A Novel Collaborative Filtering Algorithm Based on Trust

Yuhan Mao
School of Automation
Beijing Institute of Technology
Beijing, China
yuhan__mao@126.com

Abstract—**This paper proposes a novel collaborative filtering method based on trust, which combines direct and indirect trust information to further improve the precision of recommendations. First, user behaviors and user trust propagation are utilized to generate direct and indirect trust. Second, overall trust is calculated based on the direct and indirect trust obtained. Third, by combining the overall trust and similarity with certain weights, trust similarity can be calculated. To verify the performance of the proposed algorithm, the dataset in Epinions is used, and the Mean Absolute Error (MAE) of the proposed algorithm and the traditional collaborative filtering method is compared. The results show that the proposed algorithm outperforms the traditional collaborative filtering algorithm in terms of prediction accuracy.**

Keywords—trust; collaborative filtering; recommender system; user-based CF; trust similarity

I. Introduction

With tremendous development of social networks, our daily routine is tightly bond with all sorts of social applications. However, the large quantities of information tend to drag down our efficiencies in working and studying. To solve such overload problem, recommender system is proposed to filter out the unrelated information. The critical function of recommender systems is to recommend the items that are most likely to suit the appetite of a certain user by appropriately utilizing all the relative information of users, items and their possible interactions [1].

The recommendation algorithm is divided into three categories, content-based, CF-based and hybrid recommender systems [2]. Content-based recommender system works by abstracting the features of the items. However, the potential interests of a user would be hard to find out by simply applying CB method; meanwhile, some features of complicated items would be hard to generate. CF-based system performs better in this aspect as it can predict the scores of unrated items of a certain user, thus the generated recommendation results would break the constraints of a simple CB algorithm. However, data sparsity and cold start are the problem that researchers are trying to solve. Hybrid recommender systems take advantage of both the method to generate results with remarkable precision and span.

Built upon the traditional RS algorithms, researchers have proposed trust-based recommendation algorithms, which prove to be effective in generating recommendations. On the one hand, trust information can help deal with the data sparsity and cold start problem, which will make the predicted results more precise. On the other hand, trust information can be used to cut down the complexity of an algorithm, thus alleviating the time and resource cost of a recommender system.

Since this method very much fits the features of the present social platforms, it has now been an object of study for many researchers in this field. Haiyang Zhang et al. propose a trust-enriched approach for item based collaborative filtering [3], which takes the distances between users into the calculation of weights between them, however, this approach is simple and effective, but fails to take more specific trust relationships like how close two users are, as well as time stamp into consideration. Therefore, Shen Xiao et al. propose a method to incorporate trust relationships into social recommender systems [4], considering both influence of time and density of different types of interactions between users to calculate their trust relationships. However, this method doesn’t pay attention to distinguishing whether a relationship should be categorized into trust or distrust. An algorithm clustering users who trust each other into the same community and those who distrust each other into different communities are proposed in [5]. However, sometimes the design of social network itself focuses too much on the positive relations and fails to give information about if a user distrusts another one or not. A confident-based recommendation approach is proposed in [6] by combining the advantages of global confidence in dealing with data-sparsity problem and the precision of local confidence when data are no longer sparse. However, sometimes the global confidence may prove to be too far from the real interest of a new user. Another method called Merging trust in collaborative filtering is also proposed to deal with the issue of data sparsity and cold start [7]. Also, besides trust, time information can be considered in a SVD-based model in order to generate more precise prediction results [8]. Hong Zhou et al. take authors’ reputation into the establishment of trust models [9]. Trust propagation is considered when generating the indirect trust value using the Dijkstra’s algorithm in [10].

In this paper, a novel collaborative filtering algorithm based on trust (CFAT), which makes use of direct and indirect trust information is proposed. Direct trust is calculated taking three factors, which may influence the trust values between users, into consideration; meanwhile, trust propagation is also considered in order to generate the indirect trust value. Then, by setting particular weight to the combined trust value and similarity between users, a merged value called trust similarity can be got. After using an iterative method to deal with data sparsity, we make use of trust similarity to calculate the predicted ratings and generate the final results of recommendations.

II. RELATED WORK

The core of collaborative filtering is the calculation of the similarity between users or items, as users of similar interests are inclined to appreciate similar products. The critical steps for user-based CF calculation are specified as follows:

Firstly, most of the social applications have the mechanisms to let the users rate the items that they've viewed or bought, or let them specify the items that they like, based on these explicit information, a user-item rating matrix can be obtained. Then, by applying the Pearson correlation coefficient, we can calculate the similarity between arbitrary two users as long as they've rated enough common items.

$$Sim(u,v) = \frac{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_{u,v}}(r_{u,i}-\overline{r_u})^2}\cdot\sqrt{\sum_{i\in I_{u,v}}(r_{v,i}-\overline{r_v})^2}} \qquad (1)$$

$I_{u,v}$ is the set of items that have been rated by both u and v, $r_{u,i}$ is the rating given by u towards item i, $\overline{r_u}$ is the average rating given by user u. The same is with $r_{v,i}$ and $\overline{r_v}$. $sim_{u,v}$ is the similarity between user u and v.

If we regard u as the active user, that is to say, our target is to predict the rating of user u towards item i, $P_{u,i}$. We can rank the similarity between u and other relative users from high to low, and choose the K-nearest neighbors that are most similar to u. The final predicted value of $P_{u,i}$ can be obtained by aggregating the ratings of user u's nearest neighbor taking the weight into consideration.

$$P_{u,i} = \overline{r_u} + \frac{\sum_{v\in U_u} Sim(u,v)\cdot(r_{v,j}-\overline{r_v})}{\sum_{v\in U_u} Sim(u,v)} \qquad (2)$$

U_u is the set of nearest neighbors of user u. The items that get the highest ratings will be recommended to the active user u.

III. THE COLLABORATIVE FILTERING ALGORITHM BASED ON TRUST

To solve the problem of data sparsity, CFAT introduces trust value into the collaborative filtering algorithm. Fig. 1 shows the flow chart of the algorithm CFAT.

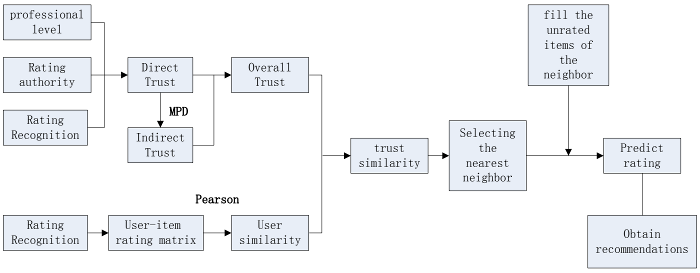


Fig. 1. The flow chart of the proposed algorithm

In this algorithm, firstly, three parameters that would affect the direct trust value between users are proposed. Second, trust propagation aiming at further completing the trust network is also taken into consideration. Third, the overall trust and similarity between two users are combined to generate the upgraded trust similarity. Then based upon that, the nearest neighbors of active user u are selected and the unrated items of his neighbors will be filled in. Finally, rating predictions are obtained according to the trust similarity and the final recommendation list can be generated.

The more specific steps are stated as follows.

A. *Creating User-item Rating Matrix*

All the rating information of users towards items can be recorded in the social networks. Therefore, we first express all these rating information in the form of matrix, active user u's rating towards a certain item i is placed in a unique position in the matrix; meanwhile, for an item that a user hasn't rated, the corresponding value in the user-item rating matrix is set to zero.

B. *Calculating User Similarity*

The user similarity between user u and v is calculated using the Pearson similarity method by (1)

C. *Calculating the Direct Trust between Users*

In this paper, we mainly propose three parameters that would influence the trust value between users: User Professional Level, User Rating Authority and User Rating Recognition.

- First is the User Professional Level.

  When a user has professional skills or experience in a certain area, it is clear that his idea will be more credible and convincing in this certain area. According to the item attributes, all the items will be divided into different topics, categories and so on. In this case, the professional user will obviously devote more energy to one or a few topics, and the ratings and comments he gives in this certain topic would certainly enjoy a much higher reputation among other users.

  We use $G_{u,i} = \{i \in I | r_{u,i} \neq 0 \cap i \in X\}$ to denote the set of items under a certain topic X that have been rated by user u. I is the collection of all items in the system, G is the set of all the items with rating records under the certain topic X, then the professional level of user u can be expressed as follows:

$$N_u = |G_{u,i}|/|G| \tag{3}$$

As can be seen, the denominator of the relative value N_u should have been $|I|$. However, since I is the collection of all items in the system, the value of $|I|$ would be too large that may lead the value of professional value to approach zero. Thus, the denominator is replaced by $|G|$ to prevent this issue. The professional level is also applied in cases that one item belongs to more than one topic.

- Second is the User Rating Authority.

User Rating Authority implies the trustworthiness of a user's rating information. One whose ratings towards items are always close to the taste of public would enjoy higher authority. To reduce the negative impact of malicious ratings, we can evaluate the quality of the ratings one by one. Users with high quality ratings will have high authority and credibility. The rating deviation of user u is denoted by D_u, and it can be expressed as follows:

$$D_u = |d_u|/|Q_u| \in [0,1] \tag{4}$$

where d_u is the set of items rated by user u that are with acceptable deviations of ratings, and $Q_u = \{i \in I | r_{u,i} \neq 0\}$ is the set of all items which have been rated by user u. The deviation of rating can be determined by a reference value. When the deviation between the user's rating and the reference value is within a certain range, such as less than a threshold ε, the deviation of rating is considered to be acceptable, and the corresponding item will be added to the set d_u.

The key issue is the selection of reference values. Usually, the mean rating obtained by an item can reflect its true quality to a great extent, thus it can be chosen as the reference value to evaluate the deviation of ratings. That is when $|r_{u,i} - \overline{r_i}| < \varepsilon$, then item i can be placed in the set $d_u (i \in d_u)$, where $\overline{r_i}$ is the mean rating of item i, ε is a constant within the range of [0,1] that acts like a threshold. The smaller the value of ε, the stricter the requirement is. That is to say, the number of items whose deviation of rating are acceptable and can be put into set d_u will be decrease as ε gets smaller. Since the number of elements in d_u decreases, the value of D_u would decrease as well. The value of ε is set to the empirical value of 0.5 as it has good experimental results, and the quality requirements are appropriate as well.

- Third is the User Rating Recognition.

The user rating recognition indicates the degree of public recognition on the ratings, and it is calculated as follows:

$$A_u = \frac{|AC_u|}{|AC_u| + |DE_u|} \tag{5}$$

where $|AC_u|$ is the number of ratings of user u accepted by the public, $|DE_u|$ is the number of ratings of user u rejected by the public.

We can see that the more ratings of a user that are accepted by the public, the higher his value of A_u would be.

Based on the definition of user professional level N_u, user rating authority D_u and user rating recognition A_u, we define the combined trust degree of user u, which can be obtained as follows:

$$T_u = w_1 N_u + w_2 D_u + w_3 A_u \tag{6}$$

where w_1 , w_2 , w_3 are the weights of N_u , D_u , A_u respectively, and $w_1 + w_2 + w_3 = 1$. There are several methods to determine the value of the weights, such as expert experience, machine learning and so on. In this paper, the weight are set to $w_1 = 0.28$, $w_2 = 0.47$, $w_3 = 0.25$ according to the expert experience.

Up to now, what we have obtained is only the general trust degree of active user u. However, in the practical situations, trust value always has two objects. Therefore, based on the trust degree of a user, we need to calculate the unique direct trust value of the user to another.

Asymmetry is a major feature of trust, which means the trust of user u towards v can be completely different from that of v to u. Thus, the direction of trust between user u and v should be distinguished when generating their trust in each other using the interactive information. The information must be gained based on the set of items commonly rated by users u and v which is denoted by $\mathrm{I}_{u,v}$. If user v's taste towards an item is similar with that of u, then u are more likely to put v into his mutual trust list. Assuming that user u asks user v for advice during establishing mutual trust relationship. The advice given by user v is more valid when it roughly the same as user u's own idea and the trust tie that links user u to user v becomes stronger. Otherwise, the advice is invalid. Then, the direct trust function between users can be defined as $\mathrm{DT}(u,v)$ using the following equation:

$$\mathrm{DT}(u,v) = \frac{|\mathrm{C}_{u,v}|}{|\mathrm{I}_{u,v}|} \tag{7}$$

where $C_{u,v}$ is the set of items with small prediction error within the span of the commonly rated items of u and v. The equation indicates that within the commonly rated item set, the more ratings of v that are similar to u, the larger the direct trust from u to v.

In this circumstance, we try to get a prediction value of user u's rating value towards item i only based on the rating towards that item given by v and compare the value with the real value to see whether the error is small enough. We denote this predicted value as $\widetilde{p_{u,i}}$:

$$\widetilde{p_{u,i}} = \bar{r_u} + T_v(r_{v,j} - \bar{r_v}) \qquad (8)$$

ε_1 is set as the threshold, when $|r_{u,i} - \widetilde{p_{u,i}}| < \varepsilon_1$, item i can be included in the set $C_{u,v}$ $(i \in C_{u,v})$. In this article, we set the value of ε_1 to 0.5.

D. *Calculating the Indirect Trust between Users*

We still take into account the model of trust propagation. Due to the asymmetry of trust, it is also a need to consider the indirect trust between users. If user u trusts user k and user k trusts user v, then it can be inferred that user u must also have a certain trust towards user v. However, the indirect trust value tends to decrease as the distance between two users increases. Therefore, we define the indirect trust between users as $IT(u,v)$, using the following equation:

$$IT(u,v) = \frac{\sum_{k \in U_{u,v}} DT(u,k) \times (DT(k,v) \times \beta_d) + \sum_{m,n \in U_{u,v}} DT(u,m) \times (DT(m,n) \times DT(n,v) \times \beta_d)}{\sum_{k \in U_{u,v}} DT(u,k) + \sum_{m,n \in U_{u,v}} DT(u,k)} \qquad (9)$$

$U_{u,v}$ is the set of users that are neighbors of both user u and v. β_d is the propagation value and is defined as:

$$\beta_d = \frac{MPD - d + 1}{MPD}, d \in [2, MPD] \qquad (10)$$

where d is the distance between two users and MPD is the maximum propagation distance of trust. In this paper, MPD is set to 3, since a too large value of MPD would again lead to information overload which would drag down the efficiency of a recommender system. Thus, when the distance between the users is $d = 2$, we can get $\beta_2 = \frac{3-2+1}{3} \approx 0.667$.

E. *Calculating the Overall Trust Between Users*

Combining the direct trust $DT(u,v)$ and indirect trust $IT(u,v)$ of user u and v, we apply a certain weight to the two different types of trusts and obtain the equation for overall trust between users as follows:

$$OT(u,v) = \lambda \times DT(u,v) + (1 - \lambda) \times IT(u,v) \qquad (11)$$

where λ is the weight of direct trust and $\lambda \in [0,1]$.

F. *Calculating the Trust Similarity*

With the trust value obtained, the trust similarity $ST(u,v)$ is obtained by weighting the overall trust $OT(u,v)$ and Pearson similarity $Sim(u,v)$ as follows:

$$ST(u,v) = \mu \times OT(u,v) + (1 - \mu) \times Sim(u,v) \qquad (12)$$

μ is the weight of $OT(u,v)$, and here it is set to 0.5.

G. *Selecting the Nearest Neighbors*

After obtaining the trust similarity between users, we can rank the similarity between the target user u and his relative users from high to low. And the first N users whose value of trust similarity rank highest in the list would be selected to form the neighbor set of the target user u. The neighbor set of user u is denoted as U_u.

H. *Predicting the Ratings*

The rating is predicted according to the identified neighbor set U_u and their rating on the target item as follows:

$$P_{u,i} = \bar{r_u} + \frac{\sum_{v \in U_u} ST(u,v) \cdot (r_{v,j} - \bar{r_v})}{\sum_{v \in U_u} |ST(u,v)|} \qquad (13)$$

$P_{u,i}$ is the final predicted rating utilizing trust similarity.

However, considering the fact that most of the selected neighbors may not have rated the target item yet, thus, in the user-item rating matrix, those positions that are needed in generating the predicted results are still set to the initial value zero. However, they have a high trust similarity with the target user, that is, they have a high weight of the predicted rating and have a greater impact on the predicted rating. If we directly take zero into the equation to generate the final predicted result, there will be a significant error. Therefore, to alleviate the error as well as making good use of the information from the users with the highest value of ST, in this paper, we apply the iterative rating prediction algorithm to fill the unrated items of these significant neighbors until the blanks in the user-item rating matrix have been filled out mostly.

I. *Obtaining Recommendations*

According to the final predictive ratings, we select the top k items whose predicted values are higher than the rest of the items and recommend them to the target user.

IV. SIMULATION

A. *Data Set*

Epinions is an online service website where users can share information such as product reviews and shopping comments with each other. Based on this, the Epinions dataset is used to verify the proposed CFAT in this paper. The dataset contains the user ratings on the items and the direct trust relationships between users. Considering the sparseness of the trust relationships, a relatively dense subset of the dataset is selected for simulation. The processed data subset contains 3000 ratings on 340 items of 560 users. In addition, the dataset also contains trust relationships among users and lists the IDs of users who have direct trust relationships.

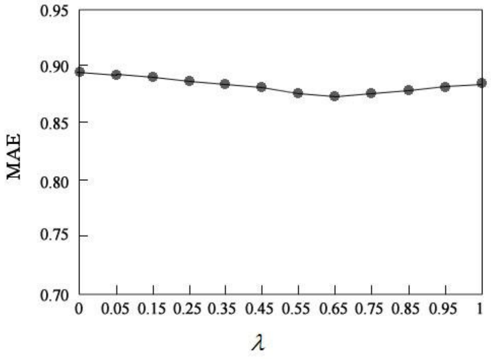
B. *Quality of Predictions(metric)*

Mean Absolute Error (MAE) is used to evaluate the prediction quality. The MAE uses the average absolute deviation between system predicted rating and the actual rating assigned by the user. The smaller the MAE value, the higher the prediction accuracy. Given all the actual ratings and predicted ratings on N items in the dataset, the MAE can be expressed as

$$\text{MAE} = \frac{\sum_{i=1}^{N} |p_{u,i} - r_{u,i}|}{N} \qquad (14).$$

C. *Simulation Results*

Fig. 2 shows the influences of λ on MAE. With the value of λ increasing, the MAE first witnesses a decrement and then begins to increase gradually. This indicates that when the single trust (directly or indirectly) takes up a too high proportion in the trust similarity calculation equation, the MAE wouldn't prove to be optimal. Drawn from the figure, the lowest MAE occurs when λ is set to 0.65, indicating that the prediction quality is the best under this value.



a.

Fig. 2. The influence of λ on MAE

To better understand the performance of the proposed algorithm CFAT, the performances of the CFAT and the traditional collaborative filtering algorithm (TCFA) are compared. The parameters μ and λ are set to 0.5 and 0.65 respectively. Fig. 3 shows the MAE of CFAT and TCFA. As can be seen, the MAE is the largest under both algorithms when the number of neighbors is 5, since it's too small to cause any low quality prediction. And the MAE gradually decreases as the number of neighboring users increases. However, the MAE changes little when the number of neighbors reaches a certain number. In addition, the MAE of CFAT is smaller than that of TCFA when the number of neighbors is set to be the same. This indicates that the prediction quality of CFAT which takes advantages of trust similarity is better than that of TCFA.

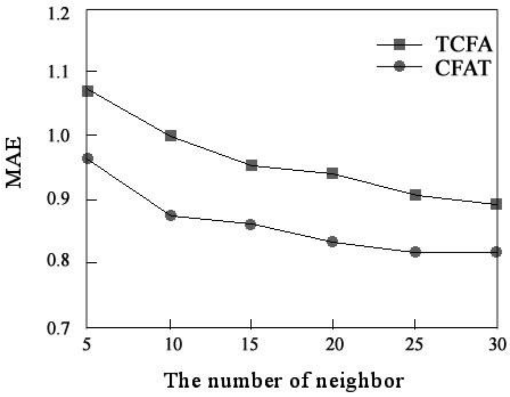


Fig. 3. The MAE of CFAT and TCFA

V. CONCLUSION

In this paper, we propose a trust-based recommendation method which is called CFAT. The trust values are obtained based on the analysis of user behavioral information and trust propagation model. Then, both direct and indirect trusts are taken into consideration to generate the overall trust value. After that, overall trust and similarity calculated by Pearson correlation coefficient are combined together to calculate trust similarity, which is then used to obtain the final rating predictions. The k items which rank highest would be recommended to the active user. Dataset from Epinion is utilized to do the simulation and MAE is chosen to test the quality of the predictions. According to the simulation results, the novel CFAT always have lower MAE value than TCFA when the number of neighbors and other possible variables are identical. Thus, the results prove that the prediction quality of CFAT outperforms that of TCFA.

REFERENCES

[1] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, Recommender system application developments: A survey[J], Decision Support Sytems, 2015 , 74 (C) :12-32

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

[3] H. Zhang, I. Ganchev, N. S. Nikokiv, and M. O. Droma," A trust-enriched approach for item-based collaborative filtering recommendations," Intelligent Computer Communication and Processing (ICCP), pp. 65-68, 2016.

[4] X. Shen, H. Long, and C. Ma,"Incorporating trust relationships in collaborative filtering recommender system," Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference, pp. 1-8, 2015.

[5] X. Ma, H. Lu, Z. Gan, and J. Zeng," An Explicit Trust and Distrust Clustering based Collaborative Filtering Recommendation Approach," Electronic Commerce Research & Applications, vol. 25, 2017.

[6] F. Gohan, FS. Aliee, H. Haghighi," A new confidence-based recommendation approach combining trust and certainty," Information Sciences, vol. 422, pp. 21-50, 2017.

[7] G. Guo, J. Zhang, and D. Thalmann," Merging trust in collaborative filtering to alleviate data sparsity and cold start," Knowledge-Based Systems, 2014, 57(2): 57-68.

[8] C. Tong, J. Qi, Y. Lian, J. Niu, and J. J.P.C. Rodrigues," TimeTrustSVD: a collaborative filtering model integrating time, trust and rating information," Future Generation Computer Systems, 2017.

[9] H. Zhou, Q. Li, and F. Zhou," Trust-aware collaborative filtering Recommendation in Reputation level," IEEE Advanced Information Technology, Electronic & Automation Control Conference, 2017: 2452-2457.

[10] X. Chen, Y. Guo, Y. Yang, and Z. Mi," Trust-based Collaborative Filtering Algorithm in Social Network," International Conference on Computer, 2016: 1-5.