A collaborative filtering recommendation algorithm based on weighted SimRank and social trust

Cite as: AIP Conference Proceedings **1839**, 020186 (2017); https://doi.org/10.1063/1.4982551
Published Online: 08 May 2017

Chang Su and Butao Zhang

A Collaborative Filtering Recommendation Algorithm Based on Weighted SimRank and Social Trust

Chang Su [a)] and Butao Zhang [b)]

College of Computer Science and Technology Chongqing University of Posts and Telecommunications Chongqing 400000, China.

[a)]changsu@cqupt.edu.cn
[b)] Corresponding author: butaozhang1990@163.com

Abstract. Collaborative filtering is one of the most widely used recommendation technologies, but the data sparsity and cold start problem of collaborative filtering algorithms are difficult to solve effectively. In order to alleviate the problem of data sparsity in collaborative filtering algorithm, firstly, a weighted improved SimRank algorithm is proposed to compute the rating similarity between users in rating data set. The improved SimRank can find more nearest neighbors for target users according to the transmissibility of rating similarity. Then, we build trust network and introduce the calculation of trust degree in the trust relationship data set. Finally, we combine rating similarity and trust to build a comprehensive similarity in order to find more appropriate nearest neighbors for target user. Experimental results show that the algorithm proposed in this paper improves the recommendation precision of the Collaborative algorithm effectively.

Key words: collaborative filtering; data sparsity; weighted SimRank; trust degree.

INTRODUCTION

In recent years, the Internet is developing rapidly, creating a large number of information data. In the face of huge amounts of data, the user cannot quickly find the information they are interested in, and Internet provider also cannot effectively recommend the new product to the user, leading to a drop in user traffic. The big data brings information overload [1] problem. To solve this problem, recommendation system plays an important role and researchers make lots of work on recommendation algorithm.

The typical recommendation system is based on collaborative filtering algorithm [2]. Recommendation system Based on the collaborative filtering has been widely used in commercial recommendation system, and has been achieved good recommendation effectiveness. However, CF algorithm exists data sparsity, cold start, malicious recommendation, etc. And these weakness restricts the development of CF to a certain extent [3].

In this paper, we focuses on how to alleviate the data sparity problem of collaborative filtering. We propose a collaborative filtering recommendation algorithm based on weighted SimRank and social trust, which can take full advantage of rating data and social trust relationship to find more appropriate nearest neighbors for target users. Rest of the paper is organized as follows. In section Ⅱ we introduce relevant work about our algorithm, and we highlight the idea of SimRank and the basic conception about social trust network. In section Ⅲ, we describe proposed algorithm in more details. In section Ⅳ, the real dataset used for evaluation is introduced and we represent the evaluation results as well as compare them with other recommendation algorithm. Finally in section Ⅴ we talk about conclusions and future works.

Materials Science, Energy Technology, and Power Engineering I
AIP Conf. Proc. 1839, 020186-1–020186-7; doi: 10.1063/1.4982551
Published by AIP Publishing. 978-0-7354-1509-6/$30.00

RELATED WORK

Collaborative filtering can be divided into two categories: memory-based collaborative filtering and model-based collaborative filtering. Memory-based collaborative filtering is calculated using the entire user-project score data set, and each user is an integral part of the score prediction process. Memory-based collaborative filtering is used to select a subset of neighbor users who have similar interest with target user and to estimate the score of the item based on the score of the neighbor user. Collaborative filtering Based on the model learns a complex model according to the training set data, and then predicts rating of target user's unrated items based on the model and the target user has been rating data.

Similarity Calculation in Collaborative Filtering

The most important part of collaborative filtering is the similarity calculation. The traditional collaborative filtering calculation equation includes Pearson similarity calculation, cosine similarity calculation and modified cosine similarity. We often use the Pearson similarity calculation equation to calculate the similarity between users according to the common item score among the users. However, when the rating information of the dataset is very sparse or there is a cold start user, Pearson similarity calculation will not effectively calculate the similarity between users. To alleviate the problem of data sparsity, researchers have made a lot of research. Reference [4] proposed a hybrid collaborative filtering recommendation algorithm based on nearest neighbor score padding. The algorithm reduces the dimensionality of the original scoring matrix and computes the similarity of the users in the low-dimensional principal component space, and reduces the algorithm complexity. Singular Value Decomposition (SVD) is used to fill in the missing value of neighbor score and reduce the sparsity of neighbor score. Jeong et al. [5] fused the user-based and project-based collaborative filtering methods to fill the unscored elements in the user-item scoring matrix using the hybrid method until the score matrix is stable. In [6], the original score matrix is filled by means of mean-filled linear regression prediction and mean-matched Bayesian classification prediction, and the accuracy of each padding method is compared. However, these improved methods have certain shortcomings, and only for a specific data set for experimental verification. In this paper, we propose an improved SimRank algorithm for computing the similarity of graphs.

The SimRank algorithm is a generalized similarity measure based on graph model and can be used to measure the similarity relation among items in structured scene. SimRank similarity core idea is: If two objects are referenced by two similar objects, then the two objects are also similar. The SimRank algorithm can alleviate the data sparsity problem in the collaborative filtering according to the transitivity of rating similarity. In the next section, we will improve the original SimRank algorithm according to the specific characteristics of the data set, in order to better compute the data between users.

Trust Network

In daily life, we tend to be more willing to accept the recommendation suggestions from the people we trust, and the trust between users influence the user's behavior to a great degree. Trust model calculation also attracts the attention of researchers in recent years. For example, Marsh introduced the trust model calculation into distributed intelligence community at the earliest [7]. John O'Donovan calculated trust values of users by the successful recommendation information of other users [8]. The page rank algorithm that Google designed whose basic idea is the more link to the page node, the higher global trust values would be obtained, has been applied many search engines [9].

In daily life, we tend to be more willing to accept the recommendation suggestions from the people we trust, and the trust between users influence the user's behavior to a great degree. Trust model calculation also attracts the attention of researchers in recent years. For example, Marsh introduced the trust model calculation into distributed intelligence community at the earliest [7]. John O'Donovan calculated trust values of users by the successful recommendation information of other users [8]. The page rank algorithm that Google designed whose basic idea is the more link to the page node, the higher global trust values would be obtained, has been applied many search engines [9].

THE PROPOSED METHOD

This paper proposes a collaborative filtering recommendation algorithm based on social trust and weighted SimRank. For the user-item scoring information in social network, we adopt the improved SimRank method, which can alleviate the problem of data sparsity according to the similar transitivity. For the trust relationship between users in social network, we use transitivity and directionality of trust to establish trust network, and integrate the similarity between users and trust Combined into a comprehensive similarity, then replace the traditional similarity calculation equation to find more appropriate neighbor users for the target user. The process of algorithm we proposed is shown in Figure 1.

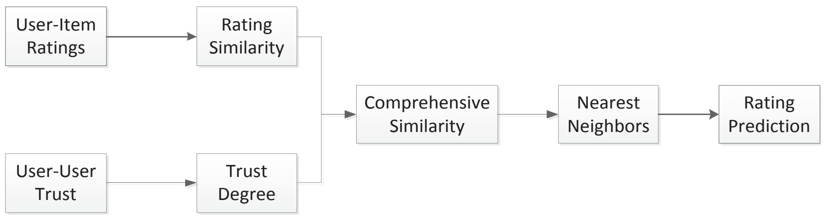

FIGURE 1. The process of proposed method.

The Weighted SimRank

As we know, original SimRank can calculate the similarity between nodes in the topological structure diagram. SimRank can be expressed as equation (1):

$$s(A, B) = \frac{C}{|O(A)|\,|O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} s(O_i(A), O_j(B)) \tag{1}$$

In the equation (1), C denotes decay coefficient and varies from 0 to 1; O (A) denotes the nodes set that node A points to and |O (A)| denotes out-degree of node A. Since the SimRank algorithm only uses the structure information of the directed graphs to compute the similarity between nodes, it does not take into account the weight information on the directed edges. So, we can't utilize original algorithm to calculate similarity. For a vivid and intuitive description, we use keywords searching and results clicking figure to represent.

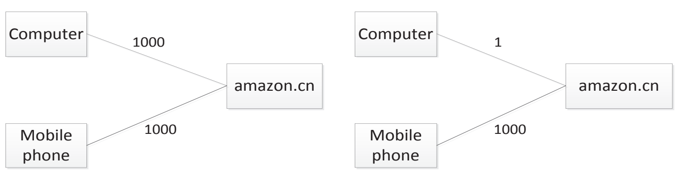

FIGURE 2. Keywords searching and results clicking

In Figure 2, computer and mobile phone are the keywords, and amazon.cn is the website clicked by users. The number 1000 and 1 represent the page view of amazon.cn. If we use original SimRank to calculate the similarity between computer and mobile, we will find the similarity between two keywords is same in the left and fight parts of Figure 2. However, we can intuitively see that the similarity in left part is higher than the right part because page view in left part is relative balanced. From this angle, we improve original SimRank algorithm through adding weight between nodes.

In the user ratings bipartite graphs, we denote the ratings as weight, and we normalize the weights on the same side. The weighted SimRank is shown as equation (2):

$$S(A, B) = \frac{C + T}{|O(A)|\,|O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} W(O_i(A))W(O_j(B))S(O_i(A), O_j(B)) \tag{2}$$

In Figure 2, we append a parameter T, which represents the ratio of common trust user amount of user A and user B to total trust user amount. Parameter T can be understood as: if two users have strong trust relationship, then the rating similarity is more similar:

$$T = \frac{TR(A) \cap TR(B)}{TR(A) \cup TR(B)} \tag{3}$$

In equation (3), TR (A) denotes the trust user set of user A, and TR (B) denotes the trust user set of user B.

The symbol W (O_i (A)) represents the normalized weights on the same edge:

$$W(O_i(A)) = \frac{R_i(A)}{\sum_{i=1}^{|O(A)|} R_i(A)} \tag{4}$$

In equation (4), Ri(A) denotes one of the weights of node A, and rating data in our paper can be expressed as the weight.

Trust Network and Trust Degree Calculation

The trust network can be expressed as a directed graph, nodes as users, and edges as trust relationships. According to the form of trust value, the trust network can be divided into two types of non-binary and non-binary. Epinions is a typical binary trust network, in which the number 0 denotes distrust and number 1 denotes trust. We can simply demonstrate the Epinions trust network from Figure 3. We can see from Figure 3 that user A has a direct trust about user B and user F. However, in real life, the trust degree of A to B may be different from the trust of A to F, that is, the relative difference of trust. In this paper, the trust between users is quantified according to the data set information.

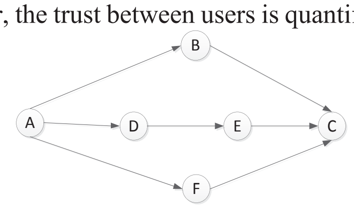

FIGURE 3. Trust network

In daily life, we can build trust relationship by the transitivity of trust. For example, user A has a direct trust to user B, and user B has a direct trust to user C. We can calculate indirect trust degree that user A towards user C by the transitivity. In trust network, we can calculate the direct trust degree that user A towards user B from equation (5):

$$DT(\mathrm{A, B}) = \frac{Ind(\mathrm{B})}{\sum_{n \in N_A} In(\mathrm{n})} \tag{5}$$

In equation (5), Ind(B) is expressed as the in-degree of user B, and we can also understand as the user set that has a direct trust to user B. The symbol NA denotes the user set that are trusted by user A.

In Figure 3, if user A has not a direct trust to user C, we can calculate the indirect trust degree that user A towards user C from equation (6):

$$ID(\mathrm{A, C}) = \frac{1}{|\mathrm{NPATH(A, C)}|} \sum_{i=1}^{|NPATH(\mathrm{A,C})|} \frac{DT_i(\mathrm{A, X_1}) * \mathrm{DT}_i(\mathrm{X_1, X_2}) * \cdots * DT_i(\mathrm{X}_m, \mathrm{C})}{PLEN(\mathrm{A, C})} \tag{6}$$

In equation (6), |NPATH (A, C)| denotes the path amount from user A to user B, PLEN (A, C) denotes path length from user A to user B, $DT_i(A,X_1)*DT(X_1,X_2)*\ldots*DT(X_m, C)$ denotes one of the indirect trust degrees that user A towards user B.

Comprehensive Similarity and Rating Predicting

We build a new comprehensive similarity equation by integrating the weighted SimRank and trust degree to find more appropriate nearest neighbors for target user. In this paper, we use the weighted harmonic mean calculation method, which is commonly used in statics, to reflect the common contribution of rating similarity and trust degree. The harmonic function is as follows:

$$H_n = \frac{\sum_{i=1}^{n} m_i}{\sum_{i=1}^{n} \frac{m_i}{X_i}} = \frac{1}{\frac{1}{m_1 + m_2 + \ldots + m_n}\left(\frac{1}{X_1} m_1 + \frac{1}{X_2} + \ldots + \frac{1}{X_n} m_n\right)} \tag{7}$$

In equation (7), Hn denotes weighted harmonic mean; mi denotes weight coefficient that weighs contribution of variate Xi to Hn, which also means the importance of Xi. In this paper, we make an equal allocation about rating similarity and trust degree, and the comprehensive similarity can be expressed as follows:

$$TR(\mathrm{A, C}) = 2 * \frac{\mathrm{S(A, C)} * \mathrm{T(A, C)}}{S(\mathrm{A, C}) + \mathrm{T(A, C)}} \tag{8}$$

In equation (8), S (A, C) denotes the rating similarity, which is calculated by weighted SimRank, and T (A, C) denotes the trust degree that user A towards user C.

Then, we can predict rating which user a score for item i by the follow aggregate function:

$$p_{\mathrm{A},i} = \bar{A} + \frac{\sum_{V \in s} TR(\mathrm{A, V})(\mathrm{R}_{v,i} - \bar{V})}{\sum_{V \in s} TR(\mathrm{A, V})} \tag{9}$$

In equation (9), $\bar{A}$ denotes average rating of user A, and $\bar{V}$ denotes average of one of the neighbors of user A. RV, i denotes the rating that neighbor user V score for item i.

EXPERIMENT RESULTS AND EVALUATION

The data set we have used in our experiment is the Epinions data, in which has trust relationship between users. Therefore, it is suitable to do social trust recommendation using Epinions. This data set contains two data information: one is 665k rating data which 49k users made for 140k items, the other is 487k trust relationship between 49k users. In this paper, we divide Epinions data into training set and testing set. The rating data is divided into training set and test set according to the ratio of 8: 2, while all the trust relationship is used for training set.

Mean Absolute Error (MAE) and Root Square Mean Error (RMSE) are recommendation assessment criteria that are most commonly used. MAE measures the accuracy of predicting rating by calculating the deviation between the users' real rating and predicting rating. RMSE is Square root of the ratio of the sum of squares of the predicted and actual deviations to the number of observations. The smaller MAE and RMSE are, the higher efficiency and quality of the algorithm is. MAE is shown in equation (10):

$$MAE = \frac{\sum_{i=1}^{N} \left| p_{\mathrm{A},i} - r_{A,i} \right|}{N} \tag{10}$$

RMSE is shown equation (11):

$$RMSE = \frac{\sqrt{\sum_{i=1}^{N} (\mathrm{p}_{\mathrm{A},i} - \mathrm{r}_{\mathrm{A},i})^2}}{N} \tag{11}$$

In equation (10) and equation (11), $P_{A,i}$ denotes the predicting rating of user *A* made for item *i*, and $r_{A,i}$ denotes the real rating. N denotes rating amount of users made for items in the test set.

Experiment Design and Analysis

We need to set the attenuation coefficient C and the similarity propagation when calculating the similarity of user rating using the improved SimRank algorithm. Literature [10] proposed that iteration can be obtained more accurate results when the coefficient C is 0.8. So, we set C to the empirical value of 0.8[11]. By the same, the similarity propagation radius and the time of iterations are empirically set to 2 and 5[11].

To verify the validity of proposed algorithm named STC, we conduct experimental contrast. The algorithm taking part in our contrast experiments are: The traditional collaborative filtering algorithm based on users (UCF), and Fusing Trust and Rating (FTRA) which improves the recommendation of collaborative filtering by fusing trust network [12]. The experiment result of MAE and RMSE is shown Figure. 4 and Figure. 5:

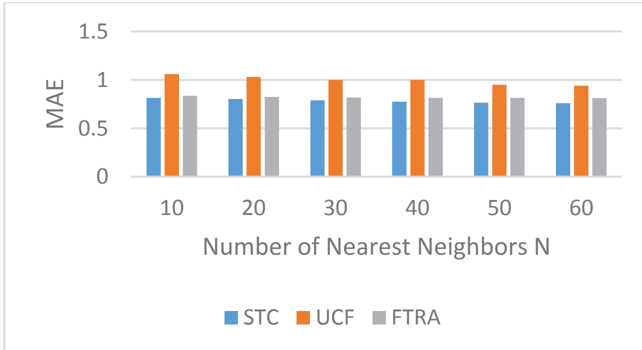

FIGURE 4. Contrast experiment in the situation of different neighbor amount

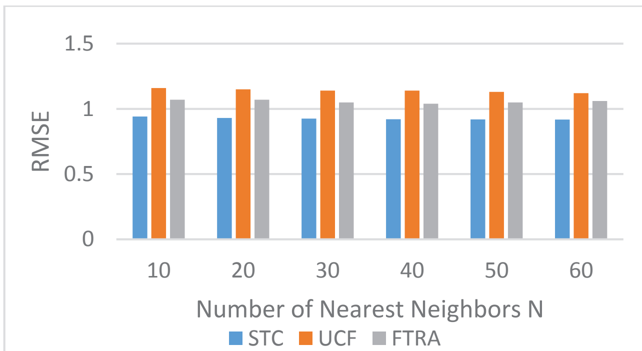

FIGURE 5. Contrast experiment in the situation of different neighbor amount

The experimental results show that the MAE and RMSE of the algorithm are smaller as the number of nearest neighbors increases. When the number of neighbors increases to a certain level, the MAE and RMSE become stable, and the MAE and RMSE of the STC algorithm proposed in this paper are less than those of other algorithms in a certain extent, which indicates that the algorithm proposed in this paper can improve the recommendation precision. From the Figure 4 and Figure 5 we can see that the recommended effect is significantly improved compared to the traditional collaborative filtering algorithm UCF. This is because we use the improved weighted SimRank to calculate the similarity between user ratings. Compared with traditional Pearson similarity calculation, weighted SimRank performs better on sparse data sets, and proposed algorithm STC contains social trust calculation, which can find more appropriate nearest neighbors for the target user. Compared with the FTRA algorithm, the proposed algorithm better reconciles the similarity and trust between users, and the similarity transitivity of weighted SimRank algorithm can find more neighbor users for the cold start target users.

CONCLUSION

To alleviate the weakness of data sparsity and cold start in traditional collaborative filtering algorithm, we propose a collaborative recommendation algorithm based on weighted SimRank and social trust. For the user-item rating information in the data set, the improved weighted SimRank is used to calculate the similarity between the users. The weighted SimRank can find the appropriate neighbor users according to the transitivity of rating similarity, especially for the cold start users. We establish the trust network, and gives the equation to calculate the trust degree among the users, and fully consider the item preference factors among the trusting users. Experiments show that the proposed algorithm can improve the recommendation accuracy to a certain extent, and prove the effectiveness of the proposed algorithm. In the future research, we will explore a better method to calculate the similarity between users and verify the validity of the proposed algorithm on other datasets.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China under contract number 61271259, the Chongqing Nature Science Foundation under contract number CSTC2016jcyjA0398, CTSC2011jjA40006, CTSC2012jjA40038 and the Research Project of Chongqing Education Commission under contract number KJ120501C.

REFERENCES

1. Xu, W., Sun, J., Ma, J., Du, W.: A personalized information recommendation system for R&D project opportunity finding in big data contexts. Journal of Network and Computer Applications 59, 362-369. 2016.
2. Adibi, P., Ladani, B.T.: A collaborative filtering recommender system based on user's time pattern activity. In: Information and Knowledge Technology (IKT), 2013 5th Conference on, pp. 252-257. IEEE 2013.
3. Elahi, H., Ricci, F.: Active Learing in Collaborative Filtering Recommender Systems. In: Hepp, M., Hoffner, Y. (eds.) 15th International Conference, EC-Web. LNBIP, vol. 188, pp. 113-124. Springer, Heidelberg 2014.
4. Leng Y J, Liang C Y, LuQ,et al. Collaborative Filtering Recommendation Algorithm Based on Neighbor Rating Imputation. Computer Engineering, 2012, 38(21): 56－58, 66.
5. Jeong B, Lee J. An Iterative Semi-explicit Rating Method for Building Collaborative Recommender Systems. Expert SystemswithApplications,2009,36(3) : 6181－6186
6. Su X Y, Khoshgoftaar T M, Greiner R .A Mixture Imputation Boosted Collaborative Filter // Proc of the 21st International Florida Artificial Intelligence Research Society Conference. Coconut Grove, USA, 2008: 312－316.
7. Marsh, S.: Formalising trust as a computational concept. Britain: University of Stir-ling.1994.
8. Donovan, J., Smyth, B., Eliciting Trust values from Recommendation Errors. In: Proceed-ings of the 18th International Florida Artificial Intelligence Research Society Conference. pp. 289-294. 2005.
9. Page, L., Bin, S., Motwani, R., Winograd, and T.: The PageRank Citation Ranking Bring Order to the Web. Technical report, Stanford Digital Library Technologies Project. 1998.
10. Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D.: Accuracy estimate and optimization techniques for SimRank computation. The VLDB Jounal 19(1), 45-66. 2010.
11. G. J eh, J. Wid om. SimRank: Ameasure o f s tructural con text similarity [C] / / Proceedings of S IGKDD. Edmonton, Alberta, Canada, 2002: 538 -543.
12. Yang, B., Zhao, P., Huang, and S.P.: Improving the Recommendation of Collaborative Filtering by Fusing Trust Network. In: Computational Intelligence and Security (CIS), 2012 Eighth International Conference on, pp. 195-199. IEEE 2012.