Service Recommendation Method Based on Collaborative Filtering and Random Forest

Lijing Xing
School of Information Science and Engineering
University of Jinan
Jinan, China
1239527150@qq.com

Bingxian Ma *
School of Information Science and Engineering
University of Jinan
Jinan, China
*Correspond Author

Delong Ma
School of Mechanical and Electrical Engineering
Soochow University
Suzhou,China

Abstract— With the development and popularization of E-commerce, more and more information services have appeared on the web. In order to meet users requirements more accurately, several service recommendation systems had been set up. Many methods have been proposed to discover users' interests for service recommendation, such as collaborative filtering and content based service recommendation. In this paper, a new service recommendation method is proposed based on user's interest, which combines collaborative filtering based on multiply users and random forest based on single user, and this fusion method uses cross validation model. This method can improve cold start and pick up speed. Experiment results show that the method can discover users' interest efficiently and is more accurate. This method can combine two basic methods so that the result is more accurate.

Keywords- Services Recommended; Collaborative Filtering; Cross Validation Model; Random Forest Model; Multiply Users

I. INTRODUCTION

Today, with the rapid development of e-commerce, there is a large number of information services on the web, for better user experience, research on the relationship among services and user's interest has been attracting great interests of industry and researchers recently. The problem mainly includes two aspects: 1) for a user, how to get his interest? 2) Which is the most suitable service that user's interested in? Services recommendation [1] is an accepted effectively approach to deal with this problem.

Many services recommendation methods had been proposed, such as content based service recommendation [1], collaborative filtering based service recommendation [2], knowledge based service recommendation [3, 4], services recommendation based on services effect [5] and service recommended based on association rules[2]. But there are still some questions need to be solved within the recommend process, such as accuracy, cold data start and sparse data. To solve those questions, a fusion method is put forward which combines random forest and collaborative filtering to discover user's interest.

The rest of this paper is organized as following: Section 2 briefly introduces service recommendation algorithm; in Section3, collaborative filtering based on multi-users is studied; random forest algorithm for service recommendation is proposed in Section4. Section 5 combines collaborative filtering and random forest method. for better services recommendation; in Section 6, the experimental results are showed and analyzed; at last, Section 7 concludes this paper..

II. RECOMMENDED METHOD

Many works had been done on content based recommended methods, which used clustering method to calculate and recommend based on the content of user's historical data.

Service recommendation based on collaborative filtering [2] is the most widely used and mature methods, the realization of collaborative filtering generally included three steps: firstly, the behavior information is used to calculate similarity among users or projects, and then the data model is set up; secondly, used high similarity of users or projects to predict user's preferences about projects; thirdly, the most likely interested project is recommended to user. But collaborative filtering had data sparseness and cold start problem, data sparseness refereed to the number of project data is too small to recommend, cold start problem is that there did not have any historical data of new users, it is very difficult to recommend.

Service recommendation method based on knowledge [4] is mainly depended on user information; utility-based recommendation would create a utility function to compute the value of services or goods for a user.

Recommendation based on association rules is an effective method to solve cold start problem, where association rules are depended on whether there existed links among several different products, such as the link between beer and diapers, now this method had been widely used in business.

Combined recommendation would select and combine above recommendation methods, such as collaborative filtering and clustering based recommendation, which could solve cold start problem in some degree.

Service recommendation has various evaluation standards. User' satisfaction refers to the degree of users satisfied with the result of service recommendation; it is the most important indicator of recommendation system. Prediction accuracy[6]mainly describe the system's ability to predict user behavior, related algorithms use offline data sets as input to output the list of recommended and coincidence rate to calculate user behavior, the greater

coincidence rate, the higher accuracy. Coverage [6] describes recommendation system ability to discover items with long tail.

III. Collaborative Filtering based on Multi-user

Collaborative filtering calculates similarity among multi-user; it is mainly depended on data records of multi-user. Historical data are used and calculated their similarity to get service recommendation results. The premise of collaborative filtering is to have multiple user historical data. Collaborative filter has many ways to discover user interest, which includes combining user-based and item-based collaborative. Here, combined user-based algorithms are used to discover user interest. Firstly, the similarity of users is calculated; secondly item similarity sorting is set up and average value can be obtained by adding parameters; at last the K nearest services will be recommended.

The main method to calculate similarity of user interest is cosine similarity, modified cosine similarity method and so on.

Cosine similarity [5, 6]

Project on m dimensions user space vectors can be as a rating of the project. The project can use the cosine of the angle. The smaller the angle, the higher the similarity. Setting users i and j in n project grading, which represents as a vector space? The similarity between users i and j is:

$$sim(\mathrm{i},\mathrm{j}) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}.\vec{j}}{\|\vec{i}\|_2 . \|\vec{j}\|_2} \tag{1}$$

Modified cosine similarity [5, 6]

Modified cosine similarity measurement method is used to reduce the average score so that the disadvantage of cosine similarity measure is changed. Setting users i and j are common grade project with l, and after users i and j score collect project, the similarity between user i and j is as follows:

$$sim = \frac{\sum_{c\in I_{ij}} (\mathrm{R}_{i,c} - \overline{R_i})(\mathrm{R}_{j,c} - \overline{R_j})}{\sqrt{\sum_{c\in I_i} (\mathrm{R}_{i,c} - \overline{R_i})^2}} \frac{1}{\sqrt{\sum_{c\in Ij} (\mathrm{R}_{j,c} - \overline{R_j})^2}} \tag{2}$$

Relative similarity [5, 6]

Setting U for users is i and j collection, similarity of the user i and j is as follows:

$$sim(\mathrm{i},\mathrm{j}) = \frac{\sum_{u\in U} (\mathrm{R}_{ui} - \overline{R_i})(\mathrm{R}_{uj} - \overline{R_j})}{\sqrt{\sum_{u\in U} (\mathrm{R}_{ui} - \overline{R_i})^2}} \frac{1}{\sqrt{\sum_{u\in U} (\mathrm{R}_{uj} - \overline{R_j})^2}} \tag{3}$$

$$\overline{R_i} = \frac{1}{U} \sum_{u\in U} R_{ui} \tag{4}$$

$$\overline{R_j} = \frac{1}{U} \sum_{u\in U} R_{uj} \tag{5}$$

To generate dynamic recommendation and more accurate recommendation, define time as timestamp for each user and resource, Time=D/T, here T is time length from the beginning to the end that user viewed a resource; D is the time length that user viewed the resource recently time which is calculated. Similarity of user i and j can be derived:

$$sim(\mathrm{i},\mathrm{j}) = \sum_{m=1}^{n} Timesim(i,j) \tag{6}$$

Here value of timestamp Time is the smaller timestamp of user i and j. The main procedure of collaborative filtering is as following:

Step1: inputting the recommend service number K;

Step2: user' data is divided into test set and train set;

Step3: the similarity of user-user is calculated and user similarity matrix is established;

Step4: sorting the similarity matrix and getting new similarity;

Step5: calculating user interested items through K nearest algorithm and recommend them.

When calculating the similarity, the history must be known, but new user do not have related data information. This paper puts forward a new ways to the problem, make full use of the information including age or job and so on. At last, it uses the information and the history to calculate the similarity. Through the similarity of attribute and historical data having different weights, it can be found that the most similar serve which is recommended. The result is accurate and improves the cold start problem.

IV. Improved RANDOM Forest Model for Single User

Use random forest model to recommend service for single user, it can classify user data and predicts whether user uses the service or not based on user's historical data. If user uses a service, the classification label is yes, otherwise is no. Every user has a number of historical data and use random forest model categorizing user's historical data. Then the label is predicted, which is yes or not. If the label is yes, the user may use the service.

Random forests are based on decision trees. Decision tree begins with the root, and then the root has two trees until producing leaf node. There are many algorithms for producing decision trees, for example, C4.5 [11, 14] and so on. Different algorithms have different ways to decide the split of trees. Decision trees have many disadvantages, for example, excessive fitting and complex classification rules. To overcome shortcomings, random forests combined with the multiple classifiers. Random forests are a kind of statistical learning theory, it is the using of heavy sampling method which extracts multiple samples from original sample, which includes many decision tree algorithm, it is a kind of nonlinear modeling tools, and it has good processing and relatively simple to implement, but the excessive noise classification and registered problems will have a fit. In this paper, CART [11, 14] and when starting to recommend service, data sparseness is the biggest question; random forest can deal with the question. Data preprocessing is the key technology of data mining, and now there are many methods of data preprocessing, treatment gap value, dirty data processing, recognition or deleting isolated point. Data preprocessing is the purpose of the data in the original file. Now use under-sampling to

solve data sparseness. Adding the weight of few classes can improve classification of accuracy. This paper uses the bagging builds random forest. Bagging is based on data fusion. The method is very stable. Bremen [14] points out that K-nearest is very stable, so bagging is equal to k-nearest. Bagging can solve unstable and increase precision.

User' data includes his history shopping records; there are many attributes to classify, such as score quality and price. Through voting decides the product label belongs to yes or not, if voting result is yes more than no, the label is yes, it means user maybe buy the product. Then recommend the product to users. The algorithm of discover users' interests is as follows:

Step1: Data sampling. Getting a user data randomly and classification's points are selected randomly.

Step2: Every record has the label .Using random forest algorithm classifies the data so that the label value is obtained. If the label is yes, it indicates user maybe use the serve.

Step3: According to the label, recommend service whose label is yes to users.

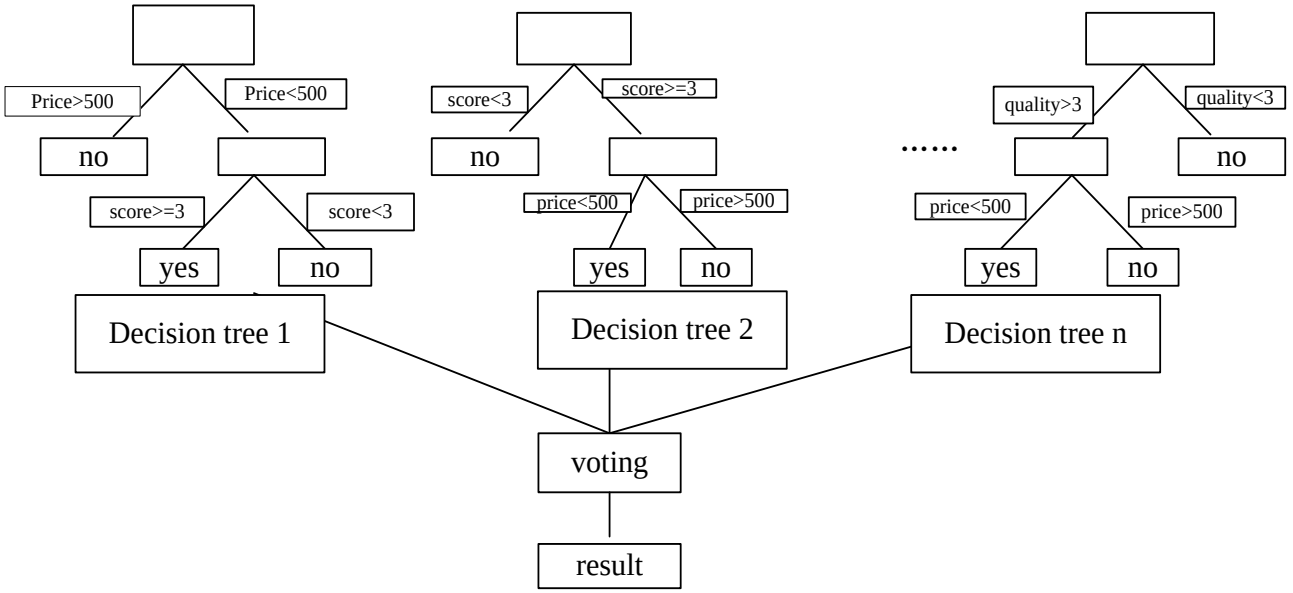

Figure 1. Application of Random Forest

V. MODEL FUSION

Put forward service recommendation method that combine collaborative filtering and random forest method. Each time randomly divide training data into five average parts and partly are used for training while the rest are used for testing. Finally collaborative filtering method is used to produce final recommendation results. The model is mainly to reduce the risk of over-fitting. Assuming there are K predictors to predict recommendation results, at last results of K predictors are fused to give a more accurate recommendation.

Use "leaving one out"[15] method to restructure training data and use it to train each predictor, and then use removed data to test each predictor, the result is set as training data for fusion model.

N-fold cross-validation method is used on the first layer to produce training data, and the data is used to construct user based recommendation and model based recommendation.

And then test trained recommendation on the corresponding cross-validation test set and test result is set as training data for the second layer of fusion model.

Cross-validation model is a very effective assessment prediction model to predict the effectiveness of the method, and is usually used five or ten times within a procedure. However, the sample size for less data with a five-time or ten-time cross-validation will be a problem that is not getting enough sample test set to split the training set and test set, the prediction will have large deviation. And then, when the sample size is small, general cross-validation is not suitable. Use Leave one out (LOO) [15] algorithm to solve this problem.

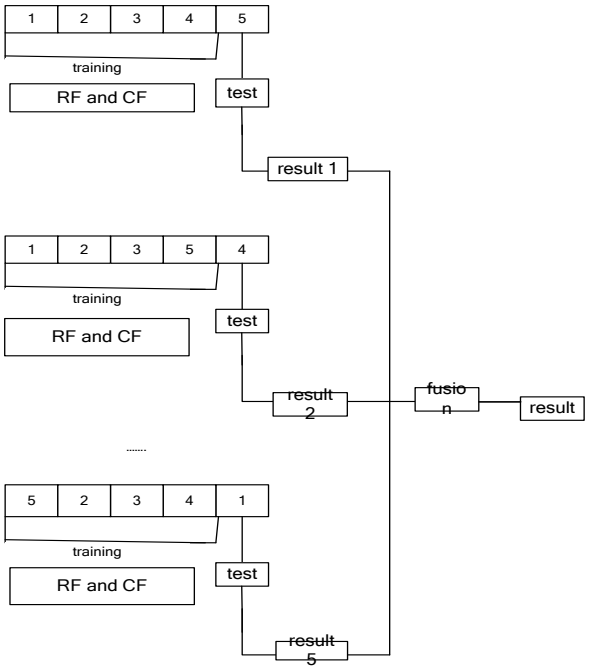

Figure 2. Method Fusion

In a word, use collaborative filtering and random forest from different views to recommend serve. The procedure of the method has following steps:

Step1: splitting testing set and training set m (m=5) parts, in which 1 / m for the testing set, and take k < = m – 1 for training set;

Step2: calculating user similarity, constructing user similarity matrix;

Step3: user similarity matrix sorting (from high to low) according to users' similarity;

Step4: using k nearest neighbor algorithm to get user interested items;

Step5: first N items are recommended based on interested degree of item;

Step6: outputting recommended service.

This is collaborative filtering recommending. Random forest is used to classify the data of user .According to the label, it decides user using serves or not. The fusion of the two methods recommended final results, according to the similarity. The combination of the two methods is effective and stable.

Experiment Analysis

The standard of evaluation is mainly statistics, use recall rate and accuracy as the evaluation standard. R (u) [6] is based on user behaviors within training set to make recommendation; T (u) [6] is behavior of users on the test set list. The recall rate is

$$\mathrm{Re}\,call = \frac{\sum_{u \in U} | R(\mathrm{u}) \cap \mathrm{T}(\mathrm{u}) |}{\sum_{u \in U} | T(\mathrm{u}) |} \tag{7}$$

The accuracy is:

$$\mathrm{Pr}\,ecision = \frac{\sum_{u \in U} | R(\mathrm{u}) \cap \mathrm{T}(\mathrm{u}) |}{\sum_{u \in U} | R(\mathrm{u}) |} \tag{8}$$

Select data from Movie Lens [12] which includes information of users and movies. There are 943 users, 1682 items and 100000 ratings about movies .When recommending the same service; compare their recall rate and accuracy of different methods.

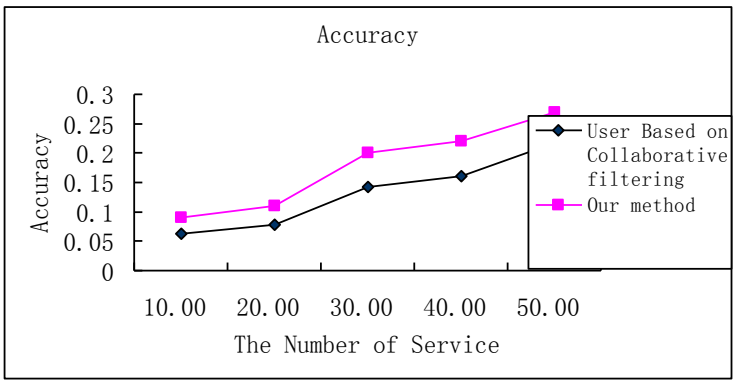

Figure 3. Comparison of Accuracy

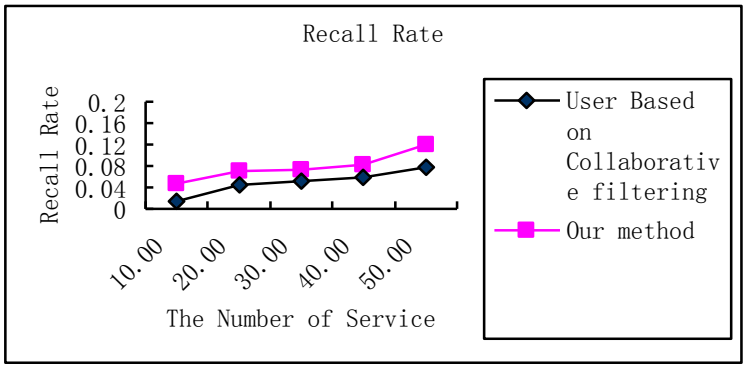

Figure 4. Comparison of Recall Rate

From the experimental result, it can be seen that accuracy and recall rate of the method has an obviously better result than collaborative filtering alone, especially when there is large number of services, the method is more accurate. And the new method can improve data sparseness question and cold start.

VI. CONCLUSION

This paper puts forward a services recommendation method that combine single user and multi users, using collaborative filtered for multi users and random forest for single user, recommended services could be get through fusion method and using cross validation model. Experimental result shows that the method has an obviously better result according to accuracy and recall rate. The combination of the two methods is effective and stable. There are many classic algorithms in machine learning which is used in recommendation, so the next step will apply machine learning widely into service recommendation.

REFERENCES

[1] ArwarB, KarypisG, KonstanJ, etal. Analysis of recommendation algorithms for E-commerce[C]. 2nd ACM Conference on Electronic Commerce, 2000,158-167.

[2] Jianguo Liu,Tao Zhou,Qiang Guo, Binghong Wang. Advances in personalized recommendation system [J]. Progress in Natural Science. 2009 (01) .

[3] WangWei-ping Liu-Ying .Recommendation algorithm based on customer behavior locus[J] .Computer Systems &Applications ,2006,15(9):35-38.

[4] Jun-zhong gu. Context-aware computing [J]. Journal of East China Normal University (natural science edition). 2010, 9 (5): 1-20.

[5] R. Salakhutdinov, A. Mnih and G. Hinton. Restricted Boltzmann Machines for collaborative filtering. Proc. 24th Annual International Conference on Machine Learning, pp.791–798, 2007.

[6] Yuxiao Zhu ,Linyuan Lv. Evaluation Summary of recommendation system [J].University of Electronic Science. 2012,(02):225-226.

[7] Ansari A, Essegaier S , Kohli R. Internet recommendation systems[J].Journal of Marketing research,2009,37(3):363-375.

[8] BREESE J S,HECKERMAN D,KADIE C. Empirical analysis of predictive algorithms for collaborative filtering; proceedings of the Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, F,2008 [C].

[9] PENNOCK D M,HORVITZ E,LAWRENCE S, et al. Collaborative filtering by personality diagnosis: A hybrid memory-and and model-based approach; proceedings of the Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence,F,2000[C]. Morgan Kaufman Publishers Inc.

[10] Perdiguero-Alonso D, Montero F E, A Kostadinova, Raga J A , Barrett J. Random Forests.a Novel Approach for Discrimination of Fish Populations Using Parasites as Biological Tags[J]. International Journal for Parasitology, 2008, 38 (12).

[11] Shani G, Gunawardana A. Evaluating recommendation systems. Proceedings of Recommender systems hand book. Springer, 2011: 257—297.

[12] Auret L, Aldrich C. Change Point Detection in Time Series Data with Random Forests[J]. Control Engineering Practice , 2010, 18(8).

[13] Breimaii L. Stacked regressions. Machine learning, 1996, 24(1):49-64.

[14] Yasushi U,Hiroyuki M.Credit Risk Evaluation of Power Market Players with Random Forest[J].Transactions on Power and Energy,2008,128(1).

[15] Wang,Y,Wang,R,Jia,H,Li,J.Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms[J]. Neural Computation 2014.

