International Journal of u- and e- Service, Science and Technology
Vol.9, No. 3 (2016), pp.27-34
http://dx.doi.org/10.14257/ijunesst.2016.9.3.04

Constructing the Core User Set for Collaborative Recommendation Based on Samples Selection Idea

Zhang Li, Yu Lei and Cao Shuyan

School of Information Technology & Management Engineering,University of International Business and Economics, Beijing,100029,China
tasummer@sina.com

Abstract

As an effective recommendation technology to solve "information overload" problem, collaborative filtering has widly attracted attention of scholars from various fields. First, this paper proposes a method for measuring the recommending ability of user based on popularity and long-tailed distribution. Then, a global core user set is constructed for recommadition based on the recommending ability and samples selection idea in data mining, aimming to take advantage of users in the different part of the long tail distribution and reduce computing complexity of the algorithm without lowing the recommended performance. Experimental results show that the algorithm is effective and can be used to solve the real-time problem and cold start recommendation.

Keywords: *collaborative filtering; core users; long-tailed distribution; samples selection*

1. Introduction

With the development of internet technology and information explosion, recommendation system as an effective tool to solve "information overload" is applied to various network platforms. Collaborative filtering(CF) is widely used in recommendation systems. The main rationale behind CF is that similar users have some interests in common. And user-based CF algorithm needs to search for the most similar users on the whole user set, which challenged by the problems of cold-start and computing complex [1]. So researchers have constantly improved CF algorithms. Some improved CF algorithms are proposed, such as matrix reduction-based CF, clustering-based CF, model-based CF algorithms and so on.Also some studies integrate social network properties into CF algorithms[2]. Howere, few studies have concerned about the impact of different users subset on the recommended properties. In this study, we mainly focus on different subset of users on the impact of collaborative recommendation performance, trying to construct a core user subset without reducing the recommending performance, while reducing the computational complexity of the algorithm.

There are two contributions of this paper,one is to study users' role in CF recommendation on different locations of the long-tailed distribution and propose a method to measure users' recommendation ability; the second is to construct the core user subset for recommending based on the users' recommendation ability and the idea of samples selection.

2. Literature Review

CF algorithms can be divided into two categories:memory-based and model-based on algorithms [2]. There are two memory-based CF methods,one is user-based,the other is item-based. This paper uses user-based CF algorithm. So we mainly review the research on it. Finding the similar users to collaborate the target users is one of the main process

ISSN: 2005-4246 IJUNESST

in user-based CF algorithm. Scholars have carried out a large number of in this process, which can be divided into two branches:

One branch is to improve user similarity measured methods. Some researches adjust *Pearson* coefficient, *Cosine* coefficient with information of time, items or users' interest in user-based CF algorithms and achieved good results[3-4].The other researches use the social network information to improve the calculation of similarity between users in the CF algorithms,such as information on friendship, tag data of social network and trustworthiness between users[5]. Accroding to the user-item diagram and user behavior data, Konstas I (2009) accessed to the friendship between users to improve the similarity, which have achieved good results and provided a new ides for the development of collaborative recommendation technology based on social networks[6].R. Zheng *et. al.* (2007) calculated the distance between two users in a social network graph, and find that similar users are closer[7]. However, with the users increased, the computational complexity is also increasing. The other branch concerns users' recommended ability in the collaborative filtering algorithm,which measured by different methods. Zhou *et. al.* (2007) introduced the diffusion theory and heat conduction into personalized recommendation and proposed a recommendation algorithm based on network resources allocated. In CF algorithm, user resources can be viewed as his recommendation capability used as weights to calculate similarity between users[8]. Zhang L. and *et. al.* (2013) measured user recommendation ability node degree in social network [9]. The results of those researches show that users have different ability.

The above studies have tried to design a new collaborative recommendation method and studied users effects on the performance of the recommendation algorithm in the user's individual level. At the same time, these studies are generally comput the similar between users or recommendation ability of user based on the whole use set. With the increasing of users, the computational complexity of the algorithms increase greatly.On the other hand, the existing research use the K most similar users of the target user to recommend in the user-based CF, and in user-item binary graph, if the degree of user node is defned as the number of evaluated items, the distribution of degree is the power law (the long-tail distribution). But the existing measuring methods of similarity (such as: pearson coefficient, cosine similarity) shows us that the degree of the user is bigger, more similar users he/she has. That it is to say the user is easier selected to recommend for the target user, *i.e.,* larger degree of users (in the head of long-tail distribution) is often used to cooperate with the target users. On the contrary,the less degree of user(in the tail of long-tail distribution) is, the less similar users,which means they are rarely used to recommend. While the above studies ignore the role of users in other parts of long-tail distribution,especially in the middle of distribution.

Inspired by the idea of the global nearest neighbor in literature [10], based on the analysis of the different part user in long-tail distribution, this study attempts to construct a global core subset of users, containing only 30-40% users, but unchanging the performance of the collaborative recommendation algorithm largely. It provides a new method to solve the real-time and cold start problem in online recommendation system.

3. Frequency of the Rated Items

The degree of user and frequency of the rated items are often used to measure the recommended ability of users [9-12]. The research result in literature [9] shows that user recommended ability is positively associated with the degree of user, while literature [11] shows that it is negatively associated with the frequency of the rated items. In this study,we use the frequency of the rated items to measure the recommended ability of users, computed by equation (1).

$$AvgFreq(u_i) = \frac{\sum_{a_j \in I_{u_i}} Freq(a_j)}{\left\| I_{u_i} \right\|} \quad (1)$$

Where $Freq(a_j)$ is the number of users who have rated item a_j, $\| I_{u_i} \|$ is the number of items rated by user u_i, *i.e.*, the degree of user u_i. This factor shows whether a user is interested in popular and high frequently rated items or non-popular and low frequently rated items. The equation shows us the degree of user is negatively associated with the frequency of the rated items and the Figure 1 also verifies this. Figure 1 shows the relationship between the degree and the frequency of the rated items of the MovieLens dataset provided by GroupLens Laboratary. From Figure 1, we can know they are negative correlation and long-tail distribution.

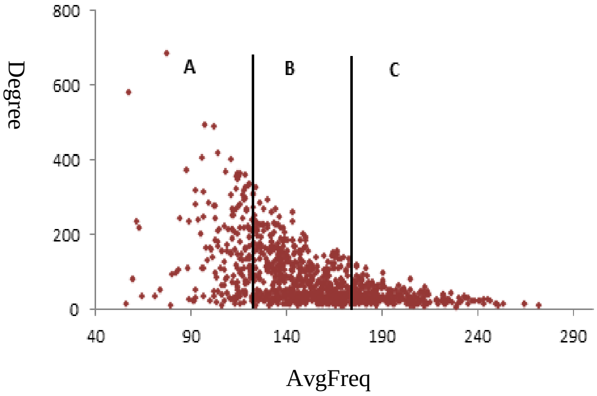

Figure 1. Distribution Scheme of all Users according to their Degree and *AvgFreq* Values

4. The Method of Constructing the Core User Subset

According to the existing research we can know that the uses in the head of long-tail distribution are more likely to be used to recommend,while information of users in the tail and middle part of the long-tail distribution has not been fully utilized. So this paper proposes a method to construct a core user subset based on samples selected idea, which purpose is to use the users in the middle part of the long-tail distribution to recommend.

According to frequency of the rated items, all users are divided into three equal subsets identified as A, B, C individually. Based on subset B, the core user subset $Quser$ is built as follows:

Step 1: select the boundary samples in subset B with the threshold R. When the distance between user $u_i(u_i \in B)$ and the center F of subset B is greater than R ,select u_i into the set Quser. At same time, for the users that distance less than R, randomly select into Quser based on the ratio 1/p. The distance is calculated by equation (2),where F is the center of subset B ,which equal to the average of all $AvgFreq(u_i)(u_i \in B)$.

$$dis\tan ce(u_i \in B) = | AvgFreq(u_i) - F | \quad (2)$$

Step 2: calculate the average similarity between user $u_i(u_i \in Quser)$ and all users in subset C. If the similarity is greater than the set threshold d_1 ,remove it from $Quser$. The similarity between users measured by cosine coefficient as shown in formula (3).

$$sim(u_i, u_j) = \frac{\| I_{u_i} \cap I_{u_j} \|}{\sqrt{\| I_{u_i} \| . \| I_{u_j} \|}} \tag{3}$$

Where $\| I_{u_i} \|$ and $\| I_{u_j} \|$ are the number of items rated by user u_i and u_j.

Step 3: calculate the average similarity between user $u_p (u_p \in A)$ and all users in subset $Quser$. If the average similarity is greater than the threshold d_2 ,select it into $Quser$. Through the above three steps,w*e g*et the final subset $Quser$ for collaborative recommendation.

5. Experiments

We used the MovieLens dataset in experiments, which is a popular dataset used by researchers and developers in the field of recommendation. The dataset contains ratings from 943 users on 1,682 movies. Figure 1 shows us that the degree and the frequency of the rated items of the dataset are long-tail distribution, which is conducive to the construction of core collection of users.

5.1. Evaluation Metrics

Recommended precision and diversity are used to evaluated the performance of the method in this study.

(1) Precision

This metric considers only the top- N items of the recommendation list. For a target user u_i, the precision of recommendation $P_{u_i}(N)$, is defined as formula (4):

$$P_{u_i}(N) = \frac{R_{u_i}(N)}{N} \tag{4}$$

Where $R_{u_i}(N)$ indicates the number of relevant items, namely the items rated by u_i in the probe set (among the N recommended items).

Averaging over all the individual precisions, we obtain the precision of the whole system, as formula (5):

$$precision(N) = \frac{1}{M} \sum_{u_i} P_{u_i}(N) \tag{5}$$

Where M is the number of target users in the system. Clearly, higher precision means higher recommendation accuracy.

(2) Diversity

It considers the uniqueness of different user's recommendation list. Given two users u_i and u_j, the difference between their recommendation lists can be measured by the Hamming distance[12], defined as formula (6):

$$diversity(N) = \sum_{i=1}^{M} \sum_{j \neq i} (1 - \frac{N_{ij}}{N}) \tag{6}$$

Where N_{ij} is the number of common objects in the top- N places of both lists.

5.2. Experimental Results and Analysis

The main purpose of the experiments is to know effects of the core subset $Quser$ on the performance of CF algorithm. The value of $AvgFreq(u_i)$ is used to measure user recommendation ability, and divide all users into divided into three equal sized sets based on recommendation ability, as shown in Figure 1. In experiments, the volume of the nearest neighbor is set $K = 10, 20, 30, 40, 50, 60, 70, 80$ respectively, finally making recommendation with top-10 items.

First, verify roles of users on the performance of collaborative recommendation in in different positions of long tailed distribution.We adopt subset A, B, C shown as figure 1 to complete user-based CF, indentified by $Aucf$, $Bucf$ and $Cucf$. That is to say selecting K nearest neighbors from subset A, B, C respectively for the target users to recommend. Experimental results are shown in Figure 2 and 3.

Figure 2 and 3 show us that users in different positions of the long tailed distribution have different effects on the performance of CF algorithm. Users in the tail have the worst recommended ability. While users in the middle part have the best recommendation precision. Also users in the head have recommendation diversity. Therefore when constructing th*e g*lobal core user subset,we should try to keep users in the subset A and B shown in Figure 1.

Secondly, use core users subset $Quser$ constructed above to implement user-based CF algorithm,identified by $Qucf$. During the process of experment,adjusting parameters R, P, d_1, d_2 to make 293 users included in subset $Quser$. The recommendation performance of $Qucf$ is compared with that of $Gucf$ algorithm based on all users (943 users) to recommend, the results as shown in figure 2 and 3.

From the exoerimental results,we can know that $Gucf$ has the better recommended precision and diversity. Although that is little worse than $Gucf$, there are only 30% users are used in $Qucf$, which similarity computation complexity is much lower than that of $Gucf$. On the other hand, $Qucf$ combining the advantages of $Aucf$ and $Bucf$ shows effects of users in the head and middle part of long-tail distribution. That means that th*e g*lobal core user subset mainly retained users in the set A and B shown in Figure 1.

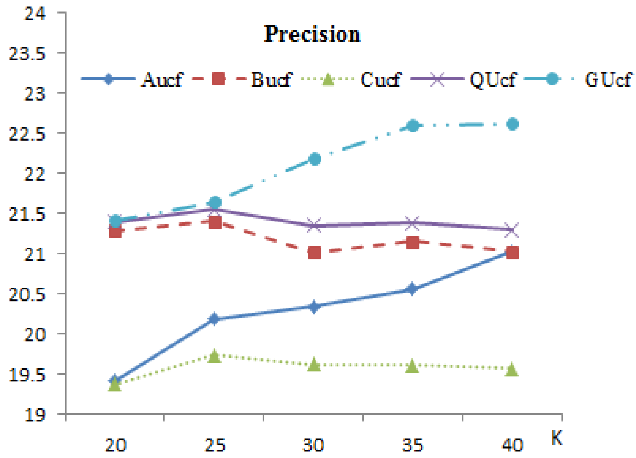

Figure 2. Comparison of Recommendation Precision

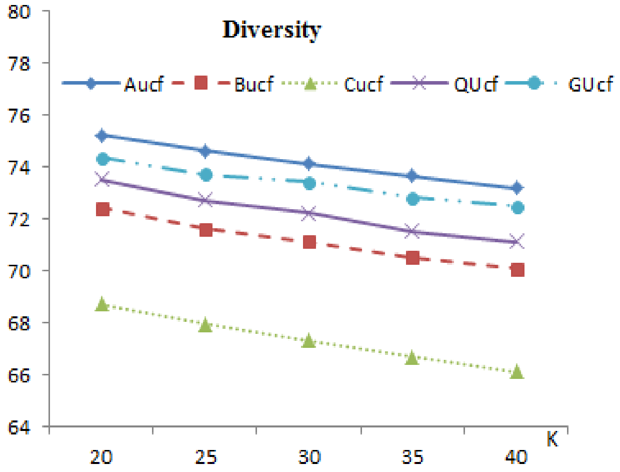

Figure 3. Comparison of Recommendation Diversity

6. Conclusions

With the development of network , the recommendation technologies are concerned by scholars in computer science, social network, e-commerce and other areas.Many recommendation algorithms are presented. But few studies have focused on the impact of users on different positions of long-tail distribution on the performance of recommendation. In this research,we study the impact of different user subset and put forward a method of constructing a core subset of users.

The experimental results show that the method proposed in the paper greatly reduces the computational complexity of the collaborative recommendation algorithm while it maintains the precision and diversity of recommendation. So the algorithm can be used in real-time recommendation environment with a great amount of users and also can be used to solve cold start problem, *i. e.*, use the core user subset to recommend for the new users of the platform.

But the algorithm proposed in this study needs to determine the best value of several parameters by experiments. They can be obtained under the online experiments, so it does not affect the recommendation real-time. Therefore, in the future work,we will validate the performance of algorithm by actual data from e-commerce platforms and propose other methods to construct the core user subset.

Acknowledgments

This work was supported by National Social Science Foundation of China (Grant No. 13BTQ027).

References

[1] X. Su and M. K. Taghi, “A Survey of Collaborative Filtering Techniques”, Advances in Artificial Intelligence, vol. 2009, no. 1, **(2009)**, pp. 1-19.

[2] C. Fidel, C. V´ictor, F. Diego and O. Vreixo, “Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems”, ACM Transactions on the Web, vol. 5, no. 1, **(2011)**, pp. 2-33.

[3] P. De Meo, E. Ferrara, G. Fiumara and A. Provetti, “Improving Recommendation Quality by Merging Collaborative Filtering and Social Relationships”, 11th International Conference on Intelligent Systems Design and Applications (ISDA), Córdoba, Spain, **(2011)** November 22-24.

[4] R. Zheng, F. Provost and A. Ghose, “Social Network Collaborative Filtering,” Preliminary Results”, Proceedings of the Sixth Workshop on eBusiness (WEB2007), Montreal, Quebec, Canada, **(2007)** December 9.

[5] T. Zou, J. Ren, M. Medo and Y. C. Zhang, “Bipartite Network Projection and Personal rRecommendation”, Physical Review E, vol. 76, no. 4, **(2007)**, pp. 046-115.

[6] A. Boumaza and A. Brun, "From Neighbors to Global Neighbors in Collaborative Filtering: An Evolutionary Optimization Approach", Proceedings of the Fourteenth International Conference on Genetic and Evolutionary Computation Conference ACM, **(2012)**, pp. 345-352.

[7] M. A. Morid, M. Shajari and A. H. Golpayegani, "Who are the Most Influential users in a Recommender System", Proceedings of the 13th International Conference on Electronic commerce, ACM, **(2011)**, pp. 19.

[8] W. Zeng, A. Zeng, H. Liu, M. S. Shang and T. Zhou, "Uncovering the information core in recommender systems", Scientific Reports, vol. 4, **(2014)**, pp. 6140.

[9] Z. Li, T. PiQiang and Q. Tao, "Using Key Users of Social Network to Solve Cold Start Problem in Collaborative Recommendation Systems", Information Technology Journal, vol. 12, no. 22, **(2013)**, pp. 7004-7008.

[10] A. Boumaza and A. Brun, "From neighbors to global neighbors in collaborative filtering: an evolutionary optimization approach", Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference, ACM, **(2012)**, pp. 345-352.

[11] A. Said, B. Fields, B. J. Jain and S. Albayrak, "User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm", Proceedings of the 2013 conference on Computer supported cooperative work, ACM, **(2013)**, pp. 1399-1408.

[12] N. Chang, M. Irvan and T. Terano, "An Item Influence-Centric Algorithm for Recommender Systems", Distributed Computing and Artificial Intelligence, 11th International Conference, Springer International Publishing, **(2014)**, pp. 553-560.

Author

Zhang Li received the Ph.D. degree in signal and information processing from Beijing University of Post and Telecommunication, Beijing, China. She is currently an associate professor of the School of Information Technology and Management Engineering, University of International Business and Economics. Her general research areas are business intelligence, data mining and recommendation systems. Her research is supported by National Social Science Foundation and University of International Business and Economics Research Fund. She has published over 30 research papers in peer-reviewed journals and conference proceedings and a book about business.