A Hotel Recommendation System Based on Collaborative Filtering and Rankboost Algorithm

GAO Huming, and LI Weili

*I*nformation management Dept.,
Tianjin University of Economics and Finance,
Tianjin 300222, China
E-mail: gaohm@tjufe.edu.cn

Abstract: A hotel recommendation system based on collaborative filtering method of clustering and Rankboost algorithm proposed in this paper, which can avoid the cold-start and scalability problems existing in traditional collaborative filtering. One can find a hotel quickly and efficiently when he uses this hotel recommendation system.

Keywords: recommendation system; collaborative filtering; Rankboost algorithm

0 Introduction

With the rapid development of network technology and the wide application of e-commerce, consumers are increasingly interested in booking hotels online. However, due to the growth of hotel amount, they have difficulties in finding the desired hotel quickly. Some e-commerce personalization recommendation systems[1] which can generate recommendations based on history buying, browsing and rating data of users have been used in real world . There are two advantages:

(1)Save the time and cost of users in finding goods and enhance the site's trust.

(2)Make potential users or site visitors into loyal customers, and improve the marketing capabilities of e-commerce website[2].

Recommendation technology is the core of a personalized recommendation system. At present, there are many recommendation technologies, such as content-based recommendation technology, the recommendation technology based on association rule, collaborative filtering technology, knowledge-based recommendation technology, effectiveness-based recommendation technology, hybrid recommendation technology and so on. Collaborative filtering technology is one of the most important technologies which are used widely and successfully. It analyzes users' data with similar features who share same personality or interests. Thus, items can be recommended to target users when the group of users prefers them as well. However, CF approaches suffer from three fundamental problems [3]:

(1) Data sparsity. Stated simply, that users do not rate on most items and hence resulted in a sparse user-item matrix. This problem often occurs when there has numerous items but too less rating values or taking place in the initial stage of recommendation system

(2) Cold-start. There are new user problem and new item problem. The cold-start of new user problem is that at begin the recommendations can not be provided because of lack of user's historical transaction data, the personal buying behavior can not be analyzed. The cold-start of new item problem is that if an item has not been rating before, it can not be recommended to users.

(3) Scalability. It is impossible for users to generate recommendations in practice if the users' and item' databases are very big.

Cluster analysis[4] is an effective data-mining technique. A collection of physical or abstract objects become multiple classes containing similar objects after being grouped. Rankboost algorithm is a method of producing highly accurate prediction rules by combining many "weak" rules which may be only

978-0-7695-4008-5/10 $26.00 © 2010 IEEE
DOI 10.1109/MMIT.2010.14

moderately accurate. This paper proposes a hotel recommendation system combining clustering-based collaborative filtering technology and Rankboost algorithm which can relieve the cold-start and scalability issue.

1 Clustering-based collaborative filtering technology

1.1 The basic idea of clustering-based collaborative filtering technology

Firstly, construct a user-item rating matrix, which contains the information of users and ratings on various items. Secondly, make the users that have similar interests into the same cluster by similarity formula. When a target user coming, determine which the cluster target user belongs to and search the target user's nearest neighbor, and then predict rating of the project he hasn't scored according to the formula. Finally, generate the list of recommendation according to the level of predicted rating.

1.2 The process of cluster-based collaborative filtering techniques

Cluster-based collaborative filtering technology combines the collaborative filtering and clustering technology, and can overcome the scalability problems. The details of clustering-based CF approaches are described later. The data source of recommendation system is D=(U,I,R), the collection of users is U={$User_1$,$User_2$,…,$User_m$}, the collection of items is I={$Item_1$,$Item_2$,…,$Item_n$}, the score matrices is R, the element R_{ij} represents the score the user i ratings the project j.

1.2.1 Similarity algorithm.

There are many similarity algorithms[5], such as the cosine similarity, person correlation, the modified cosine similarity and so on. Pearson correlation can be used to compute the similarity between users, the similarity between the user i and j sim (i, j) as follows:

$$\mathrm{sim(i,j)} = \frac{\sum_{k \in I_{ij}} (R_{i,k} - \overline{R_i})(R_{j,k} - \overline{R_j})}{\sqrt{\sum_{k \in I_{ij}} (R_{i,k} - \overline{R_i})^2 \sum_{k \in I_{ij}} (R_{j,k} - \overline{R_j})^2}}$$

Iij is item set which user i and user j had voted, Ri,k is the voting of user i for item k, which reflect the preference of user i for item k, $\overline{R_i}$, $\overline{R_j}$ is respectively the average voting of user i and user j for all item.

1.2.2 Clustering

Retrieve all of the items from the item sets I, and all of the users from the user sets U. Count the ratings of each user and make the users whose score greater than or equal the amount of P as the initial cluster center. Calculate the similarity among the initial cluster centers with similarity algorithms. If the similarity of two users is in the range of (0.7,1), they can exist in a cluster, and choose one as the new cluster center. Finally k-cluster centers can be get, denoted by (W1, W2, ..., Wk); The cluster is represent by C1, C2,..., Ck recorded by the set C = (C1, C2, ..., Ck). The algorithm of searching for a cluster's members is as follows:

repeat
for each user Ui □ U
for each cluster center Wj□ (W1, W2, ..., Wk)
Calculate the similarity sim (Ui, Wi) between Ui and Wj according to the formula (1)
end for
sim (Ui, Wm) = max (sim (Ui, W1), ..., sim (Ui, Wk))
Cm = Cm + Ui
end for
until the membership of cluster don't changes [6]

Clustering process can be carried out offline, so the time of system analysis can be saved.

1.2.3 Determine the target user's cluster and generate recommendation

Determine which cluster the target user belongs to by formula(1). According to the similarity between target user and other users to this clustering, choose the higher similarity of M1 users as target user's nearest neighbor, denoted by a set of Su, and the formula is:

$$P_{u,i} = \overline{R_u} + \frac{\sum_{m \in S_u} sim(u,m) \bullet (R_{m,i} - \overline{R_m})}{\sum_{m \in S_u} sim(u,m)}$$

$\overline{R_u}$ and $\overline{R_m}$ means respectively the average voting of target user u and user m for all items, sim(u, m) indicated the similarity between target user u and user m, Rm, i means the voting of user m for item i.

Calculate the rating target users for the unvoting items and recommend the top M2 items to the target users, namely the Top-M2 recommendation set of the target user [7].

This method has two advantages:

(1) Reduce the number of isolation points by choosing the k users as the initial cluster centers which have higher rating [8].

(2) Reduce the complexity of similarity calculation and the dimension of data; relieve the problem of collaborative filtering. It enables the system to provide higher accurate recommendation online, and increase the revenue of the site.

2 Rankboost algorithm[8]

Rankboost algorithm is a widely used sorting algorithm in statistics. Boosting is a method of producing highly accurate prediction rules by combining many "weak" rules which may be only moderately accurate. In the current setting, we use Boosting to produce a function H : X →R. The basic idea of the algorithm is as follows:

Define X as the event space and each element of X is called an event. Supposed there is a single set of features: f1, f2,…, fn, the event has a single sort in accordance with these features. Define f(x) as a function: if fi (x1)> fi (xo). x1 takes precedence over x0 when sorting according to fi. The purpose of the algorithm is to form a combination of sorting function H(x) : If H(x1) : H(x0), then x1 will take precedence over x0 sorting by H. If X is composed of a disjoint collection of samples X0, X1, the algorithm always recommend the sample X1 prior, and sort the samples belonging to X1. For the two categories of sorting problem, the algorithm is simplified in the calculation of time and space.

The hotel recommendation system can be seen as a sorting problem under more than one feature conditions. The hotel sample set is divided into two disjoint sets: the hotels consumers are satisfied with belong to a set of X1, the hotels consumers are not satisfied with belong to a set of X0. Four categories features can be defined in the hotel recommendation system, as follow: the standard of rooms, hotel services, location, cost-effection. When users choose a hotel, they must rate the four features.

Define f1 as the standard of rooms. The standard is divided into five categories, and the training set X can be divided into five subsets C1, C2, C3, C4, C5 according to the standard of rooms, the formulation of f1(xk) is as follow:

$$f_1(x_k) = \frac{M_k}{M} \Big/ \frac{N_k}{N}$$

Mk means the number the certain users choose the class K hotel, M means the number the certain user select all of hotels, Nk means the number all of the users choose the class K hotel, N means the number all the user selects of hotels。

Define f2 as the feature of hotel services, define f3 as the feature of location, define f4 as the feature of cost-effective. Those features can be set in a similar way with the standard of rooms.

The algorithm can be carried out offline.

3 Personalized Hotel Recommendation System

Booking hotel online may save time and labor which has been widely used. However consumers often have been confused when facing quite many hotels. The hotel recommendation system combining cluster-based collaborative filtering algorithm and the Rankboost algorithm can resolve this trouble. The basic construction of our sysrem is shown in Figure 2:

When a user browses, the system firstly decided whether he/she is a new user or old by retrieving the user database. If it is a new user, the system provides the target user with recommendations by using Rankboost algorithm. Or, the system provides the target user with recommendations by using the cluster-based collaborative filtering algorithm. Because the Rankboost algorithm and clustering process can be carried out offline, we can produce recommend in a very short time.

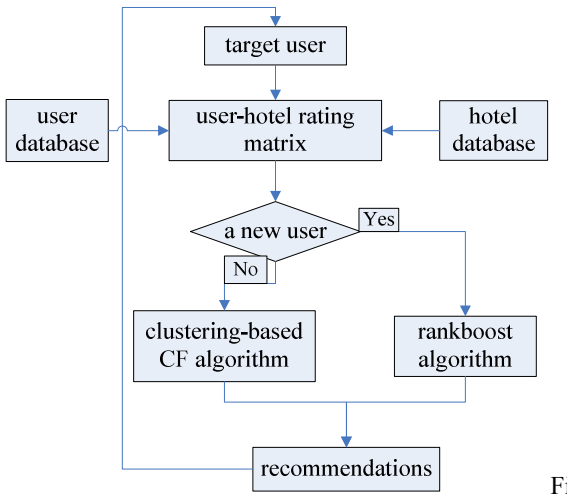

Figure 1 personalized hotel recommendation system

The system provides different recommendations for different users, and users will give actual ratings after using the system which will be automatically saved to the user database and hotel database as basic data.

4 The analysis of experimental results

Experimental data come from a hotel's site (http://www. hoteltravel.com), which extracted the voting of 500 users for 2000 hotel in specific areas 2009/3-2009/10. Provided each user vote for at least one hundred hotels., Through the qualitative and quantitative analysis, rating data are represent with integer that ranges from 0-5 points after pre-processing these data , the hotel which not be scored is represent with 0 points; then set the two threshold values, M1 = 10, M2 = 5.

Firstly, verify effectiveness of the clustering.

The user is divided into 3 to 10 categories. Find the target user's top-10 nearest neighbor set respectively in the cluster target user belongs to and in the entire user space. The feasibility of the method can be illustrated by testing the ratio that the intersection of the two nearest neighbor sets accounts for nearest neighbor set created in the entire user space [10]. The ratio under the different numbers of cluster are shown in Figure 3:

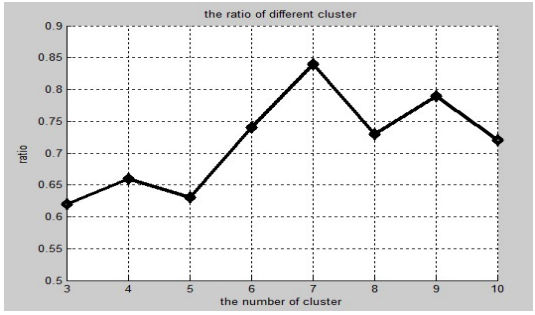

Figure 2 the percentage of similar users in different number of cluster

As can be seen from Figure 3, the ratio are all around 70%, which indicates that the majority of nearest neighbors of the target user are located in cluster the target user belongs to.

Secondly, verify the validity of system

Normally, the MAE(Mean Absolute Error) between predict values and actual values is used to measure the accuracy of system. The specific formula is:

$$MAE = \sum_{i=1}^{N} |p_i - q_i| \Big/ N$$

p_i means the target user's predictive value that derived from the algorithms, qi means the actual score of target user, N means the number of hotel being rated.

Set N=800, the quality of the recommendation is higher when MAE is smaller.

The MAE both traditional recommendation system and personalized recommendation system is shown in Figure 3 in different numbers of cluster.

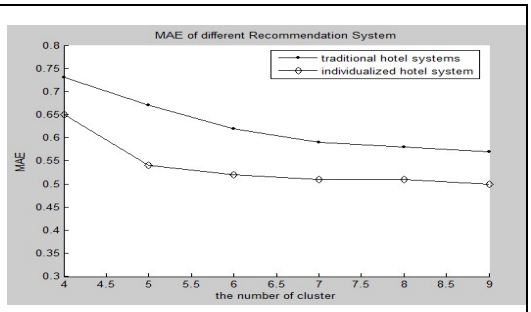

Figure 3

As can be seen from Figure 4, the hotel recommended system presents in this paper provides the target user with a high accuracy recommendation.

5 Conclusion

The information overload become much more serious in e-commerce systems, our hotel recommendation system based on collaborative filtering and Rankboost algorithm can make it convenience for the consumer to book desired hotel. There are various indicators about the satisfaction of hotel which should be considered in choosing cluster centers which should be studied in further research.

REFERENCES:

[1]Deng Xiaohui, Qi Qiang. Analysis of e-commerce recommendation system[J]. Enterprise economy, 2007, (8):116-117.

[2] Zen Chun, Xing Chunxiao, Zhou Lizhu. Personalization services technology [J]. Journal of Software, 2002, 13 (10):1953-1955.

[3] Liu Pingfeng, Nie Guihua, Chen Donglin. E-commerce recommendation system survey [J].Intelligence magazine, 2007, (9):46-47.

[4] QUANT, FUYUKII, SHINICHIH. Improving accuracy of recommender system by clustering items based on stability of user similarity [C]//Proceedings of International Conference on IAWTIC. Washington, DC: IEEE Computer Society, 2006: 61.

[5] Zha Wenqin, Liang Changyong, Cao Lei. Collaborative filtering Recommendation based on user clustering [J]. Computer technology and development, 2009, 19 (6):70-72.

[6] Wang Hongchao, Chen Weiru, LIU Jun. Of goods based on customer clustering recommended method of study [J]. Computer technology and development, 2008, 18 (7):212-214.

[7] Wang Hui, Gao Jun, Wang Zhongting. Personalization services of user-based clustering collaborative filtering recommendation [J]. Computer Applications, 2007, 27 (5):1225-1227.

[8]Freund Y, Iyer R, Schapire RE, et al. An efficient boosting algorithm for combining preferences [J]. Journal of Machine L earning Research, 2003 (4):933 - 969.

[9] Li Tao, Wang Dong, Ye leap, et al. A collaborative filtering recommendation algorithm based on user clustering [J]. Systems Engineering and Electronics, 2007, 29 (7):1178-1182.

[10] Sun Shouyi, Wang Wei. A kind of clustering collaborative filtering based on user-personalized book recommendation system [J]. Modern Information, 2007, (11):140-141.

Supported by the national natural science foundation of China(10771158) and Tianjin natural science foundation(07JCYBJC14300)