Clustering Collaborative Filtering Recommendation System Based on SVD Algorithm

Qilong Ba
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, No. 10, Xitucheng Road, Haidian District, Beijing 100876, P. R. China
baqilong1234@bupt.edu.cn

Xiaoyong Li
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, No. 10, Xitucheng Road, Haidian District, Beijing 100876, P. R. China
lxyxjtu@163.com

Zhongying Bai
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, No. 10, Xitucheng Road, Haidian District, Beijing 100876, P. R. China
bzy66@sina.com

Abstract — **The most important responsibility for every recommendation system is how to make the appropriate personalized recommendation for different customers rapidly and effectively. Collaborative filtering recommendation is one of the most popular methods among the E-commerce system, but it still remains some problems, such as "cold start" and "data sparse". At the same time, more and more users registering on make the real-time and expansibility of a system hard to be kept.**

In the paper, according to a series of problems, such as "data sparse" and bad "real-time" caused by a large number of registering users, we propose a new approach combining the clustering algorithm with SVD algorithm which is widely used in the domain of image-processing into collaborative filtering algorithm. Firstly, we classify the users by using the attributes of them. Then we decompose the rating matrix with the SVD algorithm and reunion them into new rating matrix to calculate the similarity between each pair of users. At last, we take advantage of the similarity to find the nearest neighbors in the collaborative filtering recommendation and predict the ratings of the items to make the recommendation. Our experiments show that the approach cannot only improve the "cold start" and "data sparse" problems but also increase the efficiency and scalability of a system.

Keywords-E-commerce; Collaborative filtering; SVD; Clustering

I. INTRODUCTION

In the recent years, with the springing up of the Internet technology, E-commerce has become a kind of new fashion which is developing quickly. E-commerce, which is a kind of new business model combining Internet technology with commercial activities, is the primary model among the market economy business in the 21th century. Through the E-commerce platform, people can enjoy the convenient of shopping without going out. However, with the expansion of the number of the trading on the E-commerce platform, people can't skim through all the items on the browser in a short time. And on the internet, customers can't get the detail introduction of the items from the salesmen as in reality [1]. People are facing the problem of "information overload" which specifically exists in the period of E-commerce.

Based on the problem of "information overload", the recommendation system emerged as the times require in the 1990s, such as the recommendation of news and the filtering of e-mails from Google [2]. Now, among almost all the E-commerce systems, recommendation technology consists of the important part of the sales-online, such as Amozon, Netflix, douban, taobao, etc. The mainly function of the recommendation system includes: (a) Attract the new users, that is to recommend the items to the potential new customers to convert the visitors to the purchasers. (b) Inspire the old users, that is to recommend more items to the old customers on the basis of the items they has bought before to improve the ability of cross sales on the internet. (c) Improve customers' loyalty to the website [2]. Precision, scalability and the real-time are the important factors to evaluate a recommendation system. In the paper, we utilize clustering through the characteristic value and the SVD algorithms on the collaborative filtering recommendation system, which not only improves the scalability and the precision of the system, but also solve the problem of "information overload" in a sense.

Collaborative filtering algorithm is mainly divided into three categories: user-based algorithm, item-based algorithm and model-based algorithm. Foreign scholars have made some achievements about the three kinds of algorithm. User-based algorithm is suitable for the occasion that there are not so many users. Item-based algorithm is suitable for the occasion that there are not so many items. Model-based algorithm generally takes use of the machine learning approaches, such as bayesian network, semantic indexing, which learns the information of users offline, modeling and then take advantage of the models to predict the rating and recommend

978-1-4673-5000-6/13/$31.00 ©2013 IEEE

This work is supported by the National Nature Science Foundation of China (No. 61003281), Beijing Natural Science Foundation (No. 4111002), Fok Ying Tung Education Foundation (No. 132032) and the Chinese Universities Scientific Fund (No. BUPT 2012RC0210).

the items online. However, all the approaches above have the disadvantages that the matrix that users have rated the items is sparse and too much data in the system can lead to higher algorithm complexity. According to the above problems, we propose a new algorithm combined model-based clustering algorithm with SVD algorithm which is widely used in the area of image processing. From the experiments, we find the algorithm we propose in this paper can solve the problems effectively.

The structure of this paper is as follows: section 1 introduces the recent situation about recommendation and some important recommendation categories; section 2 introduces clustering algorithm; section 3 introduces SVD algorithm; section 4 states our experiments and the relative evaluation results; section 5 draws conclusions and future work.

II. CLUSTERING ALGORITHM

A. *The concept and principle of clustering algorithm*

Clustering is an important data mining technology, which can effectively excavate the potential interests of the groups. Clustering is a kind of analysis process, which refers to polymerize the similarity of some specific real objects or some abstract objects to compose messy objects into some class [4]. It aims at analyzing the similarity among a large amount of data, and then classifying them into different clusters. The classification requires the objects in the same cluster have the largest similarity, and the objects in different clusters have the largest difference [5].

Clustering mainly includes the following steps: data preprocessing; the distance function to measure the similarity between the objects; getting the clustering cluster.

Data preprocessing mainly makes choices about the initial inputted data, including the choice of the amount of the data, the type of the data and the identification of the data. It describes an object with a suitable dimension of feature vector through extracting the characteristic attributes of the object and transforming into explicit attributes. At the same time it is necessary to find out the isolated data, in order to avoid the large deviation of clustering effect [5].

The most important step of the clustering is to balance the similarity between different objects. As describing the characteristics identification and attributes of diversity of the object determine the complexity of feature vector dimensions, it is very important to measure the similarity of different objects in the same feature space. Usually taking a distance function to measure the distance of two feature space, and then calculate the similarity between the objects. The closer two objects are, the higher similarity two objects obtain, and vice versa. Finally, it is concluded that clustering cluster through the quality assessment of clustering results. Data source is a set of mixed and disorderly random data. After calculating the similarity, theoretically similar or diverse data is found out. But another approach must be put forward to evaluate the quality of clustering.

B. *K-means algorithm*

We adopt k-means algorithm in the paper, which is one of the most popular clustering algorithms. The specific ideas as the following:

(a) Select k users as the initial cluster centers randomly, initial the cluster center with the rating of the item.

(b) For the set of the rest users, calculate the similarity between k clustering centers and every user, classify every user to the cluster which has the highest similarity to him.

(c) For the new cluster, calculate the average score of all users rating the item in the cluster, generate the new cluster center.

(d) Repeat the step (b) and (c) until the clustering center don't change anymore.

C. *User characteristic similarity*

In this paper, we use the user characteristic similarity to classify the users.

Usually, everyone has his own characteristic, such as salary, native place, gender, occupation, age, etc. According to the statistics of the consumption patterns of Chinese netizens from the famous consulting company iResearch, see Fig 1, Fig 2.

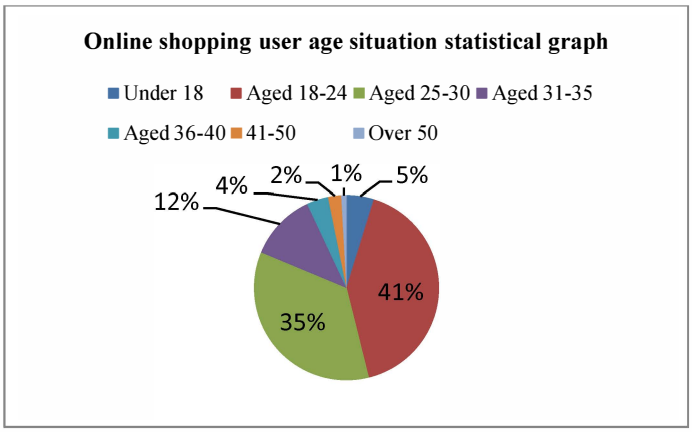

Fig 1 Online shopping user age situation statistical graph

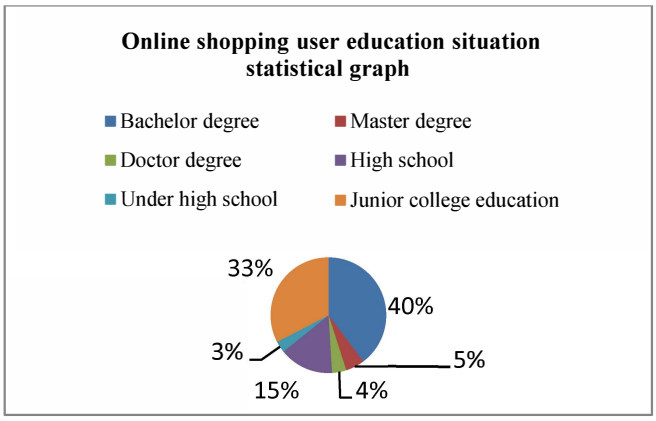

Fig 2 Online shopping user education situation statistical graph

According to related data from the movie website MovieLens in this paper, we can see that when people are

watching the film, the male users usually prefer the action movies, the female users usually prefer the emotional movies, children usually prefer the cartoon, the young people prefer the commercial blockbuster and the elderly usually prefer the family ethical movies and documentaries.

According to statistics above, we make sure that the attributes of users can classify users correctly and be persuasive. Considering the discrimination and easily acquired of the attributes, we draw a conclusion that there are three attributes (gender, age and occupation) playing a decisive role to choose the items. And we believe the users with the same characteristic have the similar consumption habits and hobbies. In [5], the author proposed a clustering approach by calculating the similarity of the user characteristic. However, considering the user's clustering center is not easy to determine, we put the three attributes as a user characteristics to calculate the user's comprehensive characteristic value in this paper. The specific calculation formula defined as: $muc(u)=\alpha S(u)+\beta A(u)+\gamma O(u)$, which $\alpha+\beta+\gamma=1$, α,β,γ present for the correlation coefficient of each attribute. muc(u) presents for the comprehensive characteristic value of user u.

S(u) presents for user u's characteristic value of gender,

$$S(u)=\begin{cases}0, & \text{user u is female}\\ 1, & \text{user u is male}\end{cases} \quad (1)$$

A(u) presents for user u's characteristic value of age,

$$A(u)=\begin{cases}0, & \text{the age of u less than 15}\\ \frac{Au-15}{40}, & \text{the age of user ubetween 15 and 55}\\ 1, & \text{the age of u more than 55}\end{cases} \quad (2)$$

O(u) presents for user u's characteristic value of occupation,

$$O(u)=\begin{cases}0, & \text{user u in leisure class}\\ 0.5, & \text{user u in culture class}\\ 1, & \text{user u in management class}\end{cases} \quad (3)$$

According to *The national occupational classification ceremony,* there are about 47 kinds of occupations separated into 8 categories. In the dataset used in the paper from MovieLens, it includes 21 kinds of occupations. We separate them into 3 large categories depending on their professionalism. They are leisure class, culture class and management class.

According to the comprehensive characteristic value calculated, we classify all the users into N classes. In the paper, we adopt 0.1 as the boundary and separate the users into 10 classes. Pseudo code as the following:

Initial the value of ten cluster center into 0.1, 0.2... 1.0
For j=1 To number of users
If user[j].gender=male Then S(j)=1
Else S(j)=0
If user[j].age≤18 Then A(j)=0
Else IF user[j].age≤55 Then A(j)=|user[j].age -15|/40
Else user[j].age≥55 Then A[j]=1
If user[j].occupation belongs to leisure class Then O(j)=0
Else IF user[j].occupation belongs to culture class Then O(j)=0.5
Else user[j].occupation belongs to management class Then O(j)=1
End If
muc(j)=αS(j)+βA(j)+γO(j)
For i=1 to 10
Compare muc(j) with center(i) to find the closest class to join in
Refresh the cluster centers
Next

According to formula above, we can separate every user into one cluster, and we can find the nearest neighbors of a user from his cluster.

III. SVD-BASED COLLABORATIVE FILTERING ALGORITHM

A. *Singular value decomposition (SVD)*

Singular Value Decomposition (SVD) is a kind of matrix decomposition technology. It deeply reveal the internal structure of the matrix [14][15]. SVD is widely used in the area of image compression, least square approach, etc[20]. It separates a m×n (assume m≥n) matrix R into three matrixes U, S, V [6]: $R = U \times S \times V^T$, U is an m×m orthogonal matrix ($UU^T=I$), V is a n×n orthogonal matrix ($VV^T=I$), S is a m×n diagonal matrix, which the elements not on the diagonal all equal to 0 and the elements on the diagonal satisfy$\sigma1\geq\sigma2\geq\cdots\geq\sigma n\geq0$, all the σn which is usually called singular value (SV)are bigger than 0 and ordered from large to small, SV can express the degree of closeness between a given matrix and another matrix which has a lower rank [19].

See Fig 3, R is a 4×6 two-dimensional matrix, and it is separated into three matrixes in Fig 4. They are 4×4 two-dimensional matrix U, 6×6 two-dimensional matrix V and 4×6 diagonal matrix S.

	1	2	3	4	5	6
1	0	1	3	0	5	2
2	2	0	2	1	5	3
3	0	0	3	2	0	1
4	4	5	0	2	0	0

Fig 3 4×6 two-dimensional matrix R

	1	2	3	4
1	0.6365	-0.2498	0.0900	-0.7241
2	0.6862	-0.1449	0.2143	0.6798
3	0.2208	-0.0182	-0.9719	0.0796
4	0.2742	0.9572	0.0374	-0.0845

	1	2	3	4	5	6
1	0.2680	0.5454	0.1793	0.5358	-0.4225	-0.3646
2	0.2179	0.6991	0.0859	-0.6014	0.0529	0.3032
3	0.4282	-0.1686	-0.6874	-0.3009	-0.4027	-0.2509
4	0.1820	0.2671	-0.5130	0.3514	0.7128	-0.0287
5	0.7180	-0.3041	0.4718	-0.1160	0.3180	-0.2342
6	0.3856	-0.1468	-0.0462	0.3519	-0.2175	0.8102

	1	2	3	4	5	6
1	9.2118	0	0	0	0	0
2	0	6.4889	0	0	0	0
3	0	0	3.2252	0	0	0
4	0	0	0	1.9068	0	0

Fig 4 They are 4×4 two-dimensional U, 6×6 two-dimensional V and 4×6 diagonal matrix S

B. *SVD-based Collaborative filtering algorithm*

Usually, an E-commerce website exists the problem of

sparse, because a user can only view, purchase and comment on very few items on it. In the most of collaborative filtering algorithms, similarity is the most important intermediate variable for predicting the ratings. Especially in the classic Pearson algorithm and Cosine algorithm, the items which are rated by different users are very important. There are two common approaches to solve the problem of sparse. One is to initial the unrated items with the default value 0, in fact people will not rate these item with the value 0, because it is too low to the item and the approach will lead to the reduction of the similarity between users. And the other one is to initial the unrated items with the average rating value of the user. The approaches invisibly increased the relationship between users, and increased the similarity between users. Above all, neither of the approaches are available. We propose an approach using SVD algorithm to solve the problem of sparse. It can improve the precision of the system by using the integrated rating matrix to calculate the similarity between users.

Algorithm:

Input: Original rating matrix R (R is a m×n two-dimensional matrix, m presents for the number of users, n presents for the number of items and the unrated items are filled with 0).

Output: The matrix PR which is full filled with the predicted ratings of the unrated items by every user in to the original rating matrix R.

(a) Calculate the average rating of every column r_i;

(b) Replace every rating $r_{ui,}$ which don't equal to 0, with r_{ui}-r_i, and get the new matrix R';

(c) Separate the matrix R' into three matrixes U, V, S with the SVD algorithm

(d) Simplify the matrix S, reset all the values which is less than 1 in the matrix S, as the singular values are sorted from large to small, delete all the columns and rows in which the values are 0 to form the new diagonal matrix S_K (It equals to keep the top K rows and columns);

(e) According to the diagonal matrix S_k, simplify matrix U and V according to the way to simplify matrix S into matrix U_K which is a m×k matrix and matrix V_k which is a k×n matrix;

(f) Calculate the square root of every value in the matrix S_k, create two new matrix A and B with the formula: $A=U_k\times S_k^{1/2}$ and $B= S_k^{1/2} \times V_k$;

(g) Fill up the matrix PR, predict all the unrated items R_{ui} in the matrix R with the formula: $PR_{ui}= \bar{R}_u+A_m\times B_n$ ($\bar{R}_u$ presents for the average rating of user u, A_m presents for the m*th* row in the matrix A, B_n presents for the n*th* column).

There are three classic collaborative filtering algorithms to calculate the similarity: Cosine similarity, Pearson correlation and Adjust cosine similarity. We use Pearson correlation to calculate the similarity in the paper, as

$$sim_{uv}=\frac{\sum_{i\in Iuv}(PRui-\bar{PRu})(PRvi-\bar{PRv})}{\sqrt{\sum_{i\in Iuv}(PRui-\bar{PRu})^2}\sqrt{\sum_{i\in Iuv}(PRvi-\bar{PRv})^2}}, \quad (1)$$

where Iuv presents for the items rated both by user u and user v, PRui and PRvi present for the rating rated by user u and user v in the matrix PR, $\bar{PRu}$ and $\bar{PRv}$ present for the average rating for user u and user v in the matrix PR. Because the matrix is formed through the decomposition and polymerization of SVD algorithm, the ratings of the items by users are substantially increased and it effectively solves the problems of inaccurate similarity caused by sparse.

According to the similarity calculated above, we find out k users who have the highest similarities with user u and we predict the ratings of unrated items in the original matrix R with the predicting formula:

$$r_{ui}= \bar{r}_u+\frac{\sum_{v\in N(u,k)\cap U(i)} simuv(Rvi-\bar{Rv})}{\sum_{v\in N(u,k)\cap U(i)}|simuv|}, \quad (2)$$

where N(u,k) presents for the set of users who have the highest similarities with user u, U(i) presents for the users who have rated item I, Rvi presents for the rating of I by user v, and $\bar{Rv}$ and $\bar{Ru}$ present the average rating of all the items rated by user v and user u.

IV. EXPERIMENTS AND ANALYSIS

The evaluation standard of recommendation usually divides into two categories: the measurement of forecasting accuracy, the measurement of classification accuracy. Mean Absolute Error (MAE) is one of the most acceptable standard to measure the statistical accuracy [7]. The mathematical expression is as follows: $MAE=\frac{\sum_{i=1}^{N}|pi-qi|}{N}$; which p_i presents the predicting rating of item i from the recommendation system, q_i presents for the real rating of item i rated by users and N presents for the number of the experiments. The smaller MAE is, the more precise a recommendation system is, and vise versa.

We adopt the 100k dataset from the movie website MovieLens, which includes about 100k ratings of 1682 movies from 943 users. In the dataset, there are three data sheet including user sheet, movie sheet and rating sheet.

After many experiment, it shows that it achieve the best classification by setting the weights α=0.2, β=0.5 and γ=0.3 in the clustering algorithm. We separate all the 943 users into 10 classes, show in Fig 5.

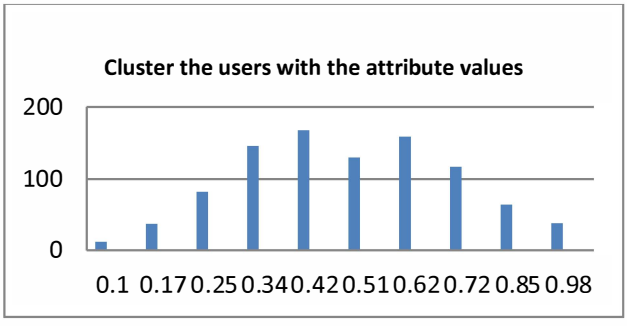

Fig 5 Separate the attribute values of users into 10 classes

We separate all the 100k ratings into 10 parts with the random function. Every part are 80%/20% splits of the data

into training and test data. By take advantage of the recordings in the training data sets to predict the ratings of the items in the test data sets, and calculate MAE values by using the real ratings and the predicted ratings, then calculate the average MAE value of ten groups of data, and finally get the precision of the algorithm.

According to KNN algorithm, considering of the different number of neighbors K, we respectively compare MAE values from the algorithm in the paper, the traditional collaborative filtering algorithm and ordinary SVD algorithm. The result is shown in Fig 6, as following.

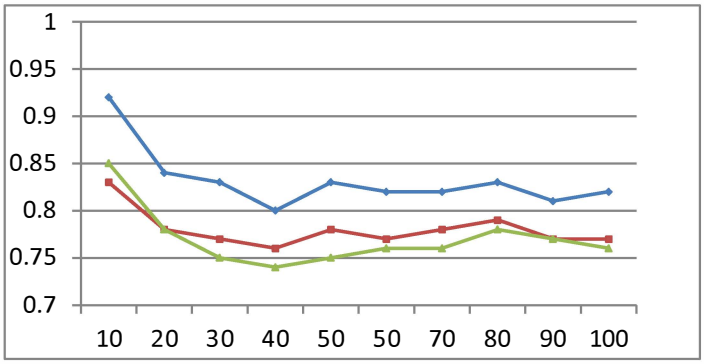

Fig 6 MAE comparison chart of three different recommendation algorithms

In Fig 6, we can notice that the algorithm proposed in the paper (the green one), which is improved from the traditional collaborative filtering algorithm and combines SVD algorithm and clustering algorithm together, is better than traditional collaborative filtering algorithm (the blue one) and ordinary SVD algorithm (the red one) as a whole. And from the choice of the number of neighbors, it reaches the best MAE value 0.74 when the number of neighbors is about 40. Totally, clustering collaborative filtering based on SVD algorithm we propose in the paper brings an improvement over the traditional algorithm. And on the basis of improving recommendation efficiency, the recommendation is satisfied.

V. CONCLUSION

With the development of the Internet technology, E-commerce has stepped into our life and turned to be an essential part of our daily life. As the core technology of E-commerce, recommendation system plays a crucial role, so the efficiency and the precision of a recommendation system decide the destiny of the recommendation system. In the paper, we improve the traditional collaborative filtering algorithm and ordinary SVD algorithm. Firstly, we cluster all the users by calculating the User characteristic value with the attributes of users, because the preference between users has a very close relationship with their gender, age and occupation. We can find the neighbors of a user through the classification using the characteristic value of the user, then reduce the dimension of the system and improve the efficiency of the system. Secondly, we add SVD algorithm which is famous in the domain of image processing on the basis of clustering algorithm proposed in the paper. The algorithm decomposes and then polymerizes commodity the rating matrix of items by the users. It effectively solves the problem of sparse of the ratings from users, then make it more accurate calculate the similarity, make the process of recommendation more suitable to the reality and make recommendation more accurate. In the age of information explosion, the algorithm proposed in the paper can lead to a better real-time and improve the efficiency of the system. At last, through preparing the MAE values in the experiments, clustering collaborative filtering recommendation system possess a better recommendation quality.

REFERENCES

[1] Songjie Gong. A flexible electronic commerce recommendation system, Physics Procedia 24(2012) 806-811

[2] J. Ben Schafer, Joseph Konstan, John Riedl. Recommender Systems in E-Commerce.

[3] Resnick P, Iacovou N, Sushak M.GroupLens: An open architecture for collaborative filtering 0f netnews[C], Proceedings of CSCW 1994, ACM SIG Computer Supported Cooperative Work, 1994.

[4] Min Gao, Zhongfu Wu, Feng Liang. Userrank for item-based collaborative filtering recommendation. Information Processing Letters, 111(2011)440-446.

[5] Golub G H, Van Loan C F. Matrix Computations (3rd edition) [M] . Johns Hopkins University Press, 1996 .

[6] Dan Kalman. A Singularly Valuable Decomposition: The SVD of a Matrix[J] . The College Mathematics Journal, 1996, 27(1) : 2- 23.

[7] erry M W, et al. Using LinearAlgebra for Intelligent Information Retrieval[J] . SIAM Review, 1995, 37(4) : 573- 595 .

[8] Tong Queue Lee, Young Park, Yong-Tae Park. A time-based approach to effective recommender systems using implicit feedback. Expert Systems with Applications, 34 (2008) 3055-3062.

[9] Long-Sheng Chen, Fei-Hao Hsu, Mu-Chen Chen, Yuan-Chia Hsu. Developing recommender systems with the consideration of product profitability for sellers. Information Sciences, 178 (2008) 1032–1048.

[10] You-Jin Park, Kun-Nyeong Chang. Individual and group behavior-based customer profile model for personalized product recommendation. Expert Systems with Applications, 36 (2009) 1932–1939.

[11] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Miguel A. Rueda-Morales. Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning, 51 (2010) 785–799.

[12] J. Ben Schafe , Joseph Konstan, John Riedl. Recommender Systems in E-Commerce. E-COMMERCE 99, Denver, Colorado, 1999 ACM 1-58113-176-3/99/0011.

[13] Resnick P, Iacovou N, Sushak M.GroupLens: An open architecture for collaborative filtering 0f netnews[C], Proceedings of CSCW 1994, ACM SIG Computer Supported Cooperative Work, 1994.

[14] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithm[C], Proc of the 10th International WorldWideWeb Conference, 2001:285—295.

[15] Mobasher B, Burke R, Sandvig J J. Model-based collaborative filtering as a defense against profile injection attacks[C], Proceedings 0f the 21st National Conference on Artificial Intelligence(AAAI'06), 2006.

[16] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J] . Journal of the American Society for Information Science, 1990, 41(6) : 391 - 407.