Summer 6-19-2015

Collaborative Filtering Recommendation Method Based on User Classification

Ting Zhu
School of Economics and Management, Xidian University, xi'an, 710071, china

Chunxiu Qin
School of Economics and Management, Xidian University, xi'an, 710071, china, cxqin@xidian.edu.cn

Collaborative Filtering Recommendation Method Based on User Classification

Ting Zhu[1], *Chunxiu Qin*[1*]
[1]School of Economics and Management, Xidian University, xi'an, 710071, china

Abstract: Collaborative filtering method is an important method of personalized recommendation, while the method often resulting in the problem of low efficiency with the increase in the number of users. To solve this problem, this paper presents a presents a personalized recommendation method with the adoption of user classification and collaborative filtering algorithm. Firstly, the huge users are classified into several groups according to a rule-based classification method. Then, on the premise of the accuracy of recommendation, the local neighbor users are discovered for users. Finally, based on the discovered local neighbors, personalized recommendation conducted. Experimental results show that with the adoption of a rule-based user classification, collaborative filtering algorithm has been significantly improved on the premise of the accuracy of recommendation.

Keywords: personalized recommendation; collaborative filtering; user classification; rule

1. INTRODUCTION

In the era of the web 2.0, network information resources are explosively growing. For example, the total number of sites has reached 1,272,704 in June 2014. Compared to the total number of sites in June 2005, it has increased by 595,204. Users need to spend a lot of time and effort to find the interesting resources in such a large amount of network resources. In order to solve this problem, personalized recommendation methods are increasingly conducted. The personalized recommendation helps people discover possibly interesting contents by analyzing the binary relationship between the user and the item (such as information, resources, goods, services), and generate personalized recommendations result, then provides different services according to different user [1]. It is different from personalized recommendation method based on resource content semantic analysis. From the user's perspective, collaborative filtering technology calculates the similarity between users and generates recommendation for target users according to the evaluation of neighbor users. Collaborative filtering can find users' interests of which users cannot be aware and has a high precision. However, with the number of users on the network increasing sharply, collaborative filtering technology needs to spend a lot of time to calculate the user similarity, which often results in the problem of low efficiency. Actually, in the last two years, the number of Internet user is increasing sharply. according to China Internet Network Development Statistics Report, the scale of china's internet users reached 632 million by the end of June 2014, an increase of 2 million people by the end of 2013.Thus, this paper attempts to adopt a rule-based user classification approach and collaborative filtering recommendation to speed the personalized recommendation process on the premise of the accuracy of recommendation.

2. RELATED WORK

The existing recommendation algorithms mainly include the content-based recommendation techniques, rules-based recommendation techniques and collaborative filtering techniques. (1) content-based recommendation techniques firstly describes users' interests by establishing their interests models, then extracts the content item feature to form the feature vector at the same time, at last, it makes recommendations based on

* Corresponding author. Email: cxqin@xidian.edu.cn(Chunxiu Qin)

the similarity of users' interests models and the item feature vector of the project[2]. Content-based recommendation techniques can handle the cold start problems of the project. While the system only can recommends users with the resources which is similar to the historical interests, and lack of mining the potential interests of users, and also can't handle the cold start problems of users. (2) Rules-based recommendation techniques use the rules to infer users some contents they are maybe interested in based on the contents users are interested in and they have visited. So the key of rule-based recommendation technique is making rules. Rules can be implemented by mining association rules[3]. Rules-based recommendation technique can improve the real-timing for it can establish models offline. But the quality of recommendation will reduce if the support and confidence of the rule are selected improperly, and the cold start problem of project cannot be handled. (3) Collaborative filtering technology makes personalized recommendation by using the interest similarity of different users. Collaborative filtering is one of the most widely strategy used in recommendation system. It includes user-based collaborative filtering [4][5] and item-based collaborative filtering [6]. User-based collaborative filtering can find neighbor users who are similar to target users by collecting user information, and then recommends target users according to the interests of neighbor users [7]. Item-based collaborative filtering selects the highest score project, calculates the similarity of items and then recommends users items which are similar to the high score items [8]. Collaborative filtering technique also has some problems. Such as the cold start problem [9] caused by new users and new items and the sparsely problem [10] caused by the increase of item which leads to the less of project score. And the most important step in collaborative filtering method is calculating the similarity between users, but there will be a huge amount of computation when finding neighbors in the whole user space with the increasingly large number of users. Therefore, in this paper we firstly classify users into some smaller groups in terms of user characteristics, such as gender, age, occupation. To accelerate the speed of the recommended, we only consider the local users in the same group, because of users within a group having greater similarity. The common classification algorithms include decision tree classification algorithm, Bayesian classification algorithm, association rules-based classification algorithm ad SVM algorithm [11]. Among them the association rules-based classification algorithm [12] combines the classifier constructor with classification of associated attributes, and finds relatively comprehensive and the classification accuracy is also higher, so we use rule-based classification algorithms to class great mount of users according to users' attributes.

3. COLLABORATIVE FILTERING RECOMMENDATION METHOD BASED ON USER CLASSIFICATION

The proposed method of collaborative filtering recommendation includes four steps: (1) rule-based user classification. It classified users with the rule-based classification method according to the user's property. (2) user-item scores. It represents the user and the user's score to the project. (3) Identification of neighbors users within groups. It computes similarity between users with the modified cosine, and then calculates the neighbor users. (4) Personalized recommendations. It predicts the target user scores for the project according to neighbor users scores for the Ungraded items, and then resulting in recommendation.

3.1 The proposed rule-based user classifications method

Firstly, the proposed method identifies some categories to form training set. Then, it discovers classification rule according to analysis of the training set. Finally, the method determine to category of the test set by the way of classification rules.

the method acquires user record, and expresses it as U1（A1，A2，A3，……，AK）, U represents a user inside, A represents the property. A series of user records can form the training set and the test set. Rule-based classification uses the "if…then…" to classify the records, and generate a rule set R, the rule set R=（r1∨r2∨…∨rk）, where ri is called classification rules.

Each classification rule can be expressed as follows:

ri：(A1 op v1)∧(A2 op v2)∧…∧(Ak op vk)）→yi, （Aj，vj）represents an attribute-value pairs, op represents Comparison Operators. If the attributes of user U1 can match with the rule antecedent, the rule is triggered and the user will be classified as category y1.

The process of the rule-based user classification can be expressed as follows（Figure.1）:

Figure 1. User classification process

Given the uneven distribution of class, we extract the rule used by decision tree and complete the user's classification.

3.2 User-item scores indication

We need show each type of user and user-item scores, this paper uses the Rm*n matrix to represent user-item scores, which consists of M users and N projects. The row vector represents the score which a user gives each item, the column vector represents scores information which different users give a certain project. As shown in Figure 2 represents the user-item matrix.

$$R_{m\times n}=\begin{pmatrix} R_{11} & R_{12} & R_{13} & \cdots & R_{1n-1} & R_{1n} \\ R_{21} & R_{22} & R_{23} & \cdots & R_{2n-1} & R_{2n} \\ R_{31} & R_{32} & R_{33} & \cdots & R_{3n-1} & R_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{m1} & R_{m2} & R_{m3} & \cdots & R_{mn-1} & R_{mn} \end{pmatrix}$$

Figure 2. User-item scores matrix

Rij represents the scores of user I to project j. scoring includes explicit and recessive score. Explicit score refer to users active score, the score is between 0-5, 0 indicates that the user did not score the project, the score of between 1-5 degree represents user's favorite degree, and different score represents different likeability. Recessive score means that the score which is gained by evaluating the time user spent on the page and the number of times user click on the page[13]. There are some difficulties on the acquisition of recessive scores, this paper only uses the explicit score to represent the score and does not consider the recessive score.

In addition, some sample of projects can be allowed to assess for users when a new user is generated. And then the new forming row vector can be added into the user-item scores matrix. So it can solve the cold start problem of users and projects to some extent.

3.3 Local neighbors user identification within groups

We can compute user similarity in each user-item scores matrix in order to obtain Local neighbors user after each groups' user-item scores matrix are obtained scores. Traditional similarity calculation has three ways: cosine similarity, modified cosine similarity, and related similarity [14], Here choose the modified cosine similarity method. Because different users have different evaluation scales in the similarity measure above, some users used to high scores, and some users used to low scores. to solve the problem, a same degree of favorite project is used in this paper. However, the modified cosine similarity improve this defect according to consider the mean score.

The similarity of user u1 and u2 is calculated with the following formula (1):

$$\text{Sin}(u_1, u_2) = \frac{\sum_{i=1}^{n}(R_{u_1,i} - \overline{R_{u_1}})(R_{u_2,i} - \overline{R_{u_2}})}{\sqrt{\sum_{i=1}^{n}(R_{u_1,i} - \overline{R_{u_1}})^2}\sqrt{\sum_{i=1}^{n}(R_{u_2,i} - \overline{R_{u_2}})^2}} \quad (1)$$

$R_{u_1,i}$ represents score of item I given by user u1 , $\overline{R_{u_1}}$ represents mean score of existing item given by user u1 in Equ.1.

It can discover local neighbor users for the target user by calculating similarity with the modified cosine similarity method in the same group. Using the k to represent the number of local neighbor user, the set of the neighbor users can be expressed as nei=（u_1,u_2,……，u_k）.3.4 Personalized recommendation

After producing the local neighbor users for the target user, it begin to generate recommendations for the target user.

The predicted score of item I by the target user u can be calculated using the following formula (2):

$$P_{u,i} = \overline{R_u} + \frac{\sum_{v \in nei} \sin(u,v) \times (R_{v,i} - \overline{R_v})}{\sum_{v \in nei} \sin(u,v)} \quad (2)$$

$p_{u,i}$ represents the prediction score of item I which produced by the target user u.

Firstly, ungraded items scores are calculated for each target user through the formula above. Then ungraded items are sorted according to the size of value. Finally, the first several projects are recommended to the target user.

4. EXPERIMENT AND EVALUATION

This experiment is conducted to simulate recommendation process and verify recommendation accuracy.

4.1 Experimental design

This experiment applies the data set on Movielens site, which consists of 100000 scores of 943 users about 1682 films, and each user evaluates at least 20 films.

In the experiment, three user's attributes (gender, age, occupation) are choosed to represent a user record which is expressed as U (id, age, gender, occupation). The specific values of the properties in Movielens site are shown in Table 1 below:

Table 1. User attributes form

age	teenagers、middle-aged and elderly
gender	F、M
occupation	administrator、artist、doctor、educator、engineer、entertainment、executive healthcare、homemaker、lawyer、librarian、marketing、none、other、programmer retired、salesman、scientist、student、technician、writer

In order to reduce the number of user categories, we compressed age into two categories including the younger group in which the user is less than 29 years old (including 29 years old) and the older group in which the user is more than 30 years old (including 30 years old). The values of property of gender include F and M. The property of occupation, include three categories which are service category(administrator、executive、librarian、marketing、doctor、educator、healthcare、lawyer、salesman), skill category(engineer、programmer、scientist、technician、artist、entertainment、writer、student), and other category(homemaker、none、other、

retired).

A training set is developed in this experiment by manual classification. The age is firstly considered, then gender and occupation in proper order. Classification process id shown in Fig.3:

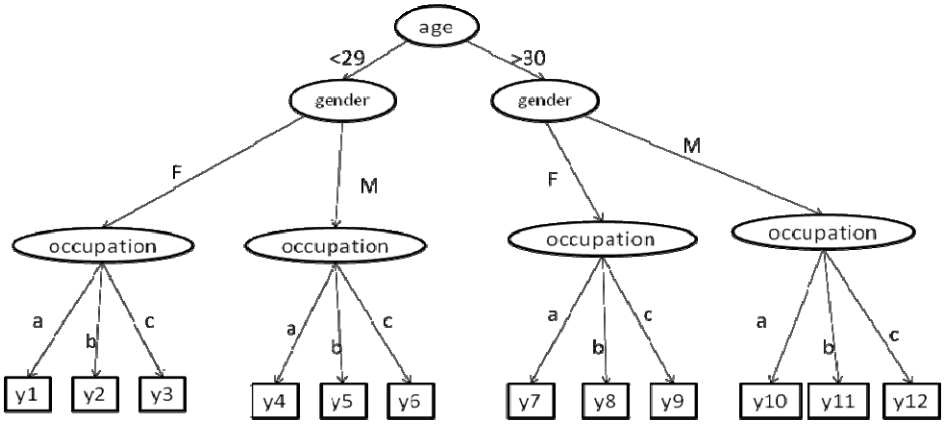

Figure 3. Decision tree of User classification

a represents service category of occupation, b skill category, and c other category. Eventually, users are classified into 12 subclasses of y1,y2,……,y12.

We use SPSS Clementine11.1 software to gain training set on the user's classification. We get explicit scores from u.data file of Movielens dataset on the user scores. The entire experimental data is divided into training and test sets. We choose 80% as training set and 20% as test set.

4.2 Experimental evaluation

(1) Rule-based classification evaluation

The quality of user classification can affect the accuracy of the final recommendation. Therefore, it is necessary to evaluate the accuracy of our proposed classification method. This paper uses accuracy, recall and F1 as evaluation standards of user classification. Confusion matrix of classification is showed in table 2.

Table 2. Confusion matrix of Classification

		predict class	
		1	0
actual class	1	TP	FN
	0	FP	TN

In table 2, "1" represents a positive class, "0" represents a negative class, and we usually refer rare class as the positive class and the majority of the class as negative class. TP (true positive) represents the number of positive class and is also predicted as a positive class. FN (false negative) represents the number of positive class and is predicted as a negative class. FP (false positive) represents the number of negative class and is predicted as a positive class. TN (true negative) represents the number of negative class and is also predicted as a negative class.

Therefore, Accuracy is expressed as $\mathrm{p} = \frac{TP}{TP + FP}$, Recall is expressed as $r = \frac{TP}{TP + FN}$. Actually, Accuracy and Recall interact with each other in practical situations, the accuracy goes highly while recall decreases and vice versa. We can combine accuracy and recall into F1 measure in order to achieve both. Thus F1 measure can ensure the relatively accuracy and recall. Among them $F_1 = \frac{2\mathrm{rp}}{r + p}$, the larger F1, the better the classification performance represents.

(2) Personalized recommendation results evaluation

As deviation of mean absolute error (MAE) is absorbed does not appear offset situation and can reflect the actual situation of the prediction error. Therefore, MAE is applied to assess the recommendation results.

MAE is presented by the following equation (3):

$$MAE=\frac{\sum\left|P_{u,i}-p_{u,i}\right|}{N} \quad (3)$$

$P_{u,i}$ represents predict score of item I given by user u , $p_{u,i}$ represents actual score of item I given by user u. N represents the total number of items. Thus, the smaller the value of MAE is, the higher the accuracy of recommendation will be.

(3) Algorithm efficiency evaluation

The algorithm efficiency is evaluated according to the complexity and time- consuming of algorithm.

4.3 Experiment result and discussion

(1) Evaluation of user classification

Table 3. User classification evaluation

category	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11	y12
accuracy	0.798	0.784	0.757	0.739	0.743	0.722	0.772	0.705	0.785	0.746	0.771	0.739
recall	0.833	0.809	0.785	0.785	0.772	0.751	0.797	0.737	0.818	0.782	0.801	0.773
F1	0.815	0.796	0.771	0.761	0.757	0.736	0.784	0.721	0.801	0.764	0.786	0.756

In table 3, it can been seen that the accuracy is the proportion of the actual positive class on predicted positive class, that is to say, the higher accuracy, the lower false positive class error rates will be. The recall is proportion of predicted positive class on positive class. The higher recall, the lower error rate of positive class being classified into negative class will be. While the F1 is harmonic mean of accuracy and recall. These above demonstrates that the rule-based user classifications method can work well according to users' attributes.

(2) Comparison of the algorithm efficiency

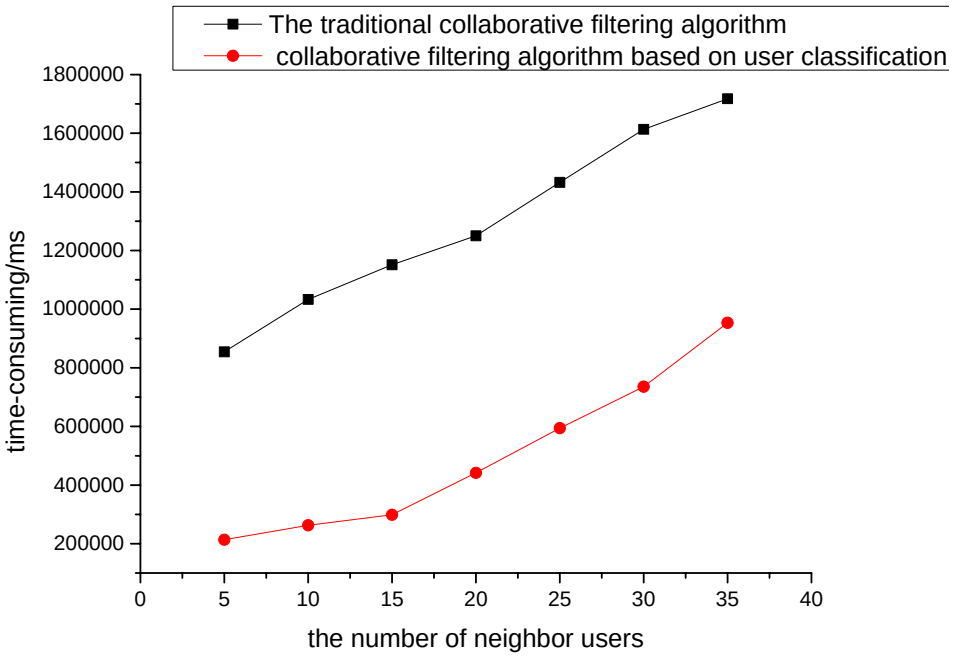

Figure 4. Comparison of paper algorithms and the traditional collaborative filtering algorithm

From figure 4, it can be seen that time-consuming of the user-classified algorithm is lower than that of unclassified algorithms in the situation of different number of neighbor users. In order to clearly and accurately compare the efficiency of two kinds of algorithms, we use time complexity to assess accurately.

The time complexity of unclassified algorithms is O(n^4*m^2) n, m is the number of users, and n is the number of items. 12 groups are running at the same time when the user-classified algorithm is running. Time-running of the algorithm is sum of time-classifying and the slowest time-running of 12 classified groups. The time complexity of decision tree Classification is O (m ^ 2.5), so the time complexity of classified algorithm is O(n^4*m^2)+O(m^2.5), namely, O(n^4*m^2). The number of users and items after classification is far less than that before classification. Thus, the efficiency of the user-classified algorithm is improved significantly.

(3) The effect of the number of neighbors k on recommendation accuracy.

The aim of this experiment is to observe the effect of the number of neighbors on recommendation accuracy. The number of neighbors initially is five, and increases in interval of 5 until it is 35. The compared total is 7 times. See Fig.5 for experimental results.

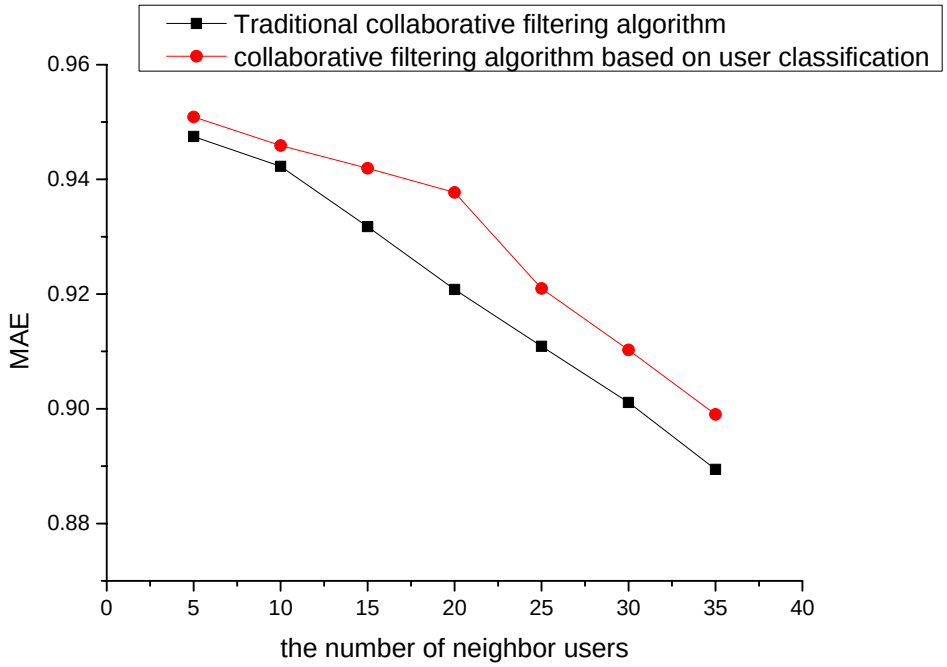

Figure 5. Contrast the paper algorithm and the traditional collaborative filtering algorithm

It can be seen in Figure 5:

1) The trends of value of MAE about the user-classified algorithm and the traditional algorithm both are decreased with the increase in the number of neighbor users, namely, the accuracy of recommendation is increased.

2) In the case of same number of neighbor users, the value of MAE about the user-classified algorithm is a little higher than the traditional algorithm, but the difference between them is small and the error rate is $w = \frac{\sum m_1 - m_2}{n}$, m_1 is the MAE value of this proposed algorithm, and m_2 is the MAE value of the traditional algorithm, n is the values number of different neighbor users. Where n equals 7, so W=0.0142. The error rate is not large. Therefore, the presented method can significantly save time-consuming in the premise of ensuring the relatively higher accuracy.

5. CONCIUSION

The collaborative filtering algorithm is widely used in the personalized recommendation to solve the problem of the modern Internet information overload. Due to the sharp increasing of users on Internet,

collaborative filtering algorithm will encounter a large amount of calculation which leads a low efficiency. To solve this problem, this paper suggests a rule-based classification method to classify great mount of user into smaller groups to reduce the time-consuming on the premise of ensuring relative accuracy The experiments show the proposed method is feasible and effective. While, the method also has some problems to be improved. For example, the efficiency of the algorithm is improved after adding a user classification, but the recommendation accuracy is reduced a little bit. In addition, the sparsely problem also exists that the recommendation will be inaccurate when users' score is less, as well as the scalability problem [15], which means that the user's interests change over time and affect the recommended results. There are the next step to study.

ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China (Project No. 71103138)

REFERENCES

[1]YANG Li-na,YAN Zhi-jun.MENG Zhaokuan.Dynamic Establishing Virtual Learning Community Based on Personalized Recommendation Strategy[J]. Modern Educational Technology,2012,01:88-92.

[2]An Yue,Li Bing.Yang Ruitai.Content-based Personalized Recommendation on Popular Micro- topic[J].journal of intelligence, 2014,02:155-160.

[3]Suo Qi,Lu Tao.Research on Recommender System Based on Association Rules[J].natural sciences journal of harbin normal university,2005,02:50-53.

[4]Y Chuan,X Jie-ping.Recommendation algorithm combining theuser-based classified regression and the item-based filtering. Processing of the International Conference on ElectronicCommerce,Proceedings-the new E-commerce:Innovations forConquering Current Barriers,Obstacles and Limitations toConducting Successful Business on the Internet . 2006,574-578.

[5]Breese J,Hecherman D,Kadie C.Empirical analysis ofpredictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in ArtificialIntelligence(UAI'98) . 1998,43-52.

[6]Sarwar B,Karypis G,Konstan J,et al.Item-based CollaborativeFiltering Recommendation Algorithms. Proc of the 10thInternational Conference on World Wide Web . 2001,285-295.

[7]FAN Bo CHENG Jiu-jun.Collaborative Filtering Recommendation Algorithm Based on User's Multi-similarity.Compter Science,2012,01:23-26.

[8]Wang Yu-bin,Meng Xiang,wu;Hu Xun.Information Aging-based Collaborative Filtering Recommendation Algorithm[J].Journal of Electronics&Information Technology,2013,10:2391-2396.

[9]Massa, P,and Avesani, P.Trust-aware Collaborative Filtering for Recommender Systems. Proceedings of International Conference on Cooperative Information Systems . 2004.

[10] Sarwar B,Karypis G,Konstan J,et al.Application of Dimensionality Reduction in Recommender System-A case study. Proceedings of the WebKDD 2000 Web Mining for E-Commerce Workshop at ACM SIGKDD . 2000.

[11]LI Ling-Li.A Review on Classification Algorithms in Data Mining[J]. Journal of Chongqing Normal University(Natural Science) ,2011,04:44-47.

[12]R.Agrawl,R.Srikant.Fast algorithms for mining association rules. proc. Of 20th Int'l Conf on Very Large Databases (VLDB'94) . 1994.

[13]Yang Li-la,Liu Ke-cheng,Yan Zhi-jun.Virtual research community knowledge sharing knowledge and personalized recommendation[J]. Educational Technology of china,2010,06:108-112.

[14]MA Hong-wei,ZHANG Guang-wei,LI Peng.Survey of Collaborative Filtering Algorithms[J].Jo urnal of Chinese Computer Sy stems,2009,07:1282-1288.

[15]Xia Pei-yong.Research on Collaborative Filtering Algorithm of Personalized Recommendation Technology[D]. University of China Ocean,2011.