International Journal of Business and Management
www.ccsenet.org/journal.html
Vol. 4, No. 7
July 2009

An Intelligent E-commerce Recommender System Based on Web Mining

Ziming Zeng
School of Information Management, Wuhan University
Wuhan 430072, China
E-mail: zmzeng1977@yahoo.com.cn

The research is supported by MOE Project of Humanities and Social Science in Chinese University (O8JC870011)

Abstract

The prosperity of e-commerce has changed the whole outlook of traditional trading behavior. More and more people are willing to conduct Internet shopping. However, the massive product information provided by the Internet Merchants causes the problem of information overload and this will reduces the customer's satisfaction and interests. To overcome this problem, a recommender system based on web mining is proposed in this paper. The system utilizes web mining techniques to trace the customer's shopping behavior and learn his/her up-to-date preferences adaptively. The experiments have been conducted to evaluate its recommender quality and the results show that the system can give sensible recommendations, and is able to help customers save enormous time for Internet shopping.

Keywords: Recommender system, Web mining, E-commerce

1. Introduction

Nowadays, the advance of Internet and Web technologies has continuously boosted the prosperity of e-commerce. Through the Internet, different merchants and customers can now easily interact with each other, and then have their transactions within a specified time. However, the Internet infrastructure is not the only decisive factor to guarantee a successful business in the electronic market. With the continuous development of electronic commerce, it is not easy for customers to select merchants and find the most suitable products when they are confronted with the massive product information in Internet. In the whole shopping process, customers still spend much time to visit a flooding of retail shops on Web sites, and gather valuable information by themselves. This process is much time-consuming, even sometimes the contents of Web document that customers browse are nothing to do with those that they need indeed. So this will inevitably influences customers' confidence and interests for shopping in Internet.

In order to provide decision support for customers, one way to overcome the above problem is to develop intelligent recommendation systems to provide personalized information services. A recommendation system is a valid mechanism to solve the problem of information overload in Internet shopping. In the shopping websites, the system can help customers find the most suitable products that they would like to buy by providing a list of recommended products. For those products that customers buy frequently, such as grocery, books and clothes, the system can be developed to reason about the customers' personal preferences by analyzing their personal information and shopping records, thus produces the sensible recommendations for them. Therefore, it is of importance to develop the high efficient learning algorithm to capture what customers need and help them what to buy. To date, collaborative filtering has been known to be the most successful technique in analyzing the customer's shopping behavior. Collaboration filtering aims to identify customers whose interests are similar to those of the current customer, and recommend products that similar customers have liked. However, despite its success, the widespread use of collaboration filtering has exposed some problems, among which there are so-called sparsity and cold-start problems, respectively.

In order to overcome the limitations of collaboration filtering, the recommender system based on web mining is proposed in the paper. It utilized a variety of data mining techniques such as web usage mining, association rule mining etc. Based on these techniques, the system can trace the customer's shopping behavior and learn his/her up-to-date preferences adaptively. Therefore, the paper is organized as follows. Section 2 provides the details of the personalized recommender system, with the recommender process relevant to the system. Section 3 gives some experimental result about the recommender quality in our system, and Section 4 gives an overall summary.

2. The Personalized Recommender System

2.1 Overview of the recommender process

The main task of the recommender system is to acquire the customers' up-to-date preferences using web mining techniques, in order to provide decision support for their Internet shopping. Figure 1 gives an overview of the personalized recommender process of the system.

We only select some member customers as the target customers for providing recommender services, considering the efficiency of the system running and maintenance. The recommender process consists of three phases as shown in figure 1. After necessary data cleansing and transformed in the form usable in the system, target customer's preferences are mined first in phase 1. In this phase, how to trace the customer's previous shopping behavior effectively in the system is very important and can be used to make preference analysis. In phase 2, different association rule sets are mined from the customer purchase database, integrated and used for discovering product associations between products. In phase 3, we use the match algorithm to match customer preferences and product associations discovered in the previous two phases, so the recommendation products list, comprising the products with the highest scores, are returned to a given target customer.

2.2 Customer preference mining

This process applies the results of analyzing preference inclination of each customer to make recommendation. To achieve this purpose, the customer preference model is constructed based on the following three general shopping steps in online e-commerce sites.

1) click-through: the click on the hyperlink and the view of the web page of the product.

2) basket placement: the placement of the product in the shopping basket.

3) purchase: the purchase of the product — completion of a transaction.

A simple but straightforward idea of mining the customer's preference is that the customer's preference can be measured by only counting the number of occurrence of URLs mapped to the product from click stream of the customers. According to three sequential shopping steps, we can classify all products into four product groups such as purchased products, products placed in the basket, products clicked through only, and the other products. It is evident to obtain a preference order between products such that {products never clicked} $\prec$ { products only clicked through} $\prec$ {products only placed in the basket} $\prec$ {purchased products}.

Supposing that c_{ij}^{c} is the total number of occurrence of click through of customer i across every product class j. Likewise, c_{ij}^{b} and c_{ij}^{p} are defined as the total number of occurrence of basket placement and purchases of customer i for products class j, respectively. c_{ij}^{c}, c_{ij}^{b} and c_{ij}^{p} are calculated from the raw click stream data as the sum over the given time period, and so reflect individual customer's behaviors in the corresponding shopping process over multiple shopping visits.

From the discussions above, the customer preferences can be acquired from the click stream data and expressed as the preference matrix $C = (c_{ij})$, which is denoted as follows:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} = (c_{ij})_{m\times n} \tag{1}$$

In formula (1), $i = 1,...,M$ (total number of target customers), and $j = 1,...,N$ (total number of product classes). In order to acquire each customer's preference about each product class, matrix element c_{ij} should be computed by formula (2), when considering the three shopping steps.

$$c_{ij} = \alpha \times \frac{c_{ij}^{c}}{\sqrt{\sum_{j=1}^{n}(c_{ij}^{c})^{2}}} + \beta \times \frac{c_{ij}^{b}}{\sqrt{\sum_{j=1}^{n}(c_{ij}^{b})^{2}}} + \gamma \times \frac{c_{ij}^{p}}{\sqrt{\sum_{j=1}^{n}(c_{ij}^{p})^{2}}} \tag{2}$$

In the formula (2), α, β, γ represent the weight adjusting coefficient corresponding to the three shopping steps. It is evident that the weights for each shopping step are not the same. It is reasonable to assign the higher weight to the purchased products than those of products only placed in the basket. Similarly, the higher weight should be give to products placed in the basket than those of products only clicked through. Therefore, we set $\alpha = 0.25$, $\beta = 0.5$, and $\gamma = 1$. In fact, the formula (2) reflects preference order among products, and hence it is the weighted sum of occurrence

frequencies in different shopping steps.

2.3 Product association mining

In this phase, we discover valuable relationships among different products by mining association rules form the customer purchase transactions. Similar to the preference mining process, association rule mining is performed at the level of the product classes. Corresponding to three general shopping steps, the association rules can be generated from three different transaction sets accordingly: purchase transaction set, basket placement transaction set and click-through transaction set. For each transaction set acquired from Web logs, there are three phases to generate associate rules: 1) Set minimum support and minimum confidence; 2) Replacing each product in transaction set with its corresponding product classes; 3) Generating association rules for each transaction set using Apriori.

After association rules are generated, the product association model can also be expressed by a matrix $P = (p_{ij})$, in which each element p_{ij} represents the association degree among the product classes in different shopping step. The matrix $P = (p_{ij})$, $i = 1,...,M$ (total number of product classes), $j = 1,...,N$ (total number of product classes) can be defines as the formula (3).

$$p_{ij} = \begin{cases} 1.0 & if\ i = j \quad \text{(within same classes)} \\ 1.0 & if\ i \overset{p}{\Rightarrow} j \quad \text{(within purchase step)} \\ 0.25 & if\ i \overset{b}{\Rightarrow} j \quad \text{(within basket placement step)} \\ 0.1 & if\ i \overset{c}{\Rightarrow} j \quad \text{(within click-through step)} \\ 0 & otherwise \end{cases} \tag{3}$$

In the formula (3), the first condition indicates that a purchase of a product in a product class implies a preference for other product within the same product class. The second condition indicates that the degree of association in the purchase step is more related to the purchasing pattern of customers than those in the basket placement, so the association degree p_{ij} for purchase can be set 1.0, which is higher than that for basket placement. In the same manner, the association degree p_{ij} for basket placement can be set 0.25, while the association degree p_{ij} for click-through is set only 0.1.

2.4 Matching algorithm for recommendation

In the preceding sections, we have built the model of customer preferences and product association defined by preference matrix and product association matrix, respectively. The final step in the recommendation process is to score each product and produce the recommendation product lists for a specific customer. This score should reflect the degree of similarity between the customer preferences and the product association. There are several methods to measure the similarity, including Pearson correlation, Euclidian distance, and cosine coefficient. In the system, we chose cosine coefficient to measure the similarity. Hence, the matching score σ_{mn} between customer m and product class n can be computed as follows:

$$\sigma_{mn} = \frac{\sum_{k=1}^{N} c_{mk} p_{kn}}{\sqrt{\sum_{k=1}^{N} c_{mk}^2} \cdot \sqrt{\sum_{k=1}^{N} p_{kn}^2}} \tag{4}$$

In the formula (4), $C^{(m)}$ is a row vector of the $M \times N$ customer preference matrix C, and $P^{(n)}$ is a row vector of the $N \times N$ product association matrix P. Here, M refers the total number of target customers and N denotes the total number of product classes. So the matching score σ_{mn} ranges from 0 to 1, where more similarity between $C^{(m)}$ and $P^{(n)}$ result in bigger value.

All products in the same product classes have identical matching scores for a given target customer. However, because matching scores are computed at the level of product classes but no at the product level, the single products must be chosen and recommended to the target customer. In the system, the chosen strategy is adopted that for all products in the same classes, those products which were purchased in the latest period would be assumed to be the most popular and the more buyable products. Therefore, we use this choice strategy to provide the recommender services for the target customers.

The whole matching algorithm for recommendation can be expressed as follows:

Algorithm Recommender_generation():

Input: customer preference matrix C, product association matrix P

Output: recommended product lists

Begin

1: Set the number of recommended products as n, the number of recommended product classes as k, such as $k < n$ and n/k is an integer;

2: Calculate the matching score σ_{mn} using the formula (4);

3: Select top-k product classes with the highest σ_{mn} as recommended product classes;

4: **for** each class **do**

5: elect top-n/k latest purchased products as the recommended products to target customer;

6: **end for**

End

3. The experiment

One important issue for evaluating the recommender quality is the extent to which recommendations with higher recommender scores are accepted preferentially over recommendations with lower scores. We address this issue by comparing the distribution of scores computed from the formula (4) for accepted recommendations with the analogous distribution for offered recommendations. The results are shown in Figure 2. The scores for the accepted recommendations are based on 120 products accepted from 50 distinct recommendation lists. The distribution for the offered recommendations is taken from about 300 recommendations made to the customers who accepted at least one recommendation during the preliminary phase of system running.

Figure 2 shows that the scores of the accepted recommendations are higher than the scores of a large number of offered recommendations. For example, 76% of the products placed onto the recommendations lists have scores below 0.1, but only 22% of the accepted recommendations fall in this lower span. The mean scores for the offered recommendations are 0.072, while the mean scores for the accepted recommendations are 0.165. The difference between the two means is 0.093, falls well within the 95% confidence interval (0.089, 0.106) computing using t-test statistical method for the difference between means. These results illustrate that the score computed using the formula (4) is indeed a useful method of a previously unbought product's appeal to the target customer.

4. Conclusions

In this paper, we have developed a product recommendation system to provide personalized information services in making a successful Internet business. The characteristics of the system can be described as follow. First, the customer preference and product association are automatically mined from click streams of customers. Second, the matching algorithm which combines the customer preference and product association is utilized to score each product and produce the recommended product lists for a specific customer. The future work will include compare the suggested methodology in our system with a standard collaborative filtering algorithm in the aspect of buying precision and other recommender performance.

References

Balabanovic M., & Shoham Y. (1997). Fab: content-based collaborative recommendation. *Communications of the ACM*, 40(3): 66-72.

Huang Z, Zeng D, Chen HC. (2007). A comparison of collaborative-filtering recommendation algorithms for e-commerce. *IEEE Intelligent System*s, 22(5): 68-78.

Lawarence R.D., Almasi,G.S., & Kotlyar V., et al. (2001). Personalization of Supermarket Product Recommendations. *Data Mining and Knowledge Discovery*, 5(1-2): 11-32.

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data Mining and Knowledge Discovery*, 5(1-2): 59-84.

Lin, W., Alvarez, S.A., & Ruiz, C. (2002). Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6(1), 83-105.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). *Analysis of recommendation algorithms for e-commerce*. Proceedings of ACM E-commerce Conference.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithm. Proceedings of the Tenth International World Wide Web Conference, pp.285-295.

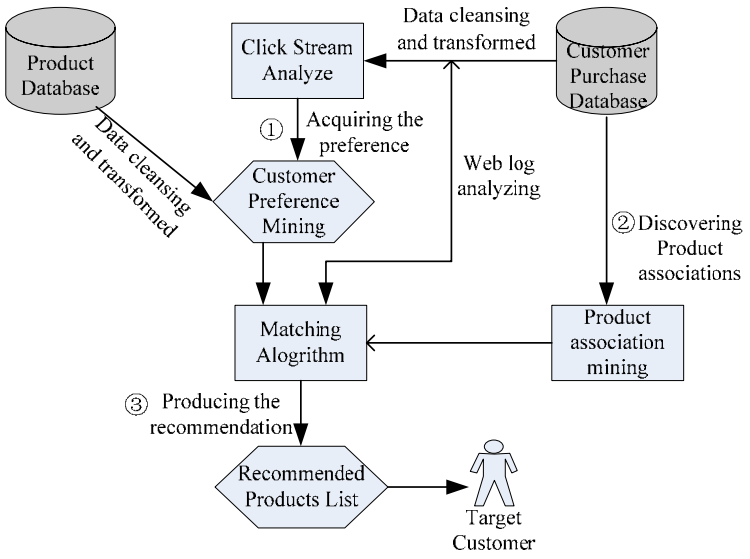

Figure 1. Overview of the recommender process of the system

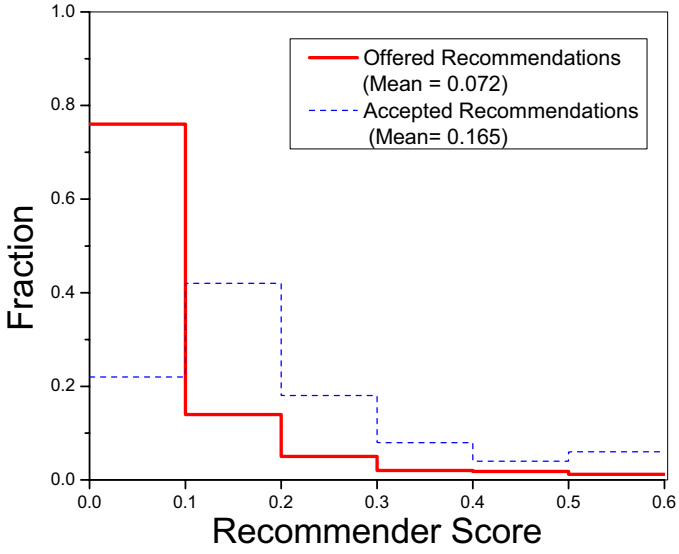

Figure 2. Distribution of scores for offered and accepted recommendations