Collaborative Filtering Recommendation Algorithm Based on User Attributes and Item Score

Chaohui Liu, Xianjin Kong, Xiang Li, and Tongxin Zhang

School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450046, China

Correspondence should be addressed to Xianjin Kong; kongxianjin9527@163.com

Received 29 June 2021; Accepted 23 February 2022; Published 22 March 2022

Academic Editor: Roberto Natella

Copyright © 2022 Chaohui Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To solve the problems of cold start and data sparseness existing in traditional collaborative filtering recommendation algorithm, a collaborative filtering recommendation algorithm based on user attributes and item scoring is proposed. Firstly, we improve the credibility of user similarity and explore the potential interests of users, a new user rating similarity calculation method is constructed by introducing confidence, item popularity, and Pearson weighting. Secondly, we construct a user attribute similarity measurement method by introducing cultural distance, age attribute similarity, and user label similarity. Finally, user rating similarity and user attribute similarity are weighted to form a new similarity measurement model. Through simulation comparison between the collaborative filtering recommendation algorithm and the traditional recommendation algorithm, our results show that the collaborative filtering recommendation algorithm can effectively improve the accuracy of recommendations and the diversity of results and effectively alleviate the problem of data sparseness.

1. Introduction

With the popularization of the Internet and the development of big data, the problem of information overload is becoming increasingly or increasing prominent. Resource recommendation, as an effective method of information filtering [1, 2], has high business value and research significance. Collaborative filtering recommendation algorithm [3, 4] is the most commonly used and successful method in the recommendation system. However, it still has a series of problems in the specific application process, such as cold start [5], data sparseness [6, 7], scalability [8], and other issues. Many researchers at home and abroad, who use different methods to improve the quality of recommendation in different degrees, have conducted in-depth and fruitful studies on data sparseness. Gao et al. [9] proposed two kinds of punishment algorithms for popular items to solve the problem of inaccurate simulation caused by popular items. Kaleli [10] used score similarity and score uncertainty difference information to search for the target users of their preference. Suryakant and Mahara [11] combined the cosine similarity, Jacquard coefficient, and other similarity calculation results for linear combination to improve the prediction accuracy. The above algorithm focuses on the user’s evaluation of items. Although it improves the recommendation accuracy to a certain extent, it neglects the influence of user attribute information on the recommendation results, which is mainly reflected in that different users with similar attributes may have similar interests and preferences.

Authors in [12] designed indicators such as similarity of interest tendency and confidence and proposed a recommendation algorithm based on user rating and attribute similarity, which alleviated the problem of data sparseness. Authors in [13] combined user attributes and confidence to calculate similarity, which alleviated the data sparseness problem to a certain extent. Authors in [14] proposed a collaborative filtering algorithm based on user attributes to solve the problems of traditional algorithm. The above studies ignored the influence of unpopular items on recommendation quality, and it was difficult to find out the potential interests of users.

To ease the problem of the algorithm, this paper proposes a recommendation algorithm called user attributes
and item score collaborative filtering recommendation algorithm (UAI-CF). Our main contributions can be summarized as follows:

1. The introduction of item popularity in the weighted Pearson correlation coefficient improves the recommendation probability of unpopular items, excavates the potential interest of users, and improves the confidence calculation method to improve the credibility of similarity among users and the recommendation accuracy.
2. The cultural distance is introduced to construct the user attribute model, and a collaborative filtering algorithm based on user attribute and item rating is proposed to effectively alleviate the data sparseness problem.
3. Through the simulation experiments on the Movielens dataset, compared with the traditional algorithm, it shows that the proposed algorithm can effectively improve the accuracy and coverage of recommendations.

2. Related Works

2.1. Classical Collaborative Filtering Algorithm. Classical collaborative filtering algorithms are divided into user-based methods, item-based methods, and model-based methods [15–17]. At present, user-based collaborative filtering has been widely studied, and the specific steps are divided into user modeling, finding the nearest neighbor, and prediction and recommendation.

2.1.1. User Modeling. Firstly, the user-item scoring matrix $R = [r_{ij}]_{m \times n}$ is established by using the user’s rating information for the item. $r_{ij}$ is the rating of user $i$ for the item $j$, and the size of the rating value is proportional to the degree of user’s interest preference. Then, according to the user’s attribute information, the user attribute matrix $A = [a_{uv}]_{m \times s}$ is established. The value of the matrix is 1 or 0, which represents whether the user has the attribute.

2.1.2. Find the Nearest Neighbor Set. Use the similarity calculation method to calculate the similarity between users, and generate the nearest neighbor user set with the highest similarity with the target user. There are three common similarity calculation methods: cosine similarity, modified cosine similarity, and Pearson similarity [18], of which Pearson similarity calculation formula is

$$\text{Sim}_p(i, j) = \frac{\sum_{x \in I_i} (r_{ix} - \bar{r}_i) \cdot (r_{jx} - \bar{r}_j)}{\sqrt{\sum_{x \in I_i} (r_{ix} - \bar{r}_i)^2} \cdot \sqrt{\sum_{x \in I_j} (r_{jx} - \bar{r}_j)^2}}, \quad (1)$$

where $i$ and $j$ represent users, $x$ represents items, $I_i$ represents the set of items rated jointly by user $i$ and $j$, $r_{ix}$ represents the rating of user $i$ on item $x$, and $\bar{r}_i$ is the average rating of user $i$ on all items.

2.1.3. Forecast Rating. According to the user similarity, the nearest neighbor set is constructed. Formula (2) is used to predict the target user’s rating towards the evaluation item and make recommendations:

$$P_{ix} = \bar{r}_i + \frac{\sum_{j \in N_i} \text{Sim}(i, j) \cdot (r_{jx} - \bar{r}_j)}{\sum_{j \in N_i} \text{Sim}(i, j)}, \quad (2)$$

where $P_{ix}$ represents the predicted score of user $i$ on item $x$, $N_i$ represents the set of the nearest neighbors of user $I$, and $\text{Sim}(i, j)$ represents the similarity between user $i$ and user $j$.

2.2. Relevant Concepts

2.2.1. Pearson Weighting Coefficient. The weighting coefficient is weighted according to Pearson correlation coefficient, and the measurement method is shown as

$$W(i, j) = \left(\frac{\sum_{x \in I_i} r_{ix} }{\sum_{x \in I_i} r_{iy} } \right) \cdot \left(\frac{\sum_{x \in I_j} r_{jx} }{\sum_{x \in I_j} r_{jy} } \right), \quad (3)$$

where $r_{ix}$ and $r_{jy}$, respectively, represent the ratings of users $i$ and $j$ to the jointly evaluated item $x$ and $y$, respectively, represent the ratings of users $i$ and $j$ to the respectively evaluated item $y$.

2.2.2. Confidence Degree. Confidence is used to measure the credibility degree of similarity between users, which is usually measured by Jaccard function, and the formula is as follows:

$$S(i, j) = \frac{|I_i \cap I_j|}{|I_i \cup I_j|}, \quad (4)$$

where $I_i$ represents the collection of items evaluated by user $I$, $|I_i \cap I_j|$ represents the number of items rated jointly by two users, and $|I_i \cup I_j|$ represents the number of unions of all evaluation items for two users, and the confidence ranges from 0 to 1.

2.2.3. Item Popularity. The popularity degree is used to measure the popularity degree of an item. A larger value indicates that the project is evaluated with fewer times and the target users may have more potential interest. The calculation formula is as follows:

$$C(x) = e^{-\lg N(x)}, \quad (5)$$

where $N(x)$ represents the number of items being evaluated, and the popularity of items is between 0 and 1.

2.2.4. Cultural Distance. Since geographic location information is a unique factor in resource recommendation, the culture is a clustered mental programming formed by groups in different regions, which is different from that of other groups. Cultural distance refers to the degree of similarity or difference between two cultures and is a relatively
standardized measurement method for quantifying cultural difference. Generally, cultural distance is used to represent the degree of cultural difference between countries [19], and the measurement method is shown as

$$CD(a, p) = \frac{\sum d_{aq} - d_{pq}}{V_q},$$

(6)

where $CD(a, p)$ represents the cultural distance between country $a$ and country $p$, $q$ represents the corresponding cultural dimension, $d_{aq}$ represents the value of country $a$ in the $q$th dimension, $V_q$ represents the variance in the $q$th dimension, and $n$ represents the number of cultural dimensions. According to Hofstede’s four cultural dimensions, power distance, uncertainty avoidance, individualism/collectivism, and masculinity/femininity, the theoretical value range of $CD$ is [0, 17.93], while the actual value range is [0.02, 8.22] [20].

3. An Approach Based on User Attributes and Item Ratings

3.1. Symbol Definition. In the recommendation system, it is assumed that there are $m$ users and $n$ items, and the user has at most $s$ attribute number. The user’s evaluation of the item is represented by the matrix $R_{m \times n}$, and the user’s attribute information is represented by the matrix $A_{m \times s}$. The user attribute $A$ matrix is analyzed to determine whether the user is a new user. If so, the similarity between users is calculated by using the user attribute similarity $\text{Sim}_A$; otherwise, the user-item rating $R$ matrix is analyzed and the similarity is calculated by using the Sim based on the user attribute and item scoring similarity. According to the similarity between users, the nearest neighbor set $N_i$ is obtained, and the value of the unrated items of target users $P_{ix}$ is predicted, and the final item recommendation scheme is obtained according to the predicted value.

For reading conveniently, the important symbols are explained in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Symbol definition.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbols</td>
</tr>
<tr>
<td>$R_{m \times n}$</td>
</tr>
<tr>
<td>$A_{m \times s}$</td>
</tr>
<tr>
<td>$\text{Sim}_A(i, j)$</td>
</tr>
<tr>
<td>$W(i, j)$</td>
</tr>
<tr>
<td>$S(i, j)$</td>
</tr>
<tr>
<td>$C(x)$</td>
</tr>
<tr>
<td>$CD = (a, p)$</td>
</tr>
<tr>
<td>$IS(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_CP(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_p(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_s(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_l(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_t(i, j)$</td>
</tr>
<tr>
<td>$\text{Sim}_a(i, j)$</td>
</tr>
<tr>
<td>$P_{ix}$</td>
</tr>
</tbody>
</table>

3.2. Similarity Calculation Based on Item Ratings. The popularity is introduced into the weighted Pearson similarity calculation to improve the influence of unpopular items on the recommendation results and fully tap the potential interest of users.

3.2.1. Confidence Calculation. Although the traditional confidence calculation formula (4) can measure the credibility of similarity between users, it has some errors. When $S(i, j) = S(i, k)$ and the items jointly rated are all 1 item, due to the difference in users’ rating habits, users are rated the same, but users and users have similar ratings on the item, the similarity of the same rating should be greater than the similarity of the similar rating, which is contradictory to the similarity of the two.

Considering that when the proportion of items jointly evaluated is the same, the greater the proportion of items with the same rating, the more credible the similarity between users. By introducing the ratio of the number of items with the same rating to the number of items with jointly rating to construct a new confidence measure function, the calculation formula is as follows:

$$IS(i, j) = S \times \frac{\sum_{i, Ij} \{F(\{r_{ix} = r_{jx}\})\}}{|I_i \cap I_j|},$$

(7)

where $F(\{ \ast \})$ is an indicator function: $[\{true, false\} \rightarrow \{1, 0\}]$ and $\sum_{i, Ij} F(\{r_{ix} = r_{jx}\})$ represents the number of the same item evaluated by user $i$ and user $j$ in the public item set.

3.2.2. Similarity Calculation Based on Item Popularity. Popular items have more user evaluation information, which cannot reflect the difference of users’ interests. On the contrary, unpopular items with less evaluation information can better reflect users’ interest preferences. Therefore, the influence of unpopular items on recommendation results should be improved. Formula (5) is introduced into the weighted Pearson similarity calculation formula (1) to obtain an improved similarity measurement formula as follows:

$$\text{Sim}_CP(i, j) = \frac{\sum_{i, Ij} (r_{ix} - \bar{r_i}) \cdot (r_{jx} - \bar{r_j}) \cdot C(x)}{\sqrt{\sum_{i, Ij} (r_{ix} - \bar{r_i})^2 \cdot \sum_{i, Ij} (r_{jx} - \bar{r_j})^2}} \times W(i, j),$$

(8)

where $W(i, j)$ is the Pearson weighting coefficient and $C(x)$ represents the popularity of an item. It can be seen that the larger the value, the greater the contribution of unpopular items to similarity, so the influence of unpopular items on the recommendation results can be improved.

3.2.3. The Score Similarity of Fusion Confidence and Popularity. According to the modified similarity calculation method and confidence, the user similarity measurement formula based on user rating is obtained:
3.3. Similarity Calculation of User Attributes. Users generally include age, location, gender, occupation, and other basic attribute information, which to a certain extent reflects the user’s interest preferences. Users with similar attributes may have similar interest preferences.

3.3.1. Age Similarity of Users. Users of different ages have similar interest preferences. Users with similar attributes may include age, location, gender, occupation, and other basic attributes. When the age gap exceeds a certain range, the impact of age on similarity becomes smaller [15].

The calculation formula of user age similarity is as follows:

\[ \text{Sim}_{\text{Age}}(i, j) = \frac{\alpha}{|a_i, \text{age} - a_j, \text{age}|} \cdot \begin{cases} \text{1}, & |a_i, \text{age} - a_j, \text{age}| \leq \alpha, \\ \text{0}, & |a_i, \text{age} - a_j, \text{age}| > \alpha, \end{cases} \]  

where \( a_i, \text{age} \) and \( a_j, \text{age} \) represent the age of the user and \( \alpha \) is the threshold value for dividing the age range. The specific value can be obtained from the experiment.

3.3.2. Location Similarity of Users. Users’ interests are affected by cultural groups and hierarchies. Users in different countries form unique groups due to different regions, races, religions, language groups, and other levels. Cultural distance can be used to measure the difference between users. The cultural distance of different user attributes is normalized, and the similarity function of user location attributes is obtained as follows:

\[ \text{Sim}_{\text{Loc}}(i, j) = 1 - \frac{\text{CD}(o_i, p_j) - \text{CD}_{\text{min}}}{\text{CD}_{\text{max}} - \text{CD}_{\text{min}}} \]  

where \( o_i \) represents the country of user \( i \) and \( \text{CD}_{\text{min}} \) and \( \text{CD}_{\text{max}} \), respectively, represent the minimum and maximum cultural distance in all data.

3.3.3. Tag Similarity of Users. Users of the same gender may have common interests; for example, males like to watch finance and economics while females like to watch soap operas. Users of the same profession generally have the same interests and hobbies, such as painters like paintings and collectors like collection. The similarity calculation formula for users’ tag attributes is as follows:

\[ \text{Sim}_{\text{Tag}}(i, j) = \frac{l}{s} \]  

where \( l \) represents the number of the same tag attributes of user \( i \) and \( j \) and \( s \) represents the total number of user tags.

Formula (13) is used to calculate the user’s overall attribute similarity according to the user’s age attribute, location attribute, and label attribute similarity:

\[ \text{Sim}_{\text{Attr}}(i, j) = \sum W_a \times \text{Sim}_{\text{Attr}}, \]  

where \( W_a \) represents the weight of the attribute similarity, with \( a \) value range of \([0,1]\), and \( \text{Sim}_{\text{Attr}} \) represents the attribute similarity.

3.4. Similarity Calculation Combining User Attributes and Item Ratings. The improved similarity calculation combines the similarity based on item rating and similarity based on user attributes. Firstly, the similarity based on item score was calculated. Secondly, the similarity based on user attributes is calculated. Then, the two similarity degrees are fused, and the calculation formula is as follows:

\[ \text{Sim}(i, j) = \beta \cdot \text{Sim}_{\text{Attr}}(i, j) + (1 - \beta) \cdot \text{Sim}_{\text{R}}(i, j) \]  

\[ \beta = 2 \times \left(1 - \frac{1}{1 + e^{-l_i}}\right) \]  

The algorithm can alleviate the problems of cold start and data sparseness. For new users, the similarity between users is calculated by using the user attribute matrix, which can effectively alleviate the cold start problem caused by new users entering the system. When the number of user rating data increases, user attributes are fused on the basis of user rating similarity to enhance the similarity between users, and the problem of low similarity between users caused by data sparseness is alleviated.

3.5. Similarity Calculation Combining User Attributes and Item Ratings. Algorithm 1 used in this study to calculate the scoring similarity has the same complexity as the traditional similarity algorithm (Pearson similarity), namely, each two user attributes needs 3 times weighting and 2 times addition. This method of calculation costs very little. When the similarity of user ratings is calculated, the weighted correlation coefficient calculation formula is multiplied by the improved confidence and popularity.

4. Experiment and Result Analysis

4.1. Dataset. The classic MovieLens dataset was used for the experiment, in which a dataset on movie evaluation is provided by GroupLens Research Group of the University of Minnesota. Among them, there are 100000 rating data of 1682 movies from 943 users, and the rating of movies is an integer between 1 and 5. The size of the rating value is proportional to the degree of users’ interest and preference, and the sparseness of the dataset is 93.7%. The ML-100K dataset provides 5 training sets and corresponding test sets, from which a pair of simulation experiments are randomly selected.
4.2. Evaluation Indicators. To better evaluate the algorithm, MAE, precision, recall, and coverage are used to measure the algorithm.

4.2.1. Average Absolute Error (MAE). Since the accuracy of prediction is inversely proportional to the size of the average absolute error (MAE), the following formula is used to measure the recommendation accuracy:

$$\text{MAE} = \frac{\sum_{i,x} |r_{ix} - \hat{p}_{ix}|}{|T|},$$ \hspace{1cm} (15)

where $T$ is the test set and $\hat{p}_{ix}$ is the predicted score of users to the item calculated by formula (2), which is the true score of users to the item.

4.2.2. Precision. Precision indicates how many of the recommended items are correctly recommended. A higher value means a higher accuracy of recommendation. The calculation formula is as follows:

$$\text{Precision} = \frac{\sum_u |R(u) \cap T(u)|}{\sum_u |R(u)|},$$ \hspace{1cm} (16)

where $R(u)$ is the recommendation items and $T(u)$ is the actual items in the test.

4.2.3. Recall. Recall indicates how many of the recommended items are predicted. The value indicates the accuracy of the recommended results. The calculation formula is as follows:

$$\text{Recall} = \frac{\sum_u |R(u) \cap T(u)|}{\sum_u |R(u)|},$$ \hspace{1cm} (17)

4.2.4. Coverage. Coverage describes the long tail taping ability of the recommendation system and whether the recommended items are effective to users. The larger the coverage value is, the more evenly the recommended items are distributed. The calculation formula is as follows:

$$\text{Coverage} = \frac{\sum_u |R(u)|}{n},$$ \hspace{1cm} (18)

where $R(u)$ is the set of items recommended by users and $n$ is the total number of items in the dataset.

4.3. Result Analysis. The threshold values are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, and the corresponding MAE values are calculated, respectively. On the ml-100 k dataset, when the value is 5, the value of MAE is the minimum. Therefore, the threshold value is set to 5.

The final similarity calculation method in this study was recorded as UAI-CF, which was compared with the traditional collaborative filtering algorithm (Pearson) and the collaborative filtering algorithm based on user features and similar confidence (UAS-CF) proposed in [15]. The three algorithms were compared from the four aspects of mean absolute error, accuracy, recall rate, and coverage rate to verify the recommendation quality of the algorithm.
In order to compare the proposed algorithm with the traditional collaborative filtering algorithm (Pearson similarity) and the differences in MAE of the proposed algorithm, the number of neighbors of the target user is taken as the variable, and the number of neighbors is increased from 5 to 50 with an interval of 5. The change of MAE values of the three algorithms is shown in Figure 1. It can be seen from Figure 1 that, with the increase of neighbor users, Pearson algorithm of MAE decreases gradually, while the UAS-CF’s and UAI-CF’s MAE increased first, and then into balance. This is due to that the target users do not like the recommended movies. When the number of neighbors increases to a certain number, the recommended movies are primarily determined by the users who are most similar to the target users. When the nearest neighbor user is the smallest, MAE takes the minimum value, indicating that the algorithm in this study can alleviate the cold start problem.

The changes in precision values of the three algorithms are shown in Figure 2. With the increase of nearby users, the accuracy of the three algorithms decreases gradually, and the UAR-CF algorithm is better than the other two algorithms.

Figure 3 shows the change of recall rate values of the three algorithms with the increase of the number of near neighbors. The recall rate of the three algorithms gradually increases with the increase of the number of near neighbors, and UAI-CF algorithm is better than the other two algorithms.

Figure 4 shows the coverage of the three algorithms increases with the increase of nearby users. This is because the increase of nearby users leads to the increase of movies to be recommended. Compared with [15] and traditional algorithms, the proposed algorithm in this study clearly shows that the coverage rate of UAI-CF algorithm increases faster, and it is superior to the other two algorithms in terms of long tail effect.

To sum up, the improved UAR-CF algorithm in this study is better than the comparison algorithm from the four aspects of mean absolute error, accuracy, recall rate, and coverage rate and is more practical.

5. Conclusion and Prospect

This study proposes a recommendation algorithm based on user attributes and item rating, which improves the traditional collaborative filtering algorithm. In the model of this study, by collecting user evaluation information on the item, considering item popularity and confidence among users, the similarity between users is calculated by integrating user
attributes. The obtained results validated the goodness of the proposal in terms of improved the recommendation accuracy to a certain extent and solved cold start and data sparseness on the recommendation results. As future work, we will try to integrate item attributes and introduce a more accurate method training model to further explore the relationship between users and items and improve the quality of recommendation.

**Data Availability**

The data used to support the findings of this study are available from the corresponding website: https://grouplens.org/datasets/movielens/100k/.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

This work was financially supported by the Key Scientific and Technological Projects in Henan Province (Grant no. 222102210079), the 13th Five Year Plan of Educational Science in Henan Province (Grant no.2020YB0149), National College Students’ innovation and Entrepreneurship Training Program (Grant no. 20210485023), and the Cooperative Education Project of Higher Education Department of Ministry of Education (Grant no.201902060014).

**References**


