ISSN: 2710-0588

Research on Hybrid Recommendation Algorithm based on Collaborative Filtering and Spearman Rank Correlation Coefficient

Shitong Zhang*, Hua Yang, Danyang Liu

Information Engineering College, Henan University of Science and Technology, Luoyang 471000, China

*Corresponding Author

Abstract

A recommendation system is an information filtering tool that helps users find the products or services they need from a large amount of information. However, collaborative filtering is quite sensitive to data sparsity and cold start problems, and it may encounter certain difficulties when handling outliers. To address the issue of handling outliers, it is necessary to study and improve the existing collaborative filtering techniques. This paper proposes a personalized recommendation algorithm that integrates collaborative filtering with the Spearman rank correlation coefficient. By combining collaborative filtering and the Spearman rank correlation coefficient, the algorithm uses the latter to handle outliers, making it more suitable for nonlinear relationships. This hybrid recommendation algorithm can better handle outliers while maintaining personalized recommendations, providing a basis and reference for personalized recommendations.

Keywords

Pearson Correlation Coefficient; Spearman Rank Correlation Coefficient; Collaborative Filtering; Improved Hybrid Recommendation Model.

1. Introduction

With the rapid development of information technologies such as the Internet, the Internet of Things (IoT), and cloud computing, these technologies have increasingly intertwined with various aspects of the human world, giving rise to the era of big data. In this era, personalized recommendation systems play an indispensable role by providing users with personalized and accurate recommendation information. Collaborative filtering algorithms, as a classic method in recommendation systems, can effectively achieve personalized recommendations by analyzing users' historical behaviors and the relationships between similar users. However, collaborative filtering still faces several challenges in handling data sparsity [1] and outliers. To address these challenges, increasing attention is being given to combining collaborative filtering with other algorithms to enhance the performance of recommendation systems. The Spearman rank correlation coefficient, as a non-parametric statistical method, is more reasonable for handling outliers and performs well in dealing with extreme cases in user rating data. This study aims to explore the integration of collaborative filtering with the Spearman rank correlation coefficient to develop a more accurate and reasonable personalized recommendation algorithm. By combining collaborative filtering with the Spearman rank correlation coefficient, we hope to retain the personalized recommendation advantages of collaborative filtering while improving its ability to handle outliers. This paper will analyze the design principles and implementation methods of this hybrid recommendation algorithm in detail and validate its performance across different scenarios through empirical research.

ISSN: 2710-0588

Through this study, we aim to provide useful insights and methods for further development in the field of recommendation systems.

2. Related Research

2.1. Collaborative Filtering Methods

Collaborative filtering methods[2] can help users mine useful information from massive data and provide personalized recommendations, currently widely used in e-commerce, movies, music, social networks, and other fields. This algorithm uses users' historical rating data to calculate the similarity between users, finds the N most similar users to the target user Q based on the similarity, and evaluates the data values of the unrated items of the target user Q through these N most similar users to determine whether to recommend them to the user. For example, establishing a collaborative filtering book recommendation method[3] to predict books of interest to user test1. If user test1 has not rated any books, the top K books with the highest collections will be recommended to user test1, see Figure 1. The similarity between users is calculated using the Pearson correlation coefficient.

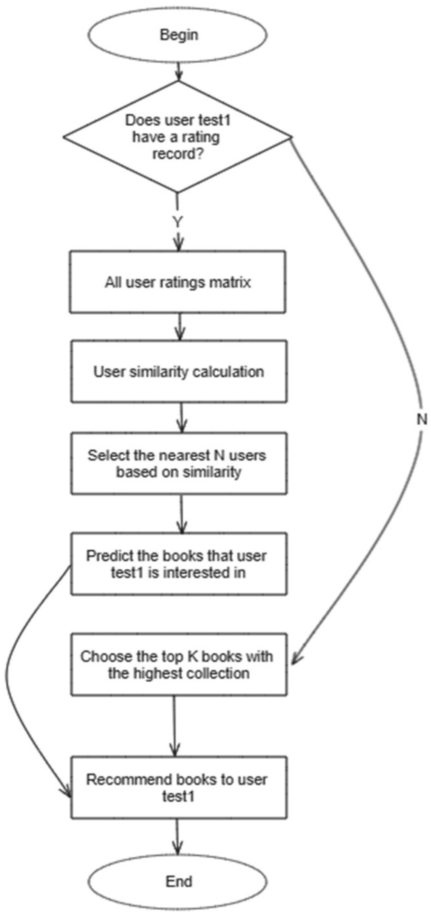

Figure 1. Flowchart of Collaborative Filtering Method

2.2. Collaborative Filtering Recommendation Method Combined with Spearman Rank Correlation Coefficient

Due to the instability of subjective user ratings, there are many extreme values that are too high or too low. Traditional collaborative filtering methods cannot solve these outlier problems, leading to reduced rationality in recommended products. Therefore, this paper proposes a collaborative filtering method that combines Spearman rank correlation coefficient for

ISSN: 2710-0588

personalized product recommendation, i.e., obtaining the rating arrays of each user and the target user, calculating the similarity using both Spearman rank correlation coefficient and Pearson correlation coefficient, and then calculating the weighted average of the two similarities. The weighted average result is used as the final similarity. Similarly, the N users closest to the target user are obtained, and other products with higher ratings from these N users are recommended to the target user. Finally, the personalized products recommended to the target user are obtained. The process of calculating similarity in the collaborative filtering method combined with Spearman rank correlation coefficient is shown in Figure 2. In Figure 2, sim1 represents the similarity calculated using the Spearman rank correlation coefficient, sim2 represents the similarity calculated using the Pearson correlation coefficient, and combine_sim represents the weighted average of sim1 and sim2.

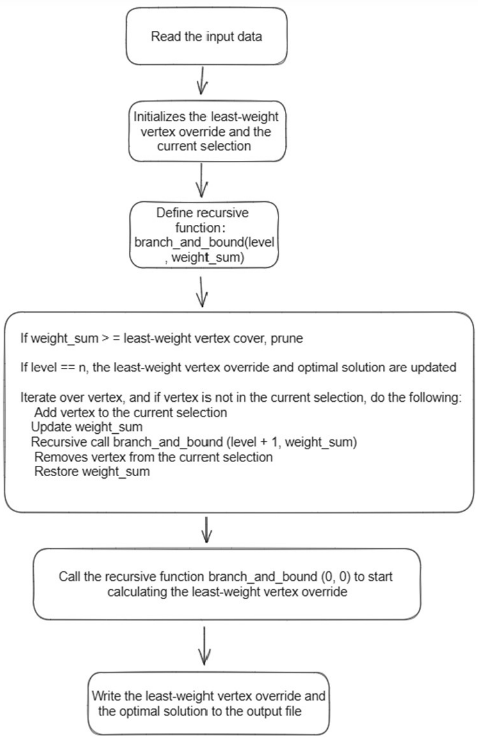

Figure 2. Similarity Calculation Process

2.2.1. Calculation of Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient is a method used to measure the non-linear correlation between two variables. It is based on the ranks of the two variables rather than their actual values. This paper adopts a user-based similarity calculation approach. Firstly, it inputs the arrays of ratings for the same items from both the target user and users who have rated the same items as the target user. For two sets of arrays, the data is sorted based on the magnitude of their values. Then, the ranks (positions) after sorting are assigned to replace the original values in the arrays. This process generates rank data for both sets of ratings. Corresponding to the rank data, the rank differences for each pair of data are calculated, which is the rank of the first variable minus the rank of the second variable. These rank differences are squared. The sum of the squares of all rank differences is then calculated and adjusted using a correction factor to obtain the Spearman rank correlation coefficient.

Spearman's rank correlation coefficient is calculated using the formula as shown in equation (1):

ISSN: 2710-0588

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}. \tag{1}$$

The Spearman rank correlation coefficient (ρ) is calculated using the formula provided, where d_i^2 represents the square of the rank differences for each pair of rating data,n is the number of rating data points, In the calculation, $\sum d_i^2$ is the sum of the squares of all rank differences,and $n(n^2-1)$is a correction factor to adjust for the sample size's influence. This formula indicates that the Spearman rank correlation coefficient is computed by normalizing the squares of the rank differences. Its values range between -1 and 1, where 1 represents a perfect positive correlation between two users, -1 represents a perfect negative correlation, and 0 indicates no correlation between the two users.

2.2.2. Calculation of Pearson Correlation Coefficient

The Pearson correlation coefficient[4] is a statistical measure used to assess the strength of a linear relationship between two variables. Firstly, the arrays of ratings for the same items from both the target user and users who have rated the same items as the target user are inputted. The ratings arrays are then standardized by subtracting the mean and dividing by the standard deviation. For the standardized data, the covariance between the two variables is calculated. The covariance indicates how the trends of two variables change together. The formula for calculating covariance is shown in equation (2):

$$\mathrm{cov}(X, Y) = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}. \tag{2}$$

In the formula, X_i and Y_i represent the data values corresponding to the ratings, $\overline{X}$and $\overline{Y}$ are the means of X and Y respectively, and n is the sample size. Dividing the covariance by the product of the standard deviations yields the Pearson correlation coefficient. Similarly, the range of Pearson correlation coefficient, also known as Pearson similarity, is [-1, 1]. It reflects the correlation between two rating vectors. When the value is -1, it indicates a negative correlation between the two user rating vectors; when the value is 0, it indicates no correlation between the two users; and when the value is 1, it indicates a positive correlation between the two vectors. The formula for calculating the Pearson correlation coefficient is shown in equation (3):

$$P(C_i, C_j) = \frac{\sum_{k \epsilon T_{i,j}} \left(R_{C_i,k} - \overline{R}_{C_i}\right)\left(R_{C_j,k} - \overline{R}_{C_j}\right)}{\sqrt{\sum_{k \epsilon T_{i,j}} \left(R_{C_i,k} - \overline{R}_{C_i}\right)^2 \cdot \sum_{k \epsilon T_{i,j}} \left(R_{j,k} - \overline{R}_{C_j}\right)^2}}. \tag{3}$$

In the formula, $\boldsymbol{P(C_i, C_j)}$represents the common score Peel similarity of the user's $\boldsymbol{C_i}$ and the user's $\boldsymbol{C_j}$, $\boldsymbol{\overline{R}_{C_i}}$ represents the average rating of the user's $\boldsymbol{C_i}$, $\boldsymbol{\overline{R}_{C_j}}$ represents the average rating of the user's $\boldsymbol{C_j}$, and $\boldsymbol{k \epsilon T_{i,j}}$ represents the common items of the user's$\boldsymbol{C_i}$ and the user's $\boldsymbol{C_j}$, and the final user similarity matrix is obtained.

2.2.3. Item Prediction

Combining Spearman rank correlation coefficient and Pearson correlation coefficient to compute the similarity values for all users, then sorting the obtained similarity values in descending order. Selecting the top N users who are most similar to the current user as the set of users most similar to the current user, and finally providing personalized recommended items for the current user. The formula for similarity calculation is shown in equation (4):

ISSN: 2710-0588

$$\mathbf{r}_{\omega} = \boldsymbol{\omega}_1 \mathbf{r}_{\mathbf{Spearman}} + \boldsymbol{\omega}_2 \mathbf{r}_{\mathbf{Pearson}}. \tag{4}$$

r_{ω}represents the weighted average correlation coefficient, $\omega_1$and $\omega_2$are the weights for the two correlatio $\omega_1 + \omega_2$=1, and $\omega_1, \omega_2 \geq 0$.

3. System Functionality Validation

3.1. Instance Validation Functionality

To validate the system functionality, actual rating data from 500 students at a domestic university on a certain online book platform is selected and applied to the designed recommendation model for experimental testing. Programming is done using Python language, and 32,000 rating records from the actual ratings of the 500 students are used as experimental data. For a specific target user, taking user K199 as an example, users with the same rating books as the target user are identified, along with each user's ratings for each book. Finally, a table of neighboring user ratings for the target user is compiled, see Table 1 (5 records of neighboring user ratings and the target user K199's rating records are randomly excerpted).

Table 1. Three Scheme comparing

Numble	"Journey to the West"	"Rickshaw Boy"	"One Hundred Years of Solitude"	"Ghosts Blowing Out the Light 2: Dragon Ridge Maze"
K099	3.3	4.0	2.9	4.5
K122	4.6	3.8	3.4	3.6
K413	4.0	4.9	3.6	4.3
K187	4.1	3.2	4.5	3.8
K332	3.5	4.5	3.4	4.3
K199	4.0	4.5	3.2	4.0

Using a hybrid recommendation algorithm that combines Spearman rank correlation coefficient and Pearson correlation coefficient, the similarity scores between each user and the target user K199 were calculated based on the obtained rating table. Taking the 5 neighboring users from Table 1 as an example, the calculated similarities are shown in Table 2

Table 2. Partial Similarity Scores of Neighboring Users to Users K199

Target User ID	Similar User ID	Similarity Score
	K122	0.877423
	K413	0.957117
K199	K187	0.856412
	K332	0.921455
	K099	0.812542

3.2. Evaluation Standard for Hybrid Recommendation Algorithm

Due to the sparsity of the rating matrix, it is challenging to assess the recommendation effectiveness using the recommendation hit rate. Therefore, this paper adopts Mean Absolute Error (MAE)[5] to measure the practical effectiveness of the hybrid recommendation algorithm. The calculation formula is shown in equation (5):

$$\boldsymbol{MAE} = \sum_{i\epsilon rec(u)} |\hat{\boldsymbol{r}}_{ui} - \boldsymbol{r}_{ui}| / |\boldsymbol{rec}(\boldsymbol{u})|. \tag{5}$$

ISSN: 2710-0588

Where rec(u) is the set of recommended books for user u, $\hat{r}_{ui}$ is the predicted rating for book ii by user u, r_{ui} is the actual rating for book i by user u.

3.3. Instance Validation Functionality

To validate the practical effectiveness of the algorithm model in recommending books, different numbers of nearest neighbors k and different numbers of recommendations N are set. k ranges from 5 to 90, and N is set to 10, 15, and 20. The rating records of 500 users are respectively inputted into the hybrid collaborative filtering algorithm model combining Spearman rank correlation coefficient and the collaborative filtering algorithm model. MAE values are calculated. The results are shown in Figures 3 and Figures 4.

From Figures 3 and 4, it can be observed that when the value of k is too small, the recommendation efficiency decreases, and when k is too large, the algorithm's computational burden increases. When k gradually increases between 10 and 50, the recommendation effectiveness significantly improves. However, after k exceeds 60, the recommendation effectiveness does not significantly improve further.Furthermore, when k is smaller, a smaller value of N leads to better recommendation effectiveness. This suggests that when the number of nearest neighbors is too small, the recommended book selection is limited and may not include most of the books that the target user likes. If too many recommendations are made in such cases, books that the user does not like may also be included.Therefore, it is recommended to choose k above 60 and N as 20. At this point, the error is minimized, and the recommendation effectiveness is maximized.

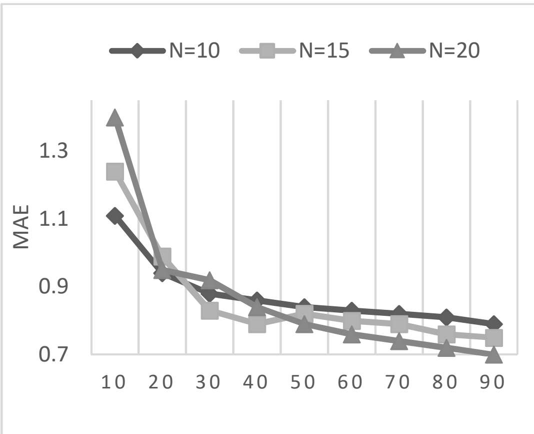

Figure 3. Comparison of MAE Values for Hybrid Algorithm

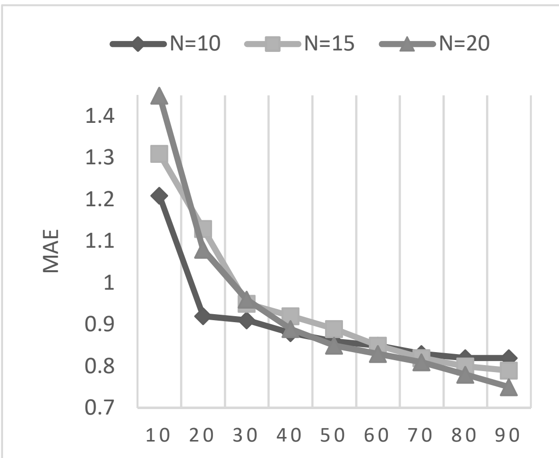

Figure 4. Comparison of MAE Values for Collaborative Filtering Algorithm

ISSN: 2710-0588

When the number N is 20, the MAE values of the collaborative filtering algorithm model incorporating the Spearman rank correlation coefficient and the MAE values of the model based on the collaborative filtering algorithm are shown in the Figure 5.

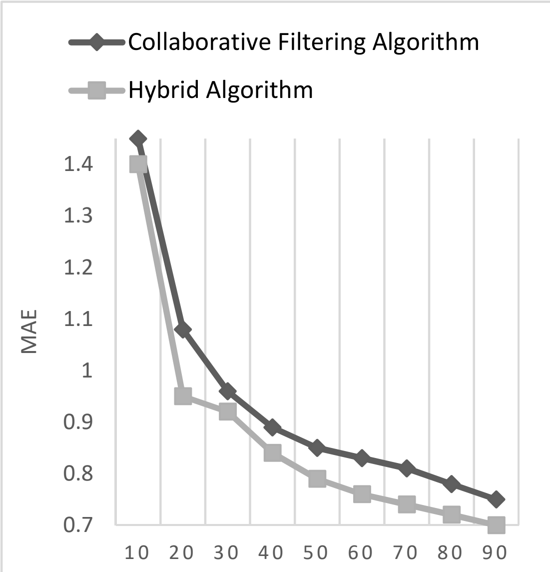

Figure 5. Comparison of MAE Values when N is 20

From Figure 5, it can be observed that the MAE value of the collaborative filtering algorithm model incorporating the Spearman rank correlation coefficient is lower compared to the MAE value of the model based solely on the collaborative filtering algorithm. This indicates a smaller error and relatively better recommendation performance.

4. Conclusion

This paper proposes a personalized recommendation algorithm based on collaborative filtering combined with the Spearman rank correlation coefficient. It overcomes the shortcomings of traditional collaborative filtering algorithms in handling outliers and can provide more accurate personalized recommendations. Through actual borrowing data, it is demonstrated that the recommendation effectiveness varies with changes in the number of neighbors and the number of recommendations, providing a basis and reference for product recommendations. However, in practical applications, it is still necessary to conduct experiments and evaluations based on more specific data characteristics and requirements to select the most suitable recommendation strategy. Continuous optimization of algorithm performance is also needed to meet user needs.

Acknowledgments

I would like to thank Ms. Bai Xiuling for her guidance in the selection of topics and the final review of the thesis, and thank Henan University of Science and Technology for the support of the Innovation and Entrepreneurship Undergraduate Training Program (No. 2023117) and the Key Science and Technology Program of Henan Province (No. 222102210177 232102211060 and No. 242102211077 242102210140).

ISSN: 2710-0588

References

[1] Tang Xinyu, Zhang Xinzheng, Liu Baoli. High-dimensional data recommendation system based on neural network and community discovery[J].Computer Applications and Software,2020,37(7):232-239.

[2] Shan Chunyu. Student achievement prediction method integrating time series and collaborative filtering[J].Information Science,2095-0691(2022)03-0069-06.

[3] Zhao Fengtao. Design of Bibliography Recommendation System for College Books Based on Collaborative Filtering Algorithm[J].Microcomputer Applications,1007-757X(2020)12-0067-03.

[4] WANG Xiaoli, SHI Gang, YANG Qingwen, et al. Research on the Relevance of Computer Science and Technology Curriculum System Based on Pearson Coefficient[J]. Wireless Internet Technology, 2019, 16(21):114-115. DOI:10.3969/j.issn.1672-6944.2019.21.053.

[5] Li Jiaojiao. Design and Implementation of University Library Management System Based on Workflow and Data Mining[D]. Jiangsu:Jiangsu University,2019(in Chinese).