Collaborative Filtering Algorithm Based on Trusted Similarity

De Meng
School of Electrical
Nanjing Normal University
Nanjing, China
1875894736@qq.com

***Abstract*—In order to solve the problem of information overload, a large number of personalized recommendation algorithms merged. Data sparsity is one of difficult problems in these algorithms. Aim at this, a novel collaborative filtering algorithm which introduces the trust into the traditional collaborative filtering recommendation algorithm is proposed in this paper. The proposed algorithm first measures the comprehensive trust by weighting the direct trust and the indirect trust, then obtains the similarity using Pearson model, and calculates the trusted similarity for prediction at last. To verify the performance of the proposed algorithm, the Mean Absolute Error (MAE) between the proposed algorithm with adaptive coordination factor and the traditional collaborative filtering recommendation algorithm with empirical factor is compared. The result shows that the proposed algorithm has better prediction accuracy.**

Keywords—information overload; data sparsity; collaborative filtering; trust.

I. INTRODUCTION

In recent years, with the rapid development of the Internet, the number of information on the network has shown a tremendous expansion trend. Although these large amounts of information provide us with a lot of convenience and usage value, the flood of information also makes us have to pay more cost when searching for useful information. It is very difficult to find useful information in such huge amount of information simply by browsing and searching. Thus, some tools are very necessary to quickly find the information which can help us make decisions. Then, personalized recommendation technology came into being. Due to the great commercial value, it has become a hot spot for scholars in all fields.

Since the first appearance of the article on personalized recommendation algorithms in the mid-1990s, the recommendation system has gradually evolved into an independent and important field in academic research. The performance of the recommended algorithm used by the recommender system is a key factor that affects the recommended effect of the personalized recommender system. The personalized recommendation algorithms can be divided into three categories: collaborative filtering recommendation algorithm, content-based recommendation algorithm and hybrid recommendation algorithm [1]. Among them, the collaborative filtering algorithm is currently the most widely used and most successful recommendation algorithm.

The basic principle of collaborative filtering algorithm is to classify users according to the difference of preferences, and give the recommendation based on the preferences of the similar user. The authors had performed a comprehensive analysis and classified a wide range of active learning strategies, along the two descriptive and discriminative dimensions: whether they are personalized or not, and how many different item selection criteria are considered [2]. A collaborative filtering approach for predicting QoS values of web services and making web service recommendation by taking advantages of past usage experiences of service users was present [3]. The authors investigated the collaborative filtering recommendation from a new perspective and present a novel typicality-based collaborative filtering recommendation method named TyCo which selects "neighbors" of users by measuring users' similarity based on their typicality degrees instead of corated items by users [4]. A novel recommendation approach for confidence modeling by combining trust and certainty factors into a single model was proposed, and it operated at the user-level and item-level to derive the user's and item's confidence values form both local and global perspectives [5]. The authors proposed a novel trust-based approach by incorporating the trusted neighbors explicitly specified by the active users in the systems aiming to improve the overall performance of recommendations and to ameliorate the data sparsity and cold-start problems [6]. Although the recommender system has achieved some success in different fields, it has some problems such as data sparsity, cold start, recommendation accuracy and scalability, which are also the focus of the research [7].

In this paper, a collaborative filtering algorithm based on trusted similarity (CFABTS) is proposed. The algorithm introduces the trust into the traditional collaborative filtering recommendation algorithm to solve the data sparsity problem. The neighbor set of the target user is selected based on the trusted similarity which is obtained from the similarity and the comprehensive trust between users. The comprehensive trust is measured by direct trust and the indirect trust. The direct trust is calculated by the Rating grade, rating authority and rating recognition, while the

indirect trust is calculated by the direct trust based on trust propagation.

II. Traditional Similarity Measure Method

Rating similarity is a method used by traditional collaborative filtering algorithms to measure the similarity between items. Cosine similarity, Pearson Correlation similarity and adjusted cosine similarity are the most commonly methods.

A. Cosine similarity

Let u and v be the two vectors in the user-item rating matrix R, and the similarity value between them is measured by the cosine value of the angle between them as follows:

$$sim(u,v) = \cos(u,v) = \frac{\sum_{i \in I_{uv}} R_{u,i} R_{v,i}}{\sqrt{\sum_{i \in I_{uv}} R_{u,i}^2} \sqrt{\sum_{i \in I_{uv}} R_{v,i}^2}} \quad (1)$$

where I_{uv} is the set of items commonly rated by users u and v, $R_{u,i}$ denote the rating of user u on item i, $R_{v,i}$ denote the rating of user v on item i, $\bar{R}_u$ and $\bar{R}_v$ are the mean ratings of users u and v , respectively.

B. Pearson Correlation Coefficient

The Pearson Correlation Coefficient (PCC) is also a widely used similarity measure. It is often used to compute linear correlation between a pair of objects [8], which is defined as follows:

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_{uv}} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{i \in I_{uv}} (R_{v,i} - \bar{R}_v)^2}} \quad (2)$$

C. Adjusted Cosine Similarity

When the users rate on the same item, some users may have higher rating, while others have lower rating. That is, users have different rating grading on the same item. Thus, the adjusted cosine similarity considers the difference of user rating grading, and it is calculated as follows:

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_u} (R_{u,i} - \bar{R}_u)^2} \sqrt{\sum_{i \in I_v} (R_{v,i} - \bar{R}_v)^2}} \quad (3)$$

III. Collaborative filtering algorithm based on trusted similarity

To improve the quality of recommendation, a new collaborative filtering algorithm based on trusted similarity (CFABTS) which combines the trust and similarity is proposed. Fig. 1 shows the basic flow chart of the proposed algorithm.

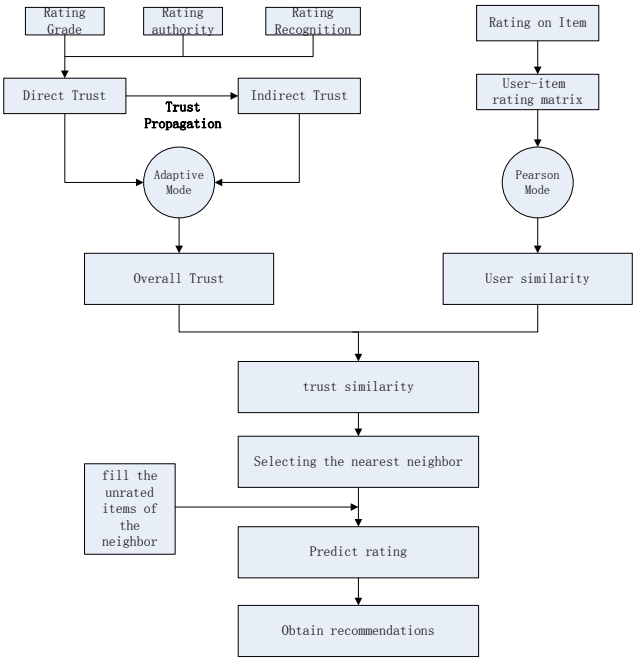

Fig. 1. Flow chart of CFABTS.

A. Calculation of Trust

The trust can be divided into direct trust and indirect trust. Direct trust refers to the direct trust relationship established through users' interactions or concerns, while indirect trust which is obtained through several direct relationships refers to the indirect relationship between users.

(a) Direct Trust between Users

To calculate the direct trust between users, following parameters need be considered: Rating grade, rating authority and rating recognition.

Rating Grade: In reality, when people ask other people's opinions, they will encounter two kinds of situations: those who are passionate and willing to make suggestions gradually gain more trust; on the contrary, those who do not like to respond to the questions gradually no longer be trusted. Similarly, some users are active in the system and willing to rate the item, while others are lazy and passive. Passive users are often reluctant to provide rating which is a contribution to the system, but only want advices from the system. Base on this, active users have more trust factors than the passive users in the collaborative filtering recommendation system. Meanwhile, more rating on the item, more valuable the rating will be. Thus, the user who makes more rating will win more trust from other users.

Define $Q_u = \{i \in I \mid R_{u,i} \neq 0\}$ as the item set rated by user u, I as the item set in the system, "$|\bullet|$" as the number of elements in the set. The Rating grade of user u can be calculated as follows:

$$N_u = |Q_u| / \max(|Q_1|, |Q_2|, ..., |Q_m|) \in [0,1] \quad (4)$$

It can be seen that N_u is a relative value of user's evaluation items, the denominator should be $|I|$, but here the maximum value that user rate the items is used. The main reason is that the $|I|$ value is too large, which will make N_u

tends to 0. The above approach can avoid this problem, and encourage users to give more rating as well.

Rating Authority: Evaluating the quality of user rating items one by one can effectively reduce the negative impact brought by malicious rating. Users with high rating quality have high authority and credibility. The rating authority of user *u* can be obtained as follows:

$$D_u = |d_u| / |Q_u| \in [0, 1] \tag{5}$$

where d_u indicates the number of items whose rating deviation is less than a certain threshold, that is, $\{i \in D_u \mid \frac{|R_{u,i} - \bar{R}_i|}{\bar{R}_i} \le \varepsilon\}$, where $\bar{R}_i$ is the mean rating of item *i*. Generally, the mean rating on the item indicates the user's real feeling to the item. Therefore, the difference between the rating of the user on the item and the mean rating on the item can indicate the quality of the evaluation of the user. ε is a constant with the range of [0,1]. The smaller the constant is, the more stringent requirements are, and the smaller corresponding D_u value is. Here, the constant ε is set as an empirical value 0. 5.

Rating Recognition: Rating recognition indicates how much a user's rating on the items has been recognized by the public. It is calculated as follows:

$$A_u = \frac{|AC_u|}{|AC_u| + |DE_u|} \tag{6}$$

where $|AC_u|$ is the number of public recognized times that the user *u* rated on item *i*, $|DE_u|$ is the number of public rejected times that the user *u* rated on item *i*. The range of A_u is [0,1].

Based on rating grade, rating authority and rating recognition, the trust degree of the user *u* can be obtained by weighting these parameters as follows:

$$T_u = w_1 N_u + w_2 D_u + w_3 A_u \tag{7}$$

where w_1, w_2 and w_3 are the weight factors with the restriction of $w_1 + w_2 + w_3 = 1$. The value of the weight factors can be taken by a variety of ways, such as expert experience, machine learning and so on. In this paper, the weight factors are setted as $w_1 = 0.28$, $w_2 = 0.47$, $w_3 = 0.25$ according to the expert experience.

One of the major features of trust is asymmetry. The user *u* and *v* should be interacted to establish trust on each other. The interaction must be based on the items rated by both user *u* and *v*.

It is assumed that during the process of establishing mutual trust, the user *u* asks the user *v* for suggestion. If the suggestion user *v* gives to user *u* is substantially the same as user *u*'s own idea, then the suggestion is regarded as a valid suggestion and the trust will be strengthened. Otherwise, regarded as a invalid suggestion. Based on this, the direct trust between user *u* and *v* can be defined as follows:

$$DT(u,v) = \frac{|C_{uv}|}{|I_{uv}|} \in [0,1] \tag{8}$$

where I_{uv} is the set of items commonly rated by users *u* and *v*, and C_{uv} is the set of items where the rating deviation between the rating of user u and the prediction rating of user v for user u is small. The definition of set C_{uv} is as follows:

$$C_{u,v} = \{i \in I_{uv} \mid \frac{|R_{u,i} - PR_{u,i}|}{R_{u,i}} < \varepsilon_1\} \tag{9}$$

where $R_{u,i}$ is the rating of user *u* on item *i*, ε_1 is the threshold of rating deviation, and it is set as 0.5, $PR_{u,i}$ is the prediction rating of user v for user *u* on item *i*, and it is defined as follows:

$$PR_{u,i} = \bar{R}_u + T_v(R_{v,i} - \bar{R}_v) \tag{10}$$

where $\bar{R}_u$ and $\bar{R}_v$ are the mean ratings of users *u* and *v*, T_v is the trust degree of user *v*.

(b) Indirect Trust between Users

Propagationis an important characteristic of trust. In a recommended system, the number of users is usually very large. Hence, the chance of direct interaction between users is very small, which leads to the fact that only a very small number of users have direct relationship with the target user. The trust propagation can find more neighbors and solve the data sparsity problem. In reality, the user usually has fewer evaluating items, and it is even rarer that two users rating on a item at the same time. Therefore, a few existing direct trust relationships should be breed, so as to broaden the trust relationship. To a certain degree, it can alleviate the data sparsity problem of traditional collaborative filtering algorithm. Assuming that user *u* trust user *k*, and user *k* trust user *v*, then we can infer that user *u* has a certain trust value for user *v* based on trust propagation. The indirect trust between user *u* and *v* is calculated by

$$IT(u,v) = \frac{\sum_{k \in U_{u,v}} DT(u,k) DT(k,v)}{\sum_{k \in U_{u,v}} DT(u,k)} \tag{11}$$

(c) Comprehensive Trust between Users

Combine the direct trust DT (u, v) and the indirect trust IT (u, v), the comprehensive trust among users can be obtained as:

$$OT(u,v) = \lambda \times DT(u,v) + (1-\lambda) IT(u,v) \tag{12}$$

where λ is the adaptive coordination factor.

After introducing the adaptive model, the algorithm can be adapted dynamically as conditions changes in real operation. The adaptive coordination factor λ is obtained as follows:

$$\lambda = \frac{DT(u,v)^2}{DT(u,v)^2 + IT(u,v)^2} \tag{13}$$

B. Trusted Similarity

The Pearson model is used to calculate similarity between two users, and obtain the user similarity matrix S_{MXM}. The comprehensive trust is used to calculate the trust value between two users, and obtain the trust matrix T_{MXM}.

The user similarity matrix and user trust matrix all have certain sparsity. By certain rules, combine the similarity matrix and user trust matrix to obtain the trusted similarity matrix, which can reduce the sparsity of the matrix to a certain extent. At the same time, it can improve the accuracy of finding neighbors, and improve user satisfaction with the recommendation. The trusted similarity is obtained as follows:

$$ST(u,v)=\begin{cases}\dfrac{2\times sim(u,v)\times OT(u,v)}{sim(u,v)+OT(u,v)}.sim(u,v)>0 \text{ and } OT(u,v)>0\\ OT(u,v), sim(u,v)=0 \text{ and } OT(u,v)>0\end{cases} \tag{14}$$

C. Rating Prediction

Based on the trusted similarity matrix, the users who have the top N highest similarity with the target predicted user are selected as the neighbor set of the target user. Since most of the selected neighbors probably didn't rating on the target item, while these neighbors have a high degree of similarity in trust with the target user, the recommendation results will have great errors if their prediction ratings are directly regarded as zero. Here, a iterative predictive rating which can fill in the unrated items in the neighborhood is proposed as follows:

$$P_{u,i}=\bar{R}_u+\frac{\sum_{k\in U} ST(u,k)\times(R_{k,i}-\bar{R}_k)}{\sum_{k\in U}|ST(u,k)|} \tag{15}$$

According to the final predictive rating, the top N items with the highest predictive rating among the non-rated projects of the target user are selected to be recommended to the user.

IV. Simulation

A. Data Set

The Epinions data set is used to validate the performance of the proposed algorithm in this paper. Epinions is an online service website where users can share some information such as product reviews and shopping comments. In addition, the users can establish their circle of friends by adding others to the personal trust list. The dataset contains the user rating information on the item and the direct trust relationship information between users. Considering the sparsity of the trust relationship in this dataset, a relatively dense subset of the dataset is selected for simulation. The processed data subset contains 3000 rating of 560 users on 340 items.

B. Performance Metric

The Mean Absolute Error (MAE) which is one of the most commonly used evaluation metric is used to verify the performance of the algorithm proposed in this paper. It measures the accuracy by the average absolute deviation between the predicted rating and the actual rating assigned by the user. The smaller the MAE value, the higher the accuracy. Given the all actual rating and predicted rating on n items in the dataset, the MAE can be expressed as:

C. Simulation Result

To understand the influences of coordination factors on the performance of the recommended model, the empirical factor (EF) with $\lambda = 0.7$ and the adaptive coordination factor (AF) proposed in this paper are compared. Fig. 2 shows the MAE of the two models. As we can see that the MAE values of the two algorithms tend to decrease first and then stabilize as the number of neighbors increases. In addition, the performance of the AF model is better than that of the EF model under different numbers of neighbors. There are two main reasons. One is that the adaptive coordination factor can automatically adjust the weights of similarities according to the similarity results so as to obtain more accurate similarity of the item. And the other is that the traditional similarity calculation method does not take into account the Rating grade, rating authority and rating recognition on the similarity. This results in the nearest neighbor may not be realistic, thus affects the quality of recommendation.

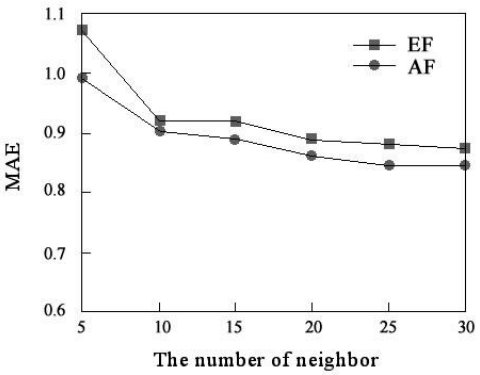

Fig. 2. The MAE of different algorithms.

V. Conclusion

To solve the data sparsity problem, a collaborative filtering algorithm based on trusted similarity which combines the trust model with the traditional collaborative filtering recommendation model is proposed in this paper. Direct trust is measured by rating grade, rating authority and rating recognition, and the indirect trust is obtained by the direct trust according to the trust propagation. Then, the direct trust and indirect trust are integrated through adaptive factor to obtain comprehensive trust. At last, the comprehensive trust and Pearson similarity is weighted to get the trusted similarity. The simulation shows that, the proposed model has better prediction accuracy than the traditional collaborative filtering model.

[1] Balabanović M., Shoham Y. Fab, "content-based, collaborative recommendation," Communications of the ACM, 40(3): 66-72, 1997.

[2] Elahi M, Ricci F, Rubens N, "A survey of active learning in collaborative filtering recommender systems," Computer Science Review, 20: 29-50, 2016.

[3] Zheng Z, Ma H, Lyu M R, et al, "Qos-aware web service recommendation by collaborative filtering," IEEE Transactions on services computing, 4(2): 140-152, 2011.

[4] Cai Y, Leung H, Li Q, et al, "Typicality-based collaborative filtering recommendation," IEEE Transactions on Knowledge and Data Engineering, 26(3): 766-779, 2014.

[5] Gohari F S, Aliee F S, Haghighi H, "A new confidence-based recommendation approach: Combining trust and certainty," Information Sciences, 422: 21-50, 2018.

[6] Guo G, Zhang J, Thalmann D., "Merging trust in collaborative filtering to alleviate data sparsity and cold start," Knowledge-Based Systems, 57: 57-68, 2014.

[7] Su X, Khoshgoftaar T M, "A survey of collaborative filtering techniques," Advances in artificial intelligence, 2009: 4, 2009.

[8] Ekstrand, M. D., Riedl, J. T., & Konstan, J. A., "Collaborative filtering recommender systems," Foundations and Trends® in Human–Computer Interaction, 4(2), 81-173, 2011.