tests are used to obtain an original graph of the network, which reduces the number of independence tests and database passes while effectively limiting the search space. Secondly, by means of the heuristic knowledge of mutual information, sort order for candidate parent nodes increases the cut-offs of B&B search tree and accelerates search process. In order to account distinctly for our algorithm, we give the definition of order-1 unilateral double-connection.

Definition 1. Given an arc between two nodes X_i and X_j in a Bayesian network, if there is another path which is the same direction as the arc, and the path only include an extra node X_k, we call this acyclic subgraph as order-1 unilateral double-connection.

The major steps of the learning algorithm are as follows.

Step 1. Given a node ordering, build the initial graph G_0 by means of order-0 CI tests, in which each arc meets to the constraint condition $I(X_i,X_j) \geq \varepsilon$ (ε is the test threshold), where

$$I(X_i,X_j) = \sum_{x_i,x_j} P(x_i,x_j)\log\frac{P(x_i,x_j)}{P(x_i)P(x_j)} \tag{8}$$

and record the mutual information of each arc in G_0.

Step 2. For every order-1 unilateral double-connection in G_0, conduct order-1 CI tests in light of Eq. (9)

$$I(X_i,X_j|X_k) = \sum_{x_i,x_j,x_k} P(x_i,x_j)\log\frac{P(x_i,x_j|x_k)}{P(x_i|x_k)P(x_j|x_k)} \tag{9}$$

and remove the invalid arc that does not pass a χ^2 test. As a result, simplify G_0 to G_1.

Step 3. For each node X_j, ascertain candidate parents $\Pi(X_j)$ of the node 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 (EB&B-MDL) technique to search from top to down, find a $\Pi(X_j)$ with the minimum MDL score and confirm the local optimized structure of X_j. Let $\pi_1 = \phi$, $p_1 = \frac{a^j-1}{2}\log N$, the main procedure of the EB&B-MDL algorithm is shown in the top of next page.

5 Empirical Study

5.1 The Performance of the EI-B&B-MDL Algorithm

To assess the performance of the proposed algorithm, we use the benchmark dataset of ALARM network, which is a

```
Algorithm: EB&B-MDL($\pi_1, p_1, MDL_1, \Pi_1$)
            /$*$ $\pi_1$ : the initial set of parents
               $p_1$ : the initial complexity description
               $MDL_1$ : the optimization score
               $\Pi_1$ : the set of parents after this search $*$/
  Begin:
  1. Compute the empirical entropy $H_1$ and
     $MDL_1 \leftarrow H_1 + \pi_1$; $\Pi_1 \leftarrow \pi_1$;
  2. if $\pi_1 = \Phi$ then $j \leftarrow 0$ else $j \leftarrow$ the last element in $\pi_1$;
  3. For $j + 1 \leq q \leq k$
      /$*$ 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
       sort ascending of candidate parents $*$/
       $p_2 \leftarrow p_1 \times a^q$;
       /$*$ update complexity description of node $*$/
       if $H_1 > p_1 \times (a^q - 1)$ then
            EB&B-MDL($\pi_2, p_2, MDL_2, \Pi_2$);
       /$*$ 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.
```

medial diagnostic system containing 37 nodes and 46 arcs. The platform used for conducting following experiments is a PC with PIV 2.0GHz CPU, 256M memory, and runs under Windows XP. The datasets were stored in an Access database.

Under the same condition, 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 2 shows the time performance of different algorithms in experiments on currency databases of Alarm. From the results, we can see that the running time of our algorithm is less than that of other algorithms over the whole scope of sample capacity; and the bigger the sample size is, the more obvious the difference is. Moreover, the fact that the running time of our algorithm increases so slowly suggests that our algorithm will be able to handle very large datasets, so the EI-B&B-MDL algorithm is promising.

5.2 Application in the Personalized Recommendation in a B2C Portal

We have applied the EI-B&B-MDL algorithm to learning the customer shopping model in a CSM-based recommendation system [4], which is a kind of personalized recommendation system used to find unsold commodities to online customers, and conduct 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.

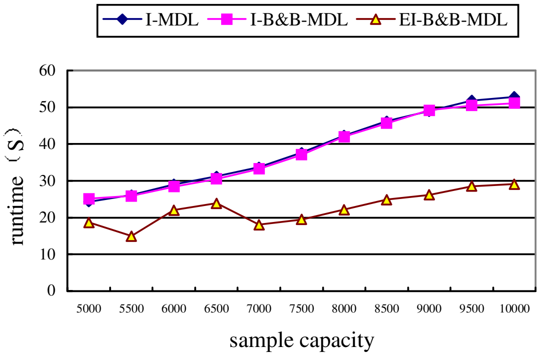

Figure 2. The comparison of time performance of different algorithms

By the probability inference of the recommendation engine, we can get a recommendation set of commodities for each online customer. Then we make use of two metrics of precision and coverage to evaluate the effectiveness of the recommendation system in a B2C portal. Because the accuracy of our learning algorithm is higher than the former algorithms, the effectiveness of the CSM-based recommendation method is improved in some sort.

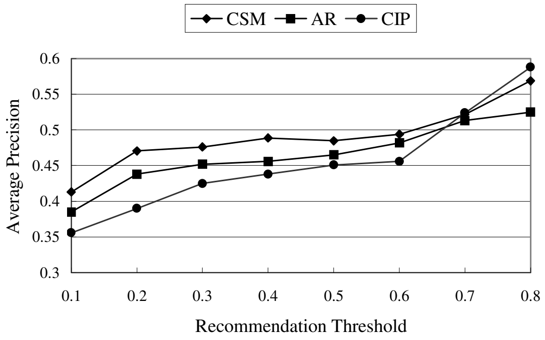

Figure 3. The curves of average precision of different recommendation methods

The curves of average precision and coverage of different recommendation methods are shown in Figures 3 and 4, respectively. From these curves, we can see that our method is better than the AR method (based on association rules) [10] and the CIP method (based on condition independence possibility) [11]. The reason is that the BN-based recommendation method virtually is a hybrid of the AR and

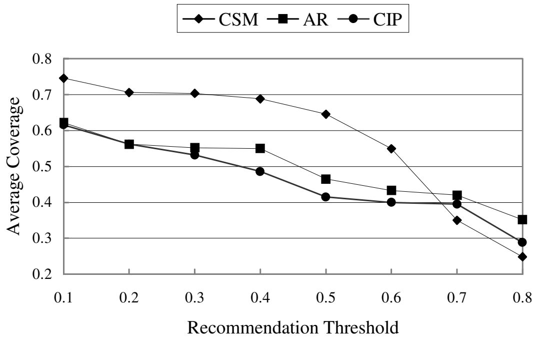

Figure 4. The curves of average coverage of different recommendation methods

CIP methods. A local structure of each node in BN can be seen as a kind of association rules, and parameters of each node denote the conditional probabilities between nodes of commodity category. As the knowledge representation of BN is compact, our method can avoid the defects of the AR method while the rules are not integrated, and can overcome the shortcomings of CIP to some extent, which uses strong constraints of conditional independence.

6 Conclusions

In this paper, we studied the personalized recommendation in a 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 possible graph of the network, which reduces the number of independence tests and database passes while effectively restricting the search space. By means of the heuristic knowledge of mutual information, the algorithm fulfills sort order of candidate parent nodes, which increases the cut-offs of B&B search tree and accelerates search process. In experiments on currency datasets of Alarm, the modified 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 modified algorithm is a powerful and efficient algorithm.

We have applied the EI-B&B-MDL algorithm to a personalized recommendation system based on CSM 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 shows that the EI-B&B-MDL is a promising data mining approach for personalized recommendation in a B2C portal. Our future work includes studying more rigorous bound expression and improving the algorithm of recommendation engine.

Acknowledgments

This work is supported by the Natural Science Foundation of China (60173014), Beijing Municipal Natural Science Foundation (4022003) and Beijing University of Technology Youth Scientific Research Foundation

References

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

[2] N. Zhong, J. Liu, Y.Y. Yao: In Search of the Wisdom Web. IEEE Computer, 35(11) (2002) 27-31.

[3] J.P. Gant, D.B. Gant: Web Portal Functionality and State Government e-Service. Proc. 35th Hawaii Inter. Conf. on System Sciences (2002).

[4] J.Z. Ji, Z.Q. Sha, C.N. Liu, N. Zhong: Online Recommendation Based on Customer Shopping Model in E-commerce, Proc. of 2003 IEEE/WIC Inter. Conf. on Web Intelligence, IEEE-CS Press (2003) 68-74.

[5] N. Zhong, J. Liu, Y.Y. Yao: Web Intelligence (WI): A New Paradigm for Developing the Wisdom Web and Social Network Intelligence. in N. Zhong et al. (eds.) Web Intelligence, Springer (2003) 1-16.

[6] N. Friedman, I. Nachman, D. Peer: Learning Bayesian Network Structures from Massive Datasets: The Sparse Candidate Algorithm. Proc. the Fifteenth Conf. on Uncertainty in Artificial Intelligence (1999) 206-215.

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

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

[9] 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.

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

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