Collaborative filtering recommendation algorithm integrating user similarity and trust

Mengge Huang1,2, Kai Cao1,2, Jingyi Zhang1,2, Chuanlin Zhang2,3 and Tao Deng1,2,3,*

1 Key Laboratory of China’s Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, 730030, P. R. China
2 Key Laboratory of Streaming Data Computing Technologies and Application, Northwest Minzu University, Lanzhou, 730030, P. R. China
3 School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, 730030, P. R. China
* Corresponding author’s e-mail: dttom@lzu.edu.cn

Abstract. The rapid development of the Internet has brought convenience to people and has also produced the problem of “information overload”. In view of the traditional collaborative filtering algorithm facing some bottlenecks to be solved, this study proposes a collaborative filtering algorithm that combines similarity and trust. First of all, in view of the large deviation of traditional similarity calculation and prediction of user ratings, this study proposes an optimized Pearson correlation coefficient calculation method; secondly, the trust relationship is established based on the user’s rating of the common project, and the trust relationship between users who do not have a direct trust relationship is established through the transfer of trust; then find the nearest neighbor set of the target user through the fusion of user similarity and trust; finally, the item is scored and predicted to generate a recommendation list. Experimental results show that the algorithm proposed in this study can effectively improve the accuracy of recommendation.

1. Introduction

As of December 2020, the number of Internet users in my country is 989 million, and the Internet penetration rate has reached 70.4%[1]. With the rapid development of the Internet, the amount of information on the Internet has increased dramatically, and the number and types of products have grown rapidly. It takes a lot of time for users to find the products they need, which causes the “information overload” problem [2-4]. In order to solve this problem, a personalized recommendation system came into being, which is widely used in the fields of e-commerce, movies and videos, music, reading, personalized mail, and advertising.

Collaborative filtering is one of the most widely used technologies in recommendation systems. The collaborative filtering recommendation algorithm discovers user preferences by mining user historical behavior data, divides users into groups based on different preferences, and recommends products with similar tastes. Collaborative filtering recommendation technologies are mainly divided into three categories, namely user-based collaborative filtering[5], item-based collaborative filtering[6], and model-based collaborative filtering[7,8]. However, collaborative filtering algorithms have problems such as cold start, data sparseness, and scalability. In the era of social networks, it is easy for users to build their own trust network, and adding user trust information to the traditional recommendation
process can effectively solve the problems faced by traditional recommendation systems. Zhang et al. [9] proposed a new trust measurement model that integrates user interaction information, preference and trust, and used an improved structural hole algorithm to identify influential users in the network and improve the performance of top-N item recommendation. Tong et al. [10] proposed an extended Pawlak conflict model on the trust mechanism to solve the problem of the consensus process, and verified the effectiveness and superiority of the proposed method through examples. Zou et al. [11] proposed a user trust recommendation algorithm based on topic-based tensor decomposition, which is used to mine the trust level of users to different friends when selecting different items.

This study first proposes an optimized similarity calculation method to measure the user’s preference for items, and secondly calculates the direct trust value between users based on the user’s common scoring items. Users who do not have a common scoring item can be indirectly through the transfer of trust. The trust relationship is built up, and finally the similarity and trust between users are combined to find the set of nearest neighbors of the target user, and the score prediction is performed to generate recommendations. The implementation results show that, compared with the traditional collaborative filtering recommendation algorithm, the method proposed in this study can effectively improve the accuracy of recommendation.

2. User similarity

2.1. Build user-item rating matrix

Assuming there are m users and n items, the user-item rating matrix is shown in Table 1.

| User  | Item 1  | Item 2  | ...... | Item n |
|-------|---------|---------|-------|--------|
| User1 | r1,1    | r1,2    | ...... | r1,n   |
| User2 | r2,1    | r2,2    | ...... | r2,n   |
| ...... | ......   | ......   | ...... | ......  |
| Useri | ri,1    | ri,2    | ...... | ri,n   |
| ...... | ......   | ......   | ...... | ......  |
| Userm | rm,1    | rm,2    | ...... | rm,n   |

Among them, the element $r_{ij}$ in the $i$-th row and the $j$-th column represents the rating of the $i$-th user on the $j$-th item.

2.2. Calculate user similarity

Calculate the user similarity according to the user-item rating matrix. There are many commonly used methods for calculating user similarity, but in related studies, cosine similarity and Pearson correlation coefficient are mostly used to measure user similarity. Therefore, this article uses the Pearson correlation coefficient to measure user similarity. The calculation formula of Pearson's correlation coefficient is shown in formula (1):

$$
Pearson(i, j) = \frac{\sum_{k=1}^{I_{i,j}} (R_{i,k} - \overline{R_i})(R_{j,k} - \overline{R_j})}{\sqrt{\sum_{k=1}^{I_{i,j}} (R_{i,k} - \overline{R_i})^2} \sqrt{\sum_{k=1}^{I_{j,k}} (R_{j,k} - \overline{R_j})^2}}
$$

Among them, $R_{i,k}$ represents the rating of user $i$ on item $k$, $\overline{R_i}$ represents the average of user $i$'s ratings on all items, $R_{j,k}$ represents the rating of user $j$ on item $k$, $\overline{R_j}$ represents the average of user $j$'s ratings on all items, and $I_{i,j}$ represents a set of common scoring items of user $i$ and user $j$.

Although the Pearson correlation coefficient solves the problem of different user rating levels, it does not consider the impact of the number of overlapping rating items between users on the similarity.
Table 2. User-item rating data.

|    | I1 | I2 | I3 | I4 | I5 | I6 | I7 | I8 |
|----|----|----|----|----|----|----|----|----|
| U1 | 2  | 1  | 4  | 5  | 3  | 4  | 5  | 2  |
| U2 | 4  | 3  |    |    |    |    |    |    |
| U3 | 2  | 5  | 3  | 4  | 5  |    |    |    |

It can be seen from Table 2 that U1 evaluated 8 items, U2 evaluated 3 items, and U3 evaluated 5 items. Using formula (1), we can get Sim(U1, U2) = 1, Sim(U1, U3) = 1, that is, the similarity between U1 and U2 and the similarity between U1 and U3 are both 1. Although the calculation is correct, U2 and U1 have 3 similarity scores, and U3 and U1 have 5 similarity scores. We think U3 and U2 are more similar to U1, so the number of items that users have evaluated together on the basis of Pearson's correlation coefficient is added to better measure the similarity between users. The improved Pearson correlation coefficient formula is shown in formula (2):

$$\text{Sim}(i, j) = \frac{2I_{i,j} - \sum_{k=1}^{I_{i,j}} (R_{i,k} - \overline{R_i})(R_{j,k} - \overline{R_j})}{\sqrt{\sum_{k=1}^{I_{i,j}} (R_{i,k} - \overline{R_i})^2} \sqrt{\sum_{k=1}^{I_{j,j}} (R_{j,k} - \overline{R_j})^2}}$$

Among them, I_{i,j} represents the total number of items evaluated by user i and user j, I_{i} represents the total number of items evaluated by user i, and I_{j} represents the total number of items evaluated by user j.

3. Calculating trust

3.1. Direct trust

There is no data that directly expresses the degree of trust in the data set, so the direct trust degree is calculated by considering the items jointly evaluated by users. Two users rate the same item and think that the two users have an exchange. According to the user's common rating items, the initial direct trust between users is calculated, and the calculation method is shown in formula (3):

$$\text{Init}_{a,b} = \frac{I_{a,b}}{I_a}$$

Among them, I_{a,b} represents the number of items evaluated by user a and user b, and I_{a} represents the number of items evaluated by user a.

If there are more common evaluation items between two users, the more communication between the two users will be. In real life, the direct trust relationship between people changes with the result of communication. If their communication and results are good, then their trust relationship will be strengthened; if their communication and results are poor, then their trust relationship will be weakened. Calculate the user’s average rating based on the user’s rating of the item. If the user’s rating for an item is greater than the user’s average rating, it means it is a positive rating; conversely, if the user’s rating for an item is less than the user’s average rating, it means it is a negative score. If user a and user b’s ratings for item i are both positive or negative, it means that the results of the communication between user a and user b are successful, that is, the users have the same opinion; if user a and user b have the same opinion on item i One of the scores is a positive score and the other is a negative score, which means that the communication results of user a and user b are failed, that is, the opinions between users are inconsistent; the expression for determining the success or failure of the communication is shown in formula (4):

$$\begin{cases} \text{success}, & (r_{a,i} - \overline{R_a})*(r_{b,i} - \overline{R_b}) \geq 0 \\ \text{failure}, & (r_{a,i} - \overline{R_a})*(r_{b,i} - \overline{R_b}) < 0 \end{cases}$$

Among them, r_{a,i} and r_{b,i} respectively represent the ratings of user a and user b on item i, and \overline{R_a} and \overline{R_b} represent the average ratings of user a and user b, respectively.
Use $s$ to represent the number of successful communication between users, and $f$ to represent the number of failed communication between users. Each time the communication between users is successful, then $s+1$ ($i \in I_s$); each time the communication between users fails, then $f+1$ ($i \in I_f$).

Calculating the scoring deviation for both parties of the exchange is convenient to measure the influence of user preference deviation on direct trust. The trust impact factor between users is added to the exchange information between users, and the direct trust between users is calculated. The calculation formula is shown in formula (5):

$$DT_{a,b} = Init_{a,b} \cdot \frac{\sum_{i \in I_s} s^*(r_{a,i} - r_{b,i})^2}{\sum_{i \in I_s} s^*(r_{a,i} - r_{b,i})^2 + \sum_{i \in I_f} f^*(r_{a,i} - r_{b,i})^2}$$ (5)

Among them, $r_{a,i}$ represents the rating of user $a$ on item $i$, $r_{b,i}$ represents the rating of user $b$ on item $i$, $I_s$ represents the set of judging items for successful communication between users, and $I_f$ represents the set of judging items for failed communication between users.

3.2. Indirect trust

In social networks, if there is a connection between users, there is a direct trust relationship between users. When there is no direct trust relationship between users, it is necessary to calculate the indirect trust relationship between users through the propagation of trust. User $a$ and user $b$ have a direct trust relationship, and user $b$ and user $c$ have a direct trust relationship. Through trust propagation, the indirect trust relationship between user $a$ and user $c$ can be inferred. If each intermediate user is regarded as user $b$, then the indirect trust calculation formula between source user $a$ and target user $c$ is shown in formula (6):

$$IT_{a,c} = \frac{\sum_{b: \text{adj}(a,c)} DT_{a,b} \cdot (DT_{b,c} \cdot \beta_d)}{\sum_{b: \text{adj}(a)} DT_{a,b}}, DT_{a,b} \geq \hat{\delta}, DT_{b,c} \geq \hat{\delta}$$ (6)

Among them, $\beta_d$ is a weight parameter to determine the optimal trust propagation distance, $\beta_d = \frac{d_{\text{max}} - d + 1}{d_{\text{max}}}$, $d_{\text{max}}$ represents the maximum trust propagation distance, according to the six-degree separation theory, the maximum value of $d_{\text{max}}$ is 6, and $d$ represents the distance from the source user to the target user. $DT_{a,b}$ and $DT_{b,c} \in [\hat{\delta}, 1]$ represent the direct trust of user $a$ and user $b$, and user $b$ and user $c$, respectively. The parameter $\alpha \in (0,1]$ is an adjustable trust threshold, only users whose direct trust degree is greater than this threshold can participate in the trust propagation process.

In summary, the trust degrees of user $a$ and user $b$ are represented by $Trust_{a,b}$, then:

$$Trust_{a,b} = \begin{cases} DT_{a,b} & d = 1 \\ IT_{a,b} & 2 \leq d \leq \text{MPD} \end{cases}$$ (7)

4. Integration of similarity and trust

From the above, the user similarity and user trust are obtained, and the two are weighted and combined to obtain the user’s comprehensive correlation. The calculation formula is shown in formula (8):

$$ST_{a,b} = \lambda Sim_{a,b} + (1 - \lambda) \cdot Trust_{a,b}$$ (8)

Among them, $\lambda$ is a balance parameter, $\lambda \in (0,1)$, and $\lambda$ is used to adjust the proportion of user similarity and user trust. According to different values of $\lambda$, different user comprehensive correlations can be obtained. $ST_{a,b}$ represents the comprehensive correlation value of user $a$ and user $b$, $Sim_{a,b}$ represents the similarity value of user $a$ and user $b$, and $Trust_{a,b}$ represents the trust value of user $a$ and user $b$.

According to the calculated comprehensive relevance of users, sort the comprehensive relevance values of the target user to other users from large to small, and select the $N$ users with the largest
comprehensive relevance value to the target user to form the nearest neighbor set \( N_a \), and then use the weighted average of \( N \) nearest neighbors' ratings of the item to predict the target user's rating of the item.

In order to minimize the error of the predicted score, the average weighting method is used to predict the target user's score. The calculation formula is shown in formula (9):

\[
P_{a,i} = \bar{R}_a + \frac{\sum_{b \in N_a} ST_{a,b} \cdot (R_{b,i} - \bar{R}_b)}{\sum_{b \in N_a} ST_{a,b}}
\]  

(9)

Among them, \( P_{a,i} \) is the predicted score of user \( a \) on item \( i \), \( R_{b,i} \) is the score of user \( b \) on item \( i \), \( \bar{R}_a \) is the average score of user \( a \), and \( \bar{R}_b \) is the average score of user \( b \).

Sort the user's predicted score values for the items from largest to smallest, and recommend the items with a value greater than a specified threshold or the top \( N \) items with the highest scores to the target user.

![Flow chart of collaborative filtering recommendation algorithm combining user similarity and trust.](image)

Figure 1. Flow chart of collaborative filtering recommendation algorithm combining user similarity and trust.

5. Experimental results and analysis

5.1. Experimental data

This study uses the Movielens data set provided by GroupLens [1 2]. The data set used in this article contains 1 million ratings of 3,952 movies from 6040 users. This data set is a rating data set. Users can rate movies with 5 different levels of scores (1~5). Each user will rate at least 20 movies, and the data sparsity is 95.81%.
5.2. Evaluation Criteria

The average absolute error (MAE) is used as the evaluation standard. MAE is the average absolute deviation between the user’s predicted score and the actual score. The smaller the value of MAE, the higher the recommendation accuracy and the better the recommendation effect. The calculation formula is shown in formula (10):

\[
MAE = \frac{\sum_{a,i} |R_{a,i} - P_{a,i}|}{n}
\]  

(10)

Among them, \(R_{a,i}\) is the actual score of user \(a\) on item \(i\), \(P_{a,i}\) is the predicted score calculated by recommendation, and \(n\) is the number of items in the data set.

5.3. Experimental results

When calculating the trust between users, the trust value is affected by the trust threshold \(\alpha\) and the maximum trust propagation distance \(d_{max}\). Therefore, experiments are performed on these two parameters to find the best combination of parameters.

First of all, an experiment is performed on the confidence threshold \(\alpha\), and the experimental results are shown in Figure 3. Sim&Trust in the figure represents a recommendation method that combines user similarity with trust, and Trust represents a recommendation method that only uses trust. The experimental result is obtained under the premise that \(d_{max}=3\) and the number of nearest neighbors is 30. It can be seen from Figure 3 that when the value of \(\alpha\) is between 0 and 0.15, the value of MAE decreases. When the value of \(\alpha\) is 0.15, the MAE value obtained by the recommendation method that combines similarity and trust is the smallest. When the value of \(\alpha\) is greater than 0.15, the value of MAE increases. Therefore, the value of \(\alpha\) can be set to 0.15.

Figure 2. A statistical graph of the number of people corresponding to the score.
Experiments on the maximum trust propagation distance $d_{\text{max}}$, the experimental results are shown in Figure 4. The experimental results are obtained under the premise that $\alpha=0$ and the number of nearest neighbors is 30. It can be seen from Figure 4 that when the value of $d_{\text{max}}$ is 3, the value of $\text{MAE}$ is the smallest.

After determining the value of the trust degree threshold $\alpha$ and the maximum trust propagation distance $d_{\text{max}}$, the weight factor $\lambda$ in formula (8) is tested, and the number of nearest neighbors is 30. Figure 5 compares the corresponding values of $\text{MAE}$ when the value of $\lambda$ is between 0 and 1.

Figure 3. MAE values corresponding to different $\alpha$.

Figure 4. MAE values corresponding to different $d_{\text{max}}$.

Figure 5. MAE values corresponding to different $\lambda$. 

It can be seen from Figure 5 that when the value of $\lambda$ is 0.4, that is, the similarity algorithm occupies a weight of 0.4, and the trust relationship occupies a weight of 0.6, the value of $MAE$ is the smallest, and the recommendation effect is the best.

After determining the above parameters, in order to verify the collaborative filtering recommendation algorithm that combines similarity and trust, the experiment selected three methods of calculating similarity of traditional collaborative filtering algorithms to compare their $MAE$ values. The algorithm includes cosine similarity (Cosin), Jaccard similarity (Jaccard) and Pearson similarity (Pearson). The experimental comparison results are shown in Figure 6.

![Figure 6. Comparison between different algorithms.](image)

It can be seen from Figure 6 that the $MAE$ value of these algorithms gradually decreases with the increase of the number of nearest neighbors, and tends to be stable when the number of nearest neighbors reaches 100, indicating that the relative maximum value will be achieved when the number of nearest neighbors is 100. Good recommendation result. It can also be seen from Figure 6 that the MAE value of the algorithm proposed in this study is significantly lower than that of the traditional method, indicating that the algorithm proposed in this study can effectively improve the accuracy of the algorithm and produce more accurate recommendation results.

6. Conclusion
This study uses improved cosine similarity to measure user preference, and divides the trust relationship between users into direct trust relationship and indirect trust relationship according to whether users have a common rating item, and then the user similarity and user trust are weighted and fused to obtain the user's comprehensive correlation, and finally the neighbor set with the highest comprehensive correlation with the target user is found. The experimental results verify the superiority of the method proposed in this study and greatly improve the high efficiency and accuracy of the recommender system. This study only uses the user's rating information, and does not consider contextual information such as time and other user attribute information, so in the next step, we will consider introducing contextual information to improve the recommendation results.

Acknowledgments
This research was supported by the National Natural Science Foundation of China [32060234], the Fundamental Research Funds for the Central Universities [31920210084].

References
[1] CNNIC. (2021) CNNIC releases the 47th "Statistical Report on China's Internet Development
[2] Feng, S., Meng, J., Zhang, J. (2021) News Recommendation Systems in the Era of Information Overload. J. WEB ENG, 20: 459–470.

[3] Gunaratne, C., Rand, W., Garibay, I. (2021) Inferring mechanisms of response prioritization on social media under information overload. Sci. Rep., 11: 1–12.

[4] Guan, Q. (2021) Information Overload and Space Governance in Digital Survival. Journal of Kunming University of Science and Technology (Social Sciences), 21: 99–107.

[5] Zhao, W., Zhang, S. (2021) Collaborative filtering algorithm based on user preference in sparse data. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), http://kns.cnki.net/kcms/detail/50.1181.N.20210316.1900.064.html.

[6] Qin, J., Zhang, Q., Wang, B. (2020) Recommendation method with focus on long tail items. Journal of Computer Applications, 40: 454–458.

[7] Ding, J., Shen, S., Jia, L., et al. (2020) Recommendation algorithm based on collaborative filtering and hybrid similarity model. J. Univ. Shanghai for Sci. Technol., 42:275–282.

[8] Chen, J., He, H., Pan, Y., et al. (2020) Collaborative filtering algorithm based on classification model and SVD. Electronic Measurement Technology, 43: 69–73.

[9] Zhang, X., Chen, X., Seng, D. (2020) Top-N recommendation algorithm combining user trust and influence[J]. Journal of Zhejiang University (Engineering Science), 2020, 54(02): 311–319.

[10] Tong, S., Sun, B., Chu, X., et al. (2021) Trust recommendation mechanism-based consensus model for Pawlak conflict analysis decision making. International Journal of Approximate Reasoning, 135: 91–109.

[11] Zou, B., Li, C., Tan, L., et al. (2014) Social Recommendations Based on User Trust and Tensor Factorization. Journal of Software, 25: 2852–2864.

[12] GroupLens. (2003) MovieLens 1M Dataset. https://grouplens.org/datasets/movielens/.