Collaborative Filtering Algorithm Based on Random Walk with Choice

Chuanmin Mi, Xiaofei Shan, Jing Ma
College of Economics and Management
Nanjing University of Aeronautics and Astronautics
Nanjing, China
cmm@ieee.org

Xin Zhang
College of Management Science and Engineering
Shandong University of Finance and Economic
Jinan, China
Zh_xin@sdie.edu.cn

***Abstract*—A brief review of the past researches on CF shows that methods for calculating users' similarities are almost Pearson Correlation or (adjusted) Cosine Similarity. This leads to same recommendations for different users because popular objects or users often win a heavier weight in the process of recommendation. Moreover, it has been increasingly recognized that the gains of the recommendation accuracy are often accompanied by the losses of the diversity. In order to walk out of the accuracy-diversity dilemma, we propose a new method named collaborative filtering based on random walk with choice which replaces the traditional Pearson Correlation or (adjusted) Cosine Similarity with random walk with choice for calculating users' similarities. Results show that our approach significantly outperforms the ordinary user-based collaborative filtering method in diversity without lowing recommendation accuracy.**

Keywords- Recommender systems; Collaborative filtering; Random walk with choice

I. INTRODUCTION

The past few years have witnessed the tremendous activity devoted to the developing of recommendation systems since the explosion of information load is far more serious than our ability to process [1]. More and more e-commercial cooperates like Amazon, Half.com, CDNOW, Netflix, and Yahoo! have adopted recommendation systems to provide customers with purchase suggestions by reference to their past purchasing records.

Although there exist many different algorithms for the design of recommender systems, the most practical and successful approach is considered as collaborative filtering [2]. The assumption of CF is that users who agreed on preferred objects in the past will tend to agree in the future [3]. The first step of CF is to calculate similarities between pairs of users from historical data and then compute discriminant scores for the candidate objects by using users' similarities. After sorting the objects by their scores, the last step is to make the recommendation from a set of candidate objects for a target user. The effectiveness of such an approach therefore largely depends on the accurate estimation of similarity between pairs of users from historical data [4]. A brief view of the past researches on CF shows that for calculating users' similarities, most of them use Pearson Correlation and (Adjusted) Cosine Similarity which will give more weight to popular objects and users. On the consequence, popular users appear more frequently in nearest neighbors' collection and objects in recommendation list for different target users are often the same ones. The basic assumption that users sharing similar preferences in history would also have similar interests in the future failed when some popular objects and users are present. Moreover, some researches pointed out that higher diversity are often on the cost of lowing accuracy [5]. It seems CF has come into the dilemma between accuracy and diversity.

Evaluation of a personalized recommendation method attracts more attention than before. Past researches tend to focus on accuracy measures such as mean absolute error [6], precision and recall [7]. But only accuracy measures are hard to evaluate a method fully and persuasively. Recent studies have increasingly recognized that diversity measures are also indispensable in order to achieve a comprehensive evaluation of a personalized method [8]. An empirical study on the relation between customers' satisfaction and recommendation diversity demonstrates that diversity plays an important part for users' feeling of a recommendation system's serviceability and usability [9]. For example, if the places recommended by a tour recommendation system are those either the customer has visited or in similar style with places the customer visited, we have to admit that the accuracy of this system is high but the recommendations are useless.

With the above understanding, we apply a method named random walk with choice to calculate the similarity of users in order to overcome the adverse influence of popular objects and users. We try to combine the advantages of both random walk with choice and CF to keep a reasonable tradeoff between the accuracy and the diversity. The rest of this paper is organized as follows. We start with covering related work in Section 2. The proposed algorithm is formally described in Section 3. The algorithm is empirically evaluated in Section 4. Finally, we conclude in Section 5 by highlighting key points of our work.

II. COLLABORATIVE FILTERING ALGORITHM BASED ON RANDOM WALK WITH CHOICE

A. Calculating users' similarity

To increase the diversity of CF, the key point lies in improving the influence of unpopular items and users. In this paper, the random walk with choice from the beginning to the end is made up of three steps shown in Fig. 1 ("△" is user and "○" is item).

© 2014. The authors - Published by Atlantis Press

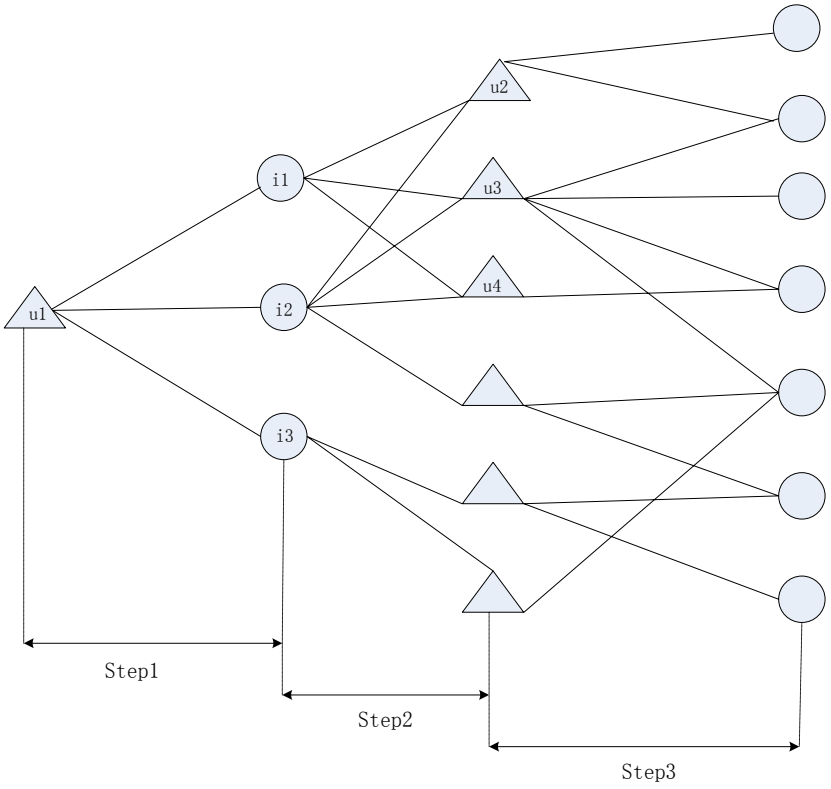

Figure 1. Three steps of random walk with choice.

In step 1, target user u_1 has rated three items $\{i_1, i_2, i_3\}$. Two candidate items will be selected from $\{i_1, i_2, i_3\}$ by generating random numbers and each one has equal probability to be selected. Supposing the selected candidate items are $C = \{i_1, i_2\}$ and then one of them will be selected by predefined strategies to be the next walking target. Here we define the first strategy for random walk with choice.

$$S_i = \frac{r_i}{k_i} \quad i \in C \tag{1}$$

where r_i is the target user's rating score for a candidate item and k_i is the item's degree (the number of users who rated the item). This strategy can be explained from two aspects. One is that item owning higher rating score can represent user's interest better than those owing lower rating score. Another is that item having lower degree can represent users' unique interest better than those owing larger degree which are so-called popular items. After step1, we will get the only option and here supposing the lucky one is i_1.

In step 2, i_1 has three fans $\{u_2, u_3, u_4\}$ besides target user. Task of step 2 is to choose candidate users from $N = \{u_2, u_3, u_4\}$ according to predefined strategies. Now we defined the second strategy for random walk with choice.

$$CN = \{u_i \mid u_i \in N, |r_i - r_0| < \min\} \tag{2}$$

where u_i is one of the users except the target user who rated the selected item i by step 1. r_i is u_i's rating score for i and r_0 is the target user's rating score for i. *Min* is a constant number we predefined. By applying the above strategies, we can get a collection of candidate users CN. If the count of the collection is more than one, it is necessary to choose the last winner from the candidate users. Supposing after step 2, we get $CN = \{u_2, u_3, u_4\}$, which means all the fans of i_1 satisfied the requirements.

In step 3, we need to select the only one from $CN = \{u_2, u_3, u_4\}$ got by step 2. First we will select two users randomly from $CN = \{u_2, u_3, u_4\}$ and then choose only one of them to be the last winner by predefined strategies. The last strategy for random walk with choice is shown in the below.

$$Neigbour = \begin{cases} u_i & if \ \ k_i < k_j \ \ u_i \in CN \\ u_j & if \ \ k_j < k_i \ \ u_j \in CN \end{cases} \tag{3}$$

where k_i, k_j are the degrees (the number of items rated by the user)of the last two users.

After the three steps, we will get the last winner and the similarity between the target user and the winner will increase. This random waking process will be carried out for hundreds of thousands of times and finally produce a similarity matrix S. $S_{i,j}$ will be added 1 if the walker starts from u_i and ends with u_j in one complete walk.

B. Predicting

We sort other users by similarity for the given user in descending order and obtain a ranking list of the users. According to the length of nearest neighbors list set before, we will remove the last few users in the ranking list and get nearest neighbors list denoted by S. Given a target user index by t and a set of candidate items (unrated ones) denoted by C_t, we calculate for each candidate $c \in C_t$ a discriminate score v_{ct} to indicate the strength of relevance between the candidate item and the target user.

$$v_{ct} = \bar{r}_t + \frac{\sum_{i \in S} s_{t,i}\left(r_{i,c} - \bar{r}_i\right)}{\sum_{i \in S} \left|s_{t,i}\right|} \tag{4}$$

where $\bar{r}_t$ is the average rating score for the target user and $\bar{r}_i$ means the same for the user denoted by i in the list of nearest neighbors. $S_{t,j}$ is the similarity between use t and user i.

Finally, we sort candidate objects for the target user in non-ascending order according to their discriminate scores and obtain a ranking list of the candidates. It is possible that a tie occurs when two or more candidate objects are assigned equal discriminant scores. In such a situation, we break the tie by putting objects with equal scores in random order.

III. THE EXAMPLE ANALYSIS

A. Validation methods

We will compare the performances of the new proposed method (RWC-CF) with the performances of two other recommendation algorithms. One of the testing algorithms is an ordinary user-based collaborative filtering algorithm (CF) whose similarity is calculated by adjusted cosine method. The other one (PL-CF) is also based on collaborative filtering but uses power law to adjust the similarity calculated by cosine method. According to the Ref.29, we set the power denoted by β as 7 (optimal value).

Three reasons can explain the choice of these two algorithms. Firstly, they have the same time complexity ($T(n) = O(n^2)$). It is possible that there exists other recommendation algorithm which outperforms RWC-CF both on accuracy and diversity. But, as we know, most of those algorithms adopt extra models such as forgetting function or trust models to improve effectiveness leading to more time cost. Secondly, they are all based on collaborative filtering and established by the same assumption that users sharing similar preferences in history would also have similar interests in the future. Thirdly, the only difference among the three algorithms lies in the similarity method and the purpose of this paper is to put up a new similarity method. Therefore, choosing these two algorithms satisfies the requirements of control test.

B. Data source

To evaluate the new algorithm, we use two benchmark datasets to carry out our analysis, MovieLens and Netflix. We run a set of experiments on MovieLens dataset (data file available at http://www.grouplens.org/) which is downloaded from the popular MovieLens site for recommending movies. MovieLens site has more than 50,000 users who have express opinions on over 3000 movies. The MovieLens dataset is a standard benchmark for recommender system techniques, containing 100,000 ratings for1682 movies by 943 users, with evaluation score ranging from 1 to 5. These movies are divided into 19 genres (action, horror, and comedy, etc.), and each movie belongs to one or several genres.

We use a repeated random sub-sampling strategy to validate the proposed approach. In each validation run, we split at random known links between objects and users into a training set that contains 80% data and a test set that contains the rest 20% data. During the experiments, we found that if the training dataset contains less than 60% data, all of the algorithms show disappointing performances because training data is not enough to train a trusted similarity among users. If the ratio between training data and test data reaches less than 8:2, we can clearly see their different performances on each criterion.

C. Results

First, we focus on MAE and Fig. 2 shows the performance of RWC-CF, CF and PL-CF on MAE using MovieLens dataset. X-axis represents the length of nearest neighbors list.

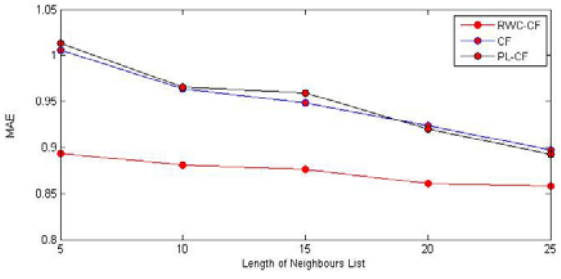

Figure 2. Performance on MAE.

From Fig. 2, we can see that RWC-CF significantly outperforms CF and PL-CF, especially when the count of neighbors is not large. This is important because length is one of the factors increasing running time of

recommendation algorithm. Real-time processing will be helpful to enhance user experiences and create new revenue. PL-CF works the worst among these three algorithms.

Then we will look at the performance of the new method on recall enhancement and precision using MovieLens dataset. Recall enhancement and precision reflect more on local properties of the accuracy in that only test objects ranked among top positions contribute to this criterion. In this paper, only if the test object which is being interested by target user appears in the recommendation list, we think this object contributes to the recall enhancement and precision. Fig. 3 and Fig. 4 show the experiment results on RE and Precision.

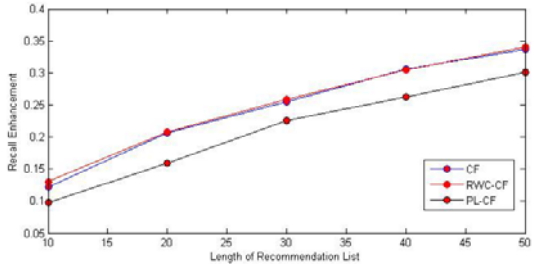

Figure 3. Performance on RE.

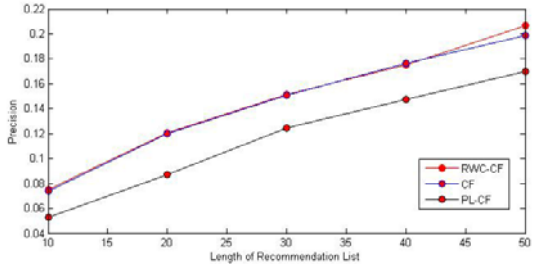

Figure 4. Performance on Precision.

What makes us disappointed is that the new method doesn't show much advantages on RE and precision as seen in Fig. 3 and Fig. 4. PL-CF demonstrates a worse performance than RWC-CF and CF. But from another perspective, if the new method outperforms CF on diversity without much loss on accuracy, it still works.

Finally, we will compare RWC-CF, PL-CF and CF on diversity measure using MovieLens dataset. The results are shown in Fig. 5. We can clearly see the positive influence of random walks with choice. By varying the length of recommendation list, diversity rate decreases. But the diversity rate can reach 0.8 with the length is 10. That is to say there are only two same recommendations of the total 10 by using the proposed method. PL-CF also shows better performance than CF on diversity. But we can't ignore the fact that PL-CF performs the worst on accuracy. The behavior of PL-CF seems to strengthen the claim that improving of diversity often takes cost of accuracy.

IV. CONCLUSION

In this paper, we have proposed a method called RWC-CF to achieve personalized recommendation by reducing the adverse effects of popular objects and popular users in the user-based collaborative filtering framework. We have demonstrated the superior performance of this approach over existing methods by large-scale validation experiments and summarized the improvements of this approach in not only the accuracy but also the diversity of recommendation results.

The success of the proposed method mainly lies in the introduction of random walk with choice in the calculating of user similarities. Certainly, the proposed method can be further investigated from the following aspects. First, we can find more reasonable strategies when using random walk. Second, we can apply this new method not only to user-based but item-based collaborative filtering algorithm. We will pursue these theoretical goals in our future work.

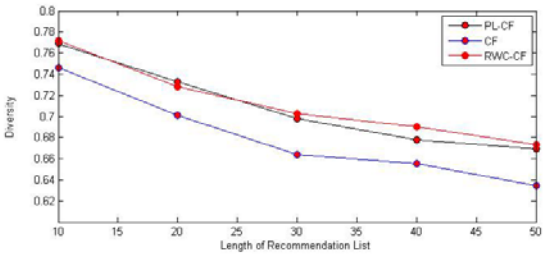

Figure 5. Performance on Diversity.

ACKNOWLEDGMENT

This work was partially supported by National Natural Science Foundation Project (71373123), the NUAA Fundamental Research Funds (NS2013080), and the Fundamental Research Funds for the Central Universities (NJ20140033).

REFERENCES

[1] Adomavicius, G., Tuzhilin, A. (2005), "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering, Vol.17, pp.734–749.

[2] T. Bogers, A. van den Bosch. (2011), "Fusing recommendations for social bookmarking web sites", International Journal of Electronic Commerce, Vol.15, pp. 31–72.

[3] Burke, R. (2002), "Hybrid recommender systems: Survey and experiments", User Modelling User-Adapted, Vol.12, pp. 331–370.

[4] G. Biau, B. Cadre, L. Rouvière. (2010), "Statistical analysis of k-nearest neighbor collaborative recommendation", The Annals of Statistics, Vol. 38, pp. 1568–1592.

[5] T. Zhou, Z. Kuscsik, J.G. Liu, M.Medo, J.R.Wakeling, Y.C. Zhang. (2010), "Solving the apparent diversity-accuracy dilemma of recommender systems", in: Proceedings of the National Academy of Sciences of the United States of America, pp. 4511.

[6] Hu Rong, Pu P. (2011), "Helping users perceive recommendation diversity", in :Proceedings of the Workshop on Novelty and Diversity in recommender Systems, New York: ACM, pp. 43 -50.

[7] Billsus, D. Pazzani, M.J. (1998), "Learning collaborative information filters", in: Proceedings of the 15th national conference on, artificial intelligence (AAAI-98).

[8] W. Wei, Q. Liu, L. Zhang. (2013), “Review on Diversity in Personalized recommender Systems”, Library and Information Service, Vol. 57 No.20, pp.127-136.

[9] D.Wei, T. Zhou, G. Cimini, P. Wu, W.P. Liu and Y.C. Zhang. (2011), “Effective mechanism for social recommendation of news”, Physica A: Statistical Mechanics and its Applications, Vol.390, pp.2117-2126.