Recommendation Algorithm of Crowdfunding Platform Based on Collaborative Filtering

Bixi Wang¹, Wenbin Wu¹, Wenfeng Zheng², Yan Liu² and Lirong Yin*

¹ Xi'an University No. 1 Science and Technology Road, Xi'an, 710065 P.R.C
² University of Electronic Science and Technology of China, Chengdu, 611731 P.R.C
³ Louisiana State University, Baton Rouge, 70803 United States
* Corresponding Author Email: lyin5@lsu.edu

Abstract. With the continuous progress of Internet technology, the crowdfunding platform has become a new way of network financing. While the generated data keeps increasing, its benefit does not increase in a proportional way, resulting in the "information overload" phenomenon. The personalized recommendation system can solve this problem by mining users' interests and preferences from a large amount of data. It has achieved success in many fields. This paper applies machine learning algorithm to build a recommendation system based on collaborative filtering. The designed personalized recommendation algorithm can provide accurate and rapid personalized recommendation services, which is convenient for users and conducive to the development of the crowdfunding platform. In addition, this paper uses the data from the crowdfunding platform in practice to complete the performance verification of the algorithm.

1. Introduction

Nowadays, with the rapid development of network technology, the crowdfunding platform is a typical representative of Internet financing. In recent years, it has become more perfect and widespread and is gradually becoming the supplement and enhancement of traditional financing methods (Thomas, Derrick, and Gillette 2004).

The data scale generated by crowdfunding grows rapidly, but the benefits generated based on data do not keep pace with the growth rate. This lag is caused by the lack of efficient methods to obtain value from massive data (Preis et al. 2012; Ding et al. 2019; Ni et al. 2019; Chen et al. 2020), which is called "Information Overload". To solve this difficulty, people design and put forward Personalized Recommendation Systems (Resnick and Varian 1997).

Personalized recommendation service has developed into an indispensable new service model for major websites. Goldberg developed the first recommendation system, Tapestry (Goldberg et al. 1992), in 1992. The system filters e-mail through the similarity of preferences of users. The key to the successful application of personalized recommendation system is the selected algorithm, which is the most important part of the system (Jianguo, Tao, and Binghong 2009; Zheng et al. 2016). Popular recommendations include content-based recommendations (Cantador, Bellogín, and Vallet 2010), Association Rules recommendations (Changchien and Lu 2001; Zheng et al. 2015), and Collaborative Filtering recommendations (Huang, Gao, and Gu 2013; Li et al. 2015). Among them, the most common and mature algorithm is the collaborative filtering algorithm.

The collaborative filtering algorithm completes the recommendation service for users according to the interests and preferences of similar users (Feng, Xian, and Feng 2004; Yin et al. 2019; Zheng et al. 2017). The advantage of collaborative filtering algorithm is that it can mine the interests and preferences and is not limited by item text information and classification information. Moreover, the
collaborative filtering algorithm will automatically obtain and complete the analysis without users give information. The idea of this technique was first proposed in 1992 and the establishment of Netflix Prize (Koren 2009) has made collaborative filtering technology more important and popular in the field of machine learning.

With the increasing demand of society, autonomous recommendation service based on collaborative filtering appears. The Group Lens system (Resnick et al. 1994) is a typical example of such a service. The system starts with a data model that requires users to provide their preferences for the project. The system analyzes the provide model and obtains the similarity between different users. Then, the traditional collaborative filtering algorithm is used to complete the recommendation service.

This paper is to complete the specific design and implementation of personalized recommendation algorithm on the crowdfunding platform. We then conduct experimental tests on the algorithm, verify the feasibility of the algorithm, and find out its problems using user rating data.

2. Dataset

The algorithm adopted in this paper is based on the user-based collaborative filtering algorithm, which needs to obtain user list, item list and user rating data of items.

Up to now, the crowdfunding data has not been publicly provided on the Internet for this study, so the data needed for the experiment need to be extracted by ourselves. Through browsing a number of crowdfunding platforms, only the roll-call time website provides user rating information. Therefore, this paper will obtain the user and project data on the roll call time platform.

After completion, the call time website has a total of 1095 crowdfunding projects, which are divided into 9 categories by the platform, namely film and television, advertising and exhibition, communications and digital, home life, smart wear, video and audio entertainment, travel positioning, culture and art, and food and beverage culture. There were 277 user-rated items, with a total of 4,645 user ratings and 3,445 users participating in the ratings. Users give scores to projects from different directions from the perspectives of innovation, design and utility, and finally take the average value.

For user support records, this article leverages the HttpClient toolkit, and these processes can all be handled programmatically. Finally, there are 1084 projects supported by users on the website of the roll call, with a total number of 172,841 person-times, 88,153 users participating in the support, and a total number of 89,427 people available on the platform.

3. Methods and Experiments

The input data of the algorithm is the user's rating data of the project, as shown in Table 1.

Table 1. User - project score data

| User ID | Item ID | Item_1 | Item_2 | ... | Item_n |
|---------|---------|--------|--------|-----|--------|
| User_1  | r_{1,1} | r_{1,2} | ...    | r_{1,n} |
| User_2  | r_{2,1} | r_{2,2} | ...    | r_{2,n} |
| ...     | ...     | ...     | ...    | ...   |
| User_m  | r_{m,1} | r_{m,2} | ...    | r_{m,n} |

The rows represent the project, the columns represent the user, and the matrix value pair corresponds to the user's score value for the project. The matrix size is users and items, as shown in Equation (1).

\[
R = \begin{bmatrix}
    r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\
    r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{m,1} & r_{m,2} & \cdots & r_{m,n} 
\end{bmatrix}
\] (1)
If two users have scored records on the same item and the scores are similar, then the two users are considered to have similar interests and preferences, so as to complete the recommendation. The nearest neighbor query needs to compare the user with all other users for user similarities. This similarity will directly affect the accuracy of the recommendation algorithm, so the calculation method of similarity is also a key research problem.

According to the comparison of similarity, Pearson correlation coefficient has the best effect in the user-based collaborative filtering algorithm. This method is adopted in this paper to obtain the similarity matrix between users. Then the correlation similarity between user \( u \) and user \( v \) is calculated as shown in Equation (2).

\[
sim(u, v) = \frac{\sum_{i \in I_u} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_v} (r_{vi} - \bar{r}_v)^2}}
\]

(2)

The rating matrix \( R \) and the user similarity matrix \( SIM \) are used to predict the user's rating of the recommended items. Rank the prediction scores from high to low, and select the top \( N \) items to recommend to users. The following introduces the generation prediction score.

In the similarity matrix, \( K \) nearest neighbor users are selected from user \( u \) to form the nearest neighbor set \( KNB_u \). Then, according to the similarity of users in \( u \) and \( KNB_u \) and the actual score value of the user in \( KNB_u \) on the recommended items, the predicted score \( \hat{r}_{u,i} \) of user \( u \) on item \( i \) can be estimated, as shown in Equation (4).

\[
\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{n \in KNB_u} sim(u, n) \times (r_{n,i} - \bar{r}_n)}{\sum_{n \in KNB_u} (\mid sim(u, n) \mid)}
\]

(4)

Experiment 1: The user score data of the crowdfunding platform was divided into two parts: the training set and the test set. To reduce the experimental error, the ratio of training set to test set is adjusted by continuous testing. The user-based collaborative filtering algorithm is adopted to test the designed algorithm. To avoiding the influence of external factors, five experiments were conducted on each parameter, as shown in Table 2.
Table 2. Collaborative filtering algorithm MAE values with different proportions of training set and test set

| Training set: Test set | 1:9 | 2:8 | 3:7 | 4:6 | 5:5 | 6:4 |
|------------------------|-----|-----|-----|-----|-----|-----|
| MAE                    | 1.26946 | 1.24384 | 1.22606 | 1.20977 | 1.19941 | 1.18779 |

| Training set: Test set | 7:3 | 8:2 | 9:1 | 19:1 | 99:1 | 190:1 |
|------------------------|-----|-----|-----|------|-----|------|
| MAE                    | 1.17631 | 1.17386 | 1.16268 | 1.15202 | 1.15749 | 1.128397 |

Table 3. Values of collaborative filtering algorithm with different number of adjacent neighbors

| K  | 5     | 15    | 25    | 35    | 45    |
|----|-------|-------|-------|-------|-------|
| MAE| 1.16722 | 1.16355 | 1.15671 | 1.15762 | 1.15545 |
| K  | 55    | 65    | 75    | 85    | 95    |
| MAE| 1.15699 | 1.15405 | 1.15213 | 1.15389 | 1.15354 |

Table 4. Classification of items to be predicted

| System advice | System is not recommended |
|---------------|---------------------------|
| Users love    | \( N_{tp} \) (True – Positive) | \( N_{fn} \) (False – Negative) |
| Users don't like | \( N_{fp} \) (False – Positive) | \( N_{m} \) (True – Negative) |
The raw data is divided into training set and test set. The main evaluation methods are precision rate and recall rate. The recommended accuracy rate (Precision) is calculated as Equation (7). The proportion of the correct number in the total number of recommendations is the recommended accuracy rate. The recommended recall rate (Recall) is calculated as the percentage of the number of items recommended by the system in the total number of items favored by the user in the recommendation list, as shown in Equation (8).

$$\begin{align*}
    \text{Precision} &= \frac{N_{tp}}{N_{tp} + N_{fp}} \\
    \text{Recall} &= \frac{N_{tp}}{N_{tp} + N_{fn}}
\end{align*}$$

Equation (7)

Equation (8)

Through the calculation of the above two indicators, it is not difficult to see that the two do not increase or decrease at the same time, they tend to change in the opposite direction. To make both the precision rate and the recall rate good, it is necessary to find an indicator that comprehensively balances the changes of the two. The commonly used index is F (F-measure), which is calculated as shown in Equation (9).

$$F\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Equation (9)

F-measure more fully represents the performance of the system. The bigger F-measure is, the better the recommendation effect will be, and the better the recommendation will be able to meet the needs of users.

![Figure 1](image1.png)

**Figure 1.** Influence of the ratio of training set and test set on the system

![Figure 2](image2.png)

**Figure 2.** Collaborative filtering algorithm value changes with the number of neighbors

The change curve in experiment 1 along with the proportion of the training set and the test set is represented on the image, and the change trend diagram can be obtained, as shown in Figure 1.

It can be seen from the above curve that the amount of training data has a significant impact on the system performance. Training-testing ratio increase from 1:9 to 19:1 range while MAE gradually narrowed. When the rate is between 19:1 and 99:1 and MAE has increased dramatically, the test set has too little data to be representative. Therefore, the ratio of training set to test set is 19:1. The K and the value of experiment 2 is taken as the coordinate and change curve is shown in Figure 2.
### Table 5. Evaluation indexes of collaborative filtering algorithm with different K

| K   | 5     | 15    | 25    | 35    | 45    |
|-----|-------|-------|-------|-------|-------|
| Precision | 0.63667 | 0.63952 | 0.64333 | 0.64542 | 0.64544 |
| Recall   | 1.00000 | 0.98528 | 0.98266 | 0.97431 | 0.96920 |
| F-measure | 0.77786 | 0.77549 | 0.77747 | 0.77641 | 0.77480 |

| K   | 5     | 65    | 75    | 85    | 95    |
|-----|-------|-------|-------|-------|-------|
| Precision | 0.64789 | 0.64791 | 0.64964 | 0.64713 | 0.64570 |
| Recall   | 0.96472 | 0.96054 | 0.96054 | 0.95931 | 0.95331 |
| F-measure | 0.77507 | 0.77370 | 0.77492 | 0.77276 | 0.76983 |

A user score above 3.5 indicates that the user is interested in the project. The accuracy rate and recall rate of the algorithm are combined to get the F index, as shown in Table 5. The change of K is represented in the coordinate system, as shown in Figure 3.

According to the test, it can be found that the algorithm has the following characteristics: The average absolute error of the algorithm based on collaborative filtering applied to the crowdfunding platform varies within the range of 1.15-1.17. The prediction accuracy rate is greater than 0.6. The influence of the change of the nearest neighbor number on the algorithm performance is not clear. According to these characteristics, the following conclusions can be drawn: The collaborative filtering algorithm designed in this chapter is feasible on the crowdfunding platform. The prediction error of the algorithm is large, which is mainly caused by the sparsity of data matrix and the cold start problem.

### 5. Conclusion

First of all, this paper uses Java algorithm to obtain the data required by the experiment on the crowdfunding platform, including user list, project list and user rating data of the project. Furthermore, these data are preprocessed for the experiment.

In this paper, a user-based collaborative filtering algorithm is designed by combining the nearest neighbor algorithm in machine learning. According to the experiment, it can be found that: When the ratio of training set to test set is 19:1, MAE reaches the minimum value. At this point, the test of the algorithm can achieve better results. In the k-mean algorithm used in the method, the change of K has no great influence on the accuracy rate. The average absolute error obtained by this algorithm on the crowdfunding platform varies within the range of 1.15-1.17, achieving good results.

The algorithm can recommend projects that the users of the crowdfunding platform are interested in, and the experiment proves that the prediction accuracy rate of the recommendation reaches over 60%, but there are also problems such as large algorithm prediction error, which are mainly caused by sparse user score data matrix and cold start of the system, so the algorithm needs to be further improved.

### 6. Reference

[1] Cantador, Iván, Alejandro Bellogín, and David Vallet. 2010. "Content-based recommendation in social tagging systems." In Proceedings of the fourth ACM conference on Recommender systems, 237-40.

[2] Changchien, S Wesley, and Tzu-Chuen Lu. 2001. 'Mining association rules procedure to support on-line recommendation by customers and products fragmentation', Expert Systems with Applications, 20: 325-35.

[3] Chen, Xiaobing, Lirong Yin, Yulin Fan, Lihong Song, Tingting Ji, Yan Liu, Jiawei Tian, and Wenfeng Zheng. 2020. 'Temporal evolution characteristics of PM2.5 concentration based on continuous wavelet transform', Science of The Total Environment, 699: 134244.
[4] Ding, Yueming, Xia Tian, Lirong Yin, Xiaobing Chen, Shan Liu, Bo Yang, and Wenfeng Zheng. 2019. "Multi-scale Relation Network for Few-Shot Learning Based on Meta-learning." In *International Conference on Computer Vision Systems*, 343-52. Springer.

[5] Feng, ZHOU Jun, TANG Xian, and G Jing Feng. 2004. 'An optimized collaborative filtering recommendation algorithm', *Journal of Computer Research and Development*, 10: 034.

[6] Goldberg, David, David Nichols, Brian M Oki, and Douglas Terry. 1992. 'Using collaborative filtering to weave an information tapestry', *Communications of the ACM*, 35: 61-70.

[7] Huang, Yue, Xuedong Gao, and Shujuan Gu. 2013. 'UARR: a novel similarity measure for collaborative filtering recommendation', *Cybernetics and Information Technologies*, 13: 122-30.

[8] Jianguo, Liu, Zhou Tao, and Wang Binhong. 2009. 'Research progress of personalized recommendation system', *Progress in Natural Science*, 19: 1-15.

[9] Koren, Yehuda. 2009. "Collaborative filtering with temporal dynamics." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 447-56.

[10] Li, Xiaolu, Wenfeng Zheng, Dan Wang, Lirong Yin, and Yali Wang. 2015. 'Predicting seismicity trend in southwest of China based on wavelet analysis', *International Journal of Wavelets, Multiresolution and Information Processing*, 13: 1550011.

[11] Ni, Xubin, Lirong Yin, Xiaobing Chen, Shan Liu, Bo Yang, and Wenfeng Zheng. 2019. "Semantic representation for visual reasoning." In *MATEC Web of Conferences*, 02006. EDP Sciences.

[12] Preis, Tobias, Helen Susannah Moat, H Eugene Stanley, and Steven R Bishop. 2012. 'Quantifying the advantage of looking forward', *Scientific reports*, 2: 1-2.

[13] Resnick, Paul, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. "GroupLens: an open architecture for collaborative filtering of netnews." In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 175-86.

[14] Resnick, Paul, and Hal R Varian. 1997. 'Recommender systems', *Communications of the ACM*, 40: 56-58.

[15] Thomas, Joshua M, Timothy R Derrick, and Jason C Gillette. 2004. 'The Effect of Knee Angle at Contact on Impacts While Running off a Raised Platform', *Medicine & Science in Sports & Exercise*, 36: S57.

[16] Yin, Lirong, Xiaolu Li, Wenfeng Zheng, Zhengtong Yin, Lihong Song, Lijun Ge, and Qingchuan Zeng. 2019. 'Fractal dimension analysis for seismicity spatial and temporal distribution in the circum-Pacific seismic belt', *Journal of Earth System Science*, 128: 22.

[17] Zheng, Wenfeng, Xiaolu Li, Jinxin Xie, Lirong Yin, and Yali Wang. 2015. 'Impact of human activities on haze in Beijing based on grey relational analysis', *Rendiconti Lincei*, 26: 187-92.

[18] Zheng, Wenfeng, Xiaolu Li, Lirong Yin, and Yali Wang. 2016. 'The retrieved urban LST in Beijing based on TM, HJ-1B and MODIS', *Arabian Journal for Science and Engineering*, 41: 2325-32.

[19] Zheng, Wenfeng, Xiaolu Li, Lirong Yin, Zhengtong Yin, Bo Yang, Shan Liu, Lihong Song, Yu Zhou, and Yanhong Li. 2017. 'Wavelet analysis of the temporal-spatial distribution in the Eurasia seismic belt', *International Journal of Wavelets, Multiresolution and Information Processing*, 15: 1750018.