An Improved Personalized Collaborative Filtering Algorithm in E-Commerce Recommender System

Yanhong Guo[1], Guishi Deng[2]

[1] Institute of Systems Engineering , Dalian University of Technology, 116023, China
(guoyanhongmail@sina.com.cn)
[2] Institute of Systems Engineering , Dalian University of Technology, 116023，China
(denggs@dlut.edu.cn)

ABSTRACT

Collaborative filtering recommender systems have become important tools of making personalized recommendations for products or services during a live interaction nowadays. However, there are still some drawbacks and challenges for CF based recommender system such as prediction accuracy, scalability and sparsity. This paper points out that from a certain angle, the predictions these systems produce are not really personalized ones which lead to the above problems. After the analysis of the traditional collaborative filtering algorithm, the authors then proposes a new personalized recommender algorithm based on traditional CF algorithm to improve the recommender system. At last the effectiveness and superiority of the proposed novel algorithm is proved by four experiments using both Cosine correlation similarity and Pearson correlation similarity in this paper.
Keywords: collaborative filtering, recommender system, personalized algorithm, Cosine correlation, Pearson correlation

1. INTRODUCTION

With the fast development of Internet and e-business, recommender systems have been widely used in e-business, which apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction.
Collaborative filtering (CF) is the most widely and successfully used algorithm in all kinds of recommender algorithms. [1] Most research on collaborative filtering focus on different models to improve CF recommender system and seldom have attempted to challenge the algorithm itself. From certain angle, just because the traditional collaborative filtering algorithm is not a real personalized one that lead it to many problems such as sparsity and scalability. In this paper, the authors propose a novel personalized algorithm based on the traditional CF algorithm. At last, the authors prove its superiority through experiments.

The paper is organized as follows: Section 1 is the introduction of recommender system and collaborative filtering algorithm in e-business. Section 2 is the brief overview of collaborative filtering algorithm and issues with an emphasis on its not real personalizing recommender strategy. Details of the propsed personalized CF algorithm are provided in Section 3. Section 4 describes the dataset, evaluation metrics, and experimental results. Experiments are run on open dataset associated with Movielens rating dataset. Conclusions are presented in Section 5. The authors should mainly present Section 3 and Section 4 .

2. GENERAL REVIEW of COLLABORATIVE FILTERING ALGORITHM IN E-COMMERCE

2.1 Basic idea of collaborative filtering

CF algorithm predicts a person's affinity for items or information by connecting that person's recorded preferences with that of a community of people and sharing ratings between likeminded persons. The basic idea could be shown using figure 1[4]. In this figure three users have all shown an interest in assets A, B& C (for instance they have all rented videos A B C). This high level of overlap indicates that these users have similar tastes. Further it seems a safe bet to recommend assets D and E to User 2 because they are 'endorsed' by Users 1 and 3 that have similar interests to User 2.

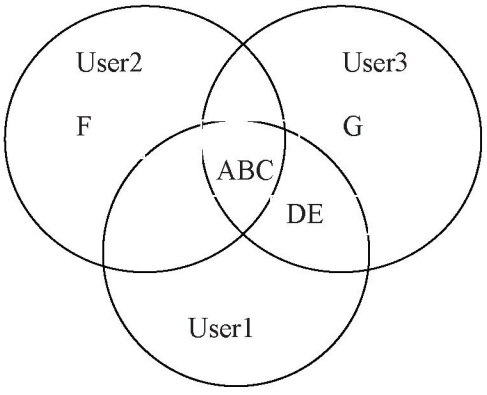


Fig.1: A Venn diagram showing interests of three users in assets ABCDEFG

CF algorithm was first put forward in 1992, [3] is one of the earliest implementations. GroupLens of the

1-4244-0451-7/06/$20.00 ©2006 IEEE

University of Minnesota use CF to recommend movies, DVDs , and receive great success in 1996 in internet. After that, CF has been widely applied in recommendation systems in e-commerce.
One of the most strengths of CF is that, if enough data is available, good quality of recommendations could be produced without any representations of the assets being recommended.

2.2 Steps of Collaborative Filtering

Most CF algorithms can be separated into three steps as addressed by Herlocker, Konstan, Borchers, and Riedl (1999)[4]:
(1) Similarity Weight: weight all users with respect to similarity with the active user, which refer to the users whose preferences are to be predicted;
(2) Selecting Neighborhoods: select those users used to make prediction;
(3) Rating Normalization and Prediction Making: normalize and calculate the weighted sum of selected users' ratings, then make predictions based on that.

Table 1: Matrix of Users ratings

	$Item_1$	$Item_2$	……	$Item_n$
$User_1$	4	5		4
……			R_{ij}	
$User_n$	0	4		3

Table 1 shows the sample of matrix of users ratings, where R_{ij} indicates the user i votes for item j and R_{ij} is between 0 and 5.
In similarity weight step, similarity could be calculated by Pearson-correlation or Cosine correlation, such as (1) and (2)[5]. Where $\overline{R}_a$ and $\overline{R}_u$ indicate the average rating of user i and the active user.

$$sim(a,i) = \sum_{y \in R_a \cap R_i} \frac{(R_{a,y} - \overline{R}_a)(R_{u,y} - \overline{R}_i)}{\sqrt{\sigma_a \sigma_i}} \quad (1)$$

$$sim(a,i) = \cos(a,i) = \frac{\overline{a} \bullet \overline{i}}{\|\overline{a}\| \|\overline{i}\|} \quad (2)$$

In Prediction Making step, the prediction of the active user on every item could be calculated as equation (3) or (4)[5]. The difference between (3) and (4) is that equation (4) reflects the different rating style of different persons. For example, some persons like to vote very high for every item but others would like to vote low. Equation (4) could give exact ratings according to its history votes.

$$p_{a,y} = \frac{\sum_{u,y \in v_u} sim(a,u) R_{u,y}}{\sum_{u,y \in v_u} |sim(a,u)|} \quad (3)$$

$$P_{a,y} = \overline{R}_a + \frac{\sum_{u,y \in vu} sim(a,u)(R_{u,y} - \overline{R}_u)}{\sum_{u,y \in vu} |sim(a,u)|} \quad (4)$$

2.3 Problems of traditional collaborative filtering

From the above, we could find some problems as follows[6]:
(1) When there is only one neighbor, for example, i has ever vote for item j as $R_{i,j}$, then the prediction of this item i is only related with $R_{i,j}$, no matter the similarity between the user i and the active user is high or low. But it is not the case in real world because there is no meaningful to make predictions when two persons have not the similar preference or taste.
(2) Even when we choose neighbors group, but because the rating matrix is sparse, there is often only one user vote for a item. Then if the $R_{i,j}$ is very high, the possibility of this item being recommended to other users is very high. For example if $R_{i,j}$ is equal to 5, then all users should get a prediction 5 on item j. Then this item should be the popular one in the recommendation list. As known to us, it is not accuracy, even very extreme, because this evaluation is only based on one person's opinion.
(3) On the contrary, if the $R_{i,j}$ is very low, then it is very possible that it never be recommended to other users. This is another case in the extreme.

The reason leads to the above problems is in that the prediction algorithm is not really personalized. The recommendation is produced based on neighbors group. But in application, we usually make recommendation only based on one person who with the similar opinion. If the recommendations are not enough, then we search for another person's help until we satisfied with the results.

3. THE PROPOSED NEW METHOD

First we give equation of making the prediction only based on one person who has high similarity with the active user.

$$p_{ai,y} = \begin{cases} \overline{p}_a + sim(i,a)(p_{i,y} - \overline{p}_i) & p_{i,y} \neq 0 \\ 0 & p_{i,y} = 0 \end{cases} \quad (5)$$

Where $\overline{p}_a$ and $\overline{p}_i$ indicate the average rating of the active user and user i in database. $p_{ai,y}$ means the predictive vote of active user for item y based on user i. Of course user i is one neighbor of the active user's neighbor aggregation.
Generally speaking, we always select those users who

have high similarity with the active user. High similarity means it could not be negative. So we could deduce （3） as follows:

$$p_{a,y} = \frac{\sum_{i \in u'} sim(i,a) \times p_{ai,y}}{\sum_{i \in u'} |sim(i,a)|}$$

$$= \frac{\sum_{i \in u'} sim(i,a) \times (\overline{p}_a + sim(i,a) \times (p_{i,y} - \overline{p}_i))}{\sum_{i \in u'} sim(i,a)}$$

$$= \frac{\sum_{i \in u'} \overline{p}_a \times sim(i,a) + sim(i,a)^2 \times (p_{i,y} - \overline{p}_i)}{\sum_{i \in u'} sim(i,a)}$$

$$= \frac{\overline{p}_a \times \sum_{i \in u'} sim(i,a) + \sum_{i \in u'} sim(i,a)^2 \times (p_{i,y} - \overline{p}_i)}{\sum_{i \in u'} sim(i,a)}$$

$$= \overline{p}_a + \frac{\sum_{i \in u'} sim(i,a)^2 \times (p_{i,y} - \overline{p}_i)}{\sum_{i \in u'} sim(i,a)} \quad (6)$$

Here U is the aggregate of neighbors for the active user and u' is the sub aggregate of U which means that the users who have all vote for one same item y. So $u' \subseteq U$. We could safely conclude that the final prediction of the active user is not linear correlation with the users' votes which accord with the real world from （6） and （8）. This is like Matthew theory that richer gets richer. For another, it avoids the problems we mentioned above in 2.3.

Then we could also get （8） with the same way. The difference between （5） and （7） is that the former considers the rating style of different person and the later is not. For example some people are not strict so they prefer to vote for each item higher than the strict ones.

$$p_{ai,y} = \begin{cases} sim(i,a) \times p_{i,y} & p_{i,y} \neq 0 \\ 0 & p_{i,y} = 0 \end{cases} \quad (7)$$

$$p_{a,y} = \frac{\sum_{i \in u'} sim(i,a) \times p_{ai,y}}{\sum_{i \in u'} |sim(i,a)|}$$

$$= \frac{\sum_{i \in u'} sim(i,a) \times sim(i,a) \times p_{i,y}}{\sum_{i \in u'} sim(i,a)}$$

$$= \frac{\sum_{i \in u'} sim(i,a)^2 \times p_{i,y}}{\sum_{i \in u'} sim(i,a)} \quad (8)$$

4. EXPERIEMENTS, RUSULTS AND ANALYSIS

4.1 Dataset

To confirm the effectiveness of the proposed personalized CF algorithm, we choose the open dataset of Movielens (http://www. Movielens. umn.edu) which is collected by the GroupLens Research Project at the University of Minnesota as our experimental dataset[7][8][9]. Movielens is a CF based online recommender system for the purpose of recommending favorite movies. Each user in this system had to rate at least 20 movies.

We choose 1000 users and 3952000 ratings randomly from the whole dataset and of the 1000 users, 800 were used as the training group, and the remaining users were used as the test group.

4.2 Evaluation metric

We also choose MAE (Mean Absolute Error) as our evaluation metric which is most widely used in CF algorithm evaluation field. The accuracy of the MAE, expressed as Equation (9), is determined by the absolute value of the difference between the predicted value and real value of user evaluation.

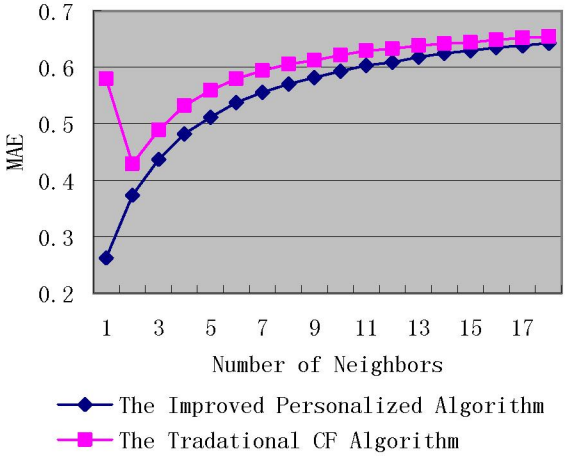
$$MAE = \frac{1}{m_a} \sum_{j \in p_a} | p_{a,j} - v_{a,j} | \quad (9)$$

In Equation (9), $p_{a,j}$ is the predicted preference, $v_{a,j}$ the real preference, and m_a the number of items that have been evaluated by the new user.

Formally, the lower the MAE is, the more accurately the recommendation engine predicts user ratings[8].

4.3 Experimental results

Graph 1: MAE of Tradational CF and The Improved Personalized CF(Cosine correlation)

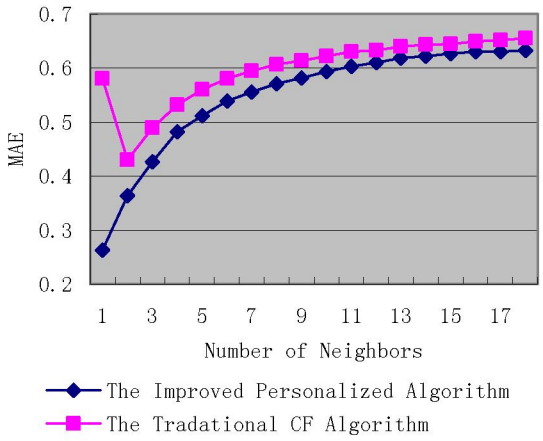


To confirm the effectiveness of the proposed algorithm, we run four experiments in all as follows: equation (6)

with Cosine correlation similarity and Pearson correlation similarity; equation (8) with Cosine correlation similarity and Pearson correlation similarity.

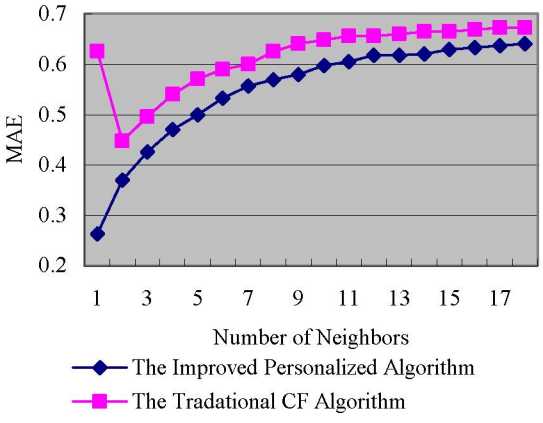
Graph 1 shows the MAE of traditional CF algorithm and our proposed personalized CF algorithm based on equation (6) with cosine correlation similarity.

Graph 2: MAE of Tradational CF and The Improved Personalized CF(Pearson correlation)



Graph 2 shows the MAE of traditional CF algorithm and our proposed personalized CF algorithm based on equation (6) with Pearson correlation similarity.

Graph 3: MAE of Tradational CF and The Improved Personalized CF(Cosine Correlation)

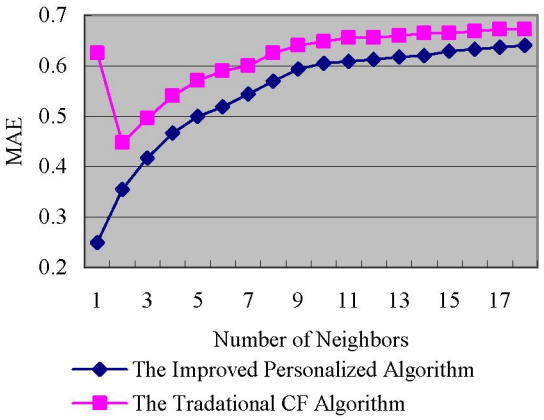


Graph 3 shows the MAE of traditional CF algorithm and our proposed personalized CF algorithm based on equation (8) with Cosine correlation similarity.

Graph 4 shows the MAE of traditional CF algorithm and our proposed personalized CF algorithm based on equation (8) with Pearson correlation similarity.

The four experiments prove the superiority from different angles.

Graph 4: MAE of Tradational CF and The Improved Personalized CF(Pearson Correlation)



4.4 Analasis

From the graphs we could find that:

(1) No matter how many neighbors we choose;
(2) No matter Cosine similarity or Pearson correlation similarity is used;
(3) No matter whether the prediction equation consider the different users rating style or not;

the line of traditional CF is higher than the line of our proposed one, which indicates that the later accuracy of prediction is higher than the former one, especially when we choose one neighbor, the MAE of the traditional one is dramatically high.

For further consideration, why the larger the number of neighbors is, the shorter the distance of the traditional line between the proposed one is?

That is because in the algorithm we proposed, the neighbors we choose is based on the sequence of the similarity of the active user. When we choose one neighbor, the neighbor's similarity with the active user is much high and the prediction equation becomes as follows:

$$p_{a,y} = \bar{p}_a + \frac{\sum_{i \in u'} sim(i,a)^2 \times (p_{i,y} - \bar{p}_i)}{\sum_{i \in u'} sim(i,a)}$$

$$= \bar{p}_a + \mathrm{sim}(i,a) \times (p_{i,y} - \bar{p}_i) \qquad (10)$$

$$p_{a,y} = \frac{\sum_{i \in u'} sim(i,a)^2 \times p_{i,y}}{\sum_{i \in u'} sim(i,a)}$$

$$= \mathrm{sim}(i,a) \times p_{i,y} \qquad (11)$$

On the contrary, MAE of traditional collaborative filtering is much high when we choose one neighbor. That is because when we choose one neighbor, the prediction produced by traditional collaborative filtering is the same as the neighbor's rating. This situation is not accord to the real world just as section 2.3 mentioned. So the prediction error is very high.

When we choose two or more than two neighbors for the active user, the distinction becomes smaller and smaller until the two lines approaches together nearly.

5. CONCLUTIONS AN D FUTURE WORK

In this paper, we have proposed a new collaborative filtering algorithm which is a real personalized one compared with the traditional collaborative filtering algorithm. The four experiments with different similarity all prove the superiority of our proposed algorithm. The new algorithm could get much better prediction than traditional one when we choose one neighbor. And our proposed algorithm show superiority always though when we choose more and more neighbors, the MAE of our proposed algorithm and the traditional one becomes closer and closer.

Our future work includes the application of the novel personalized algorithm in prototype system such as dissertation recommender system and e-commerce recommender system. We would also explore the difference of the proposed algorithm between dissertation recommender system and e-commerce recommender system. The improvement of the proposed algorithm using clustering method such as SOM (self-organization map) clustering, genetic clustering and fuzzy logic etc should be investigated.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China under grants No.70272050.

REFERENCES

[1] Yanhong Guo , Guishi Deng , "An improved Collaborative Filtering based E-Commerce Recommendation System with Case-based Reasoning", Proceedings of the International Conference on Serveice System and Service Management, pp.780-784, 2004.

[2] Conor Hayes, Pádraig Cunningham, Barry Smyth, "A case-based reasoning view of automated collaborative filtering ", Proceedings of the International Conference on Case-Based Reasoning proceedings, pp. 234-248, 2001.

[3] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. "Using collaborative filtering to weave an information tapestry", Communications of the ACM, Vol.35, No.12, pp61–70, 1992.

[4] Jon Herlocker, Joseph Konstan, Al Borchers, John Riedl, "An algorithmic framework for performing collaborative filtering", Proceedings of the 1999 Conference on Research and Development in Information Retrieval ,pp263-266, 1999.

[5] Resnick,Varian , "Recommender systems", Communications of the ACM, Vol. 40, No. 3, pp 56 C58,1997.

[6] Joseph A. Konstan, Loren G. Terveen, John T. Riedl, "Evaluating Collaborative Filtering Recommender Systems", ACM Transactions on Information Systems", Vol. 22, No. 1, pp 5–53, 2004.

[7] Su-Jeong Ko， "Prediction of Preferences through Optimizing Users and Reducing Dimension in Collaborative Filtering System", IEA/AIE 2004, pp. 1259-1268, 2004.

[8] Peng Han, Bo Xie, Fan Yang, Ruimin Shen,"A scalable P2P recommender system based on distributed collaborative filtering", Expert Systems with Applications , Vol. 27, pp 203–210,2004.

[9] Yu Lia, Liu Lu, Li Xuefeng, "A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce", Expert Systems with Applications , Vol. 28, pp 203–210,2005.