Application of Improved Collaborative Filtering Algorithm in Recommendation of Batik Products of Miao Nationality

Ning Ding*, Jian Lv* and Lai Hu*

Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang, Guizhou, China

*Corresponding author e-mail: dingning19@163.com, *1505011295@qq.com, bLaifinial@163.com

Abstract. Recommendation system has been widely used in various e-commerce websites, but there are few personalized recommendations for ethnic handicrafts. And the national handicraft recommendation belongs to the small batch recommendation behavior, so the recommended performance is limited by the sparsity of the data in the scoring matrix. To solve this problem, an improved collaborative filtering algorithm is proposed, in which the user-based collaborative filtering algorithm is nested in the collaborative filtering algorithm based on gradient descent. First, the scoring prediction value is filled into the unscoring items in the scoring matrix, and then the recommendation is given by using the collaborative filtering method based on gradient descent. The proposed method is applied to the recommendation of wax dyeing products of Miao nationality and compared with the existing traditional methods. The results show that the improved algorithm greatly improves the recommendation performance, reduces the recommendation error and improves the accuracy.

1. Introduction

The rapid development of information technology, mobile Internet and cloud computing not only provide abundant information resources for human beings, but also bring human beings into a serious situation of "information overload" [1], resulting in the real needs of users drowned in a large number of irrelevant information and diversified commodities. Therefore, the recommendation system emerges, and the recommendation based on Collaborative Filtering (CF) has become the most widely used recommendation algorithm because of its accuracy, simplicity and effectiveness [2]. However, data sparseness has always been a difficult problem to be solved in collaborative filtering algorithms. To solve this problem, researchers at home and abroad have put forward various improvement methods and solutions. The problem of data sparseness proposed in the paper [3] will adversely affect collaborative filtering from two aspects: inaccurate neighbor search and low neighbor score. In reference [4], a robust constraint information embedding method is proposed to construct the relational matrix, which reduces the sparsity of the matrix. Literature [5] proposes a method to improve scoring similarity by combining user scoring trust and user preference trust, and to reduce data sparsity and improve accuracy. A collaborative filtering algorithm based on missing value iterative prediction filling is proposed in reference [6], which not only reduces the data sparsity, but also improves the accuracy of user similarity calculation.
Based on the above analysis, in view of these small-scale national handicrafts recommendation behavior, this paper takes batik products as an example, and proposes an improved collaborative filtering algorithm based on gradient descent for national handicrafts recommendation method. Through the developed batik products scoring system to obtain user scores, establish a user-product scoring matrix. Before applying the collaborative filtering algorithm based on gradient descent. At first, we find out the user union of the current product and another product, and the user set of the current product that has not been rated, then predict the current product rack by the user rack collaborative filtering algorithm, so as to reduce the sparsity of the data rack and improve the accuracy of the prediction results.

2. Background study

2.1. Collaborative filtering algorithm

2.1.1. Basic concepts. The CF algorithm was proposed by Goldberg et al in 1972 and applied to e-mail recommendation system [7]. The CF algorithm uses the past behavior or opinions of existing user groups to predict the most likely favorite items of current users or the rating of item interest. CF algorithm is divided into model-based collaborative filtering algorithm and neighborhood-based collaborative filtering algorithm. The neighborhood-based collaborative filtering algorithm is divided into user-based collaborative filtering algorithm (User-based CF) and item-based collaborative filtering algorithm (Item-based CF).

2.1.2. Application Process. Collaborative filtering algorithm does not rely on the characteristics and attributes of users and the project itself, but through the analysis of historical data of user preferences, to find the relationship between the current user and the rest of the users [8].

The core of CF algorithm is composed of three steps: (1) collecting user preference scoring data and establishing user-item scoring matrix; (2) through similarity calculation of users (or items), finding the nearest neighbor set of current users (or items), and selecting top - N nearest neighbors; (3) Predict the current user's rating of the target item according to the rating data of the nearest neighbor set, and make recommendation.

2.2. Gradient descent method

Gradient Descent is a first-order optimization algorithm. The calculation process is to solve the minimum value along the direction of gradient descent [9]. Its basic loss function formula is shown in Equation (1):

\[
J(\theta_0, \theta_1) = \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2
\]

Where, \(x_i\) represents the \(i\)-th sample characteristic, \(y_i\) represents the output corresponding to the \(i\)-th sample, \(h_\theta(x_i)\) is a hypothetical function.

3. Algorithm Proposal and Model Construction

3.1. Algorithm description

The traditional collaborative filtering algorithm is to calculate the similarity between users according to the user’s scoring of items, thus predicting the user’s scoring of unscoring items. Because the number of user scoring matrices of Miao batik products is far less than that of e-commerce or movie recommendation websites, sparse matrices are the key factors limiting the accuracy of batik products recommendation. Combining the User-based CF algorithm with Gradient Descent algorithm, the User-
based CF algorithm is used to pre-process the scoring matrix to increase the number of scoring matrices, and then the collaborative filtering algorithm based on gradient descent is used to predict the scoring.

Note that $I_i$ and $I_j$ represent the set of users who have scored the articles $i$ and $j$, respectively, and $C_{ij}$ is the union of $I_i$ and $I_j$, i.e., $C_{ij} = I_i \cup I_j$; find the user set $D$, i.e., $D = C_{ij} - I_i$, which has not scored product $i$ in the set $D$, and then use the User-based CF algorithm to predict the user's score of product $i$ in the set $D$ and fill it back to increase the intensity of the scoring matrix.

3.2. Data preprocessing

3.2.1. Establishing User-Project Scoring Model. All of the user's scoring records can be considered as a user-item scoring matrix $R$, which includes $m$ users and $n$ items. Each row vector represents a user's rating set of $m$, and each column vector represents a rated set of $n$ items. $r_{ui}$ is used to represent the user $u$'s rating of item $i$. The scoring matrix is represented by Equation (2).

$$ R = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m,1} & r_{m,2} & \cdots & r_{m,n} \end{bmatrix} $$

(2)

3.2.2. Similarity calculation. In the user-based collaborative filtering algorithm, the core is to calculate the similarity between users and obtain the current user's nearest neighbor, commonly used Pearson correlation coefficient, cosine similarity and modified cosine similarity to calculate the measure [10]. Because the modified cosine similarity measure is better in the case of sparse data, this paper uses the modified cosine similarity to process the initial scoring matrix. The formula is shown in Equation (3).

$$ Sim(u_a, u_b) = \frac{\sum_{i \in I} (r_{u_a,i} - \bar{r})(r_{u_b,i} - \bar{r})}{\sqrt{\sum_{i \in I}(r_{u_a,i} - \bar{r})^2} \sqrt{\sum_{i \in I}(r_{u_b,i} - \bar{r})^2}} $$

(3)

Where, $u_a$ and $u_b$ represent any two users in the user space, $Sim(a, b)$ represents similarity between the two users, $I$ represents batik products, $I$ represents product sets, $r_{u_a,i}$ and $r_{u_b,i}$ represent ratings of the products $i$ by the users $u_a$ and $u_b$, respectively, and $\bar{r}$ represents average ratings of the products $i$.

3.2.3. Prediction of Batik Product Score. The user whose similarity value is greater than 0 is taken as the nearest neighbor set $D_{Nu}$ of the current user in the $D_v$, and the prediction score of the current user $u_a$ to the current product $i$ is calculated by formula (4).

$$ Pr(u_a, i) = r_a + \frac{\sum_{u_b \in D_{Nu}} Sim(u_a, u_b) \times (r_{u_b,i} - \bar{r})}{\sum_{u_b \in D_{Nu}} Sim(u_a, u_b)} $$

(4)
Where, $r_u$ represents the average score of all articles rated by the user $u$, and $r_{ui}$ is the score of the product $i$ rated by the user $u$.

3.3. Collaborative filtering algorithm based on gradient descent

There are three main parameters of collaborative filtering algorithm based on gradient descent. They are user $j$’s scoring $y_{ji}$ on item $i$, vector $\theta_{ij}$ composed of item’s eigenvalue $x_{ij}$, vector $\theta_{ij}$ composed of user’s preference degree $\theta_{ij}$ for item’s eigenvalue. Because of the obvious characteristics of batik products, the gradient descent method of collaborative filtering algorithm for scoring prediction, the results are more accurate, the recommendation effect is more satisfactory to users.

There are four steps to apply the algorithm to the recommendation of batik products: (1) getting the score of the product $i$ from the current user $j$ and constructing the data set; (2) randomly initializing the eigenvalue $x_{ij}$ of the product $i$ to form the eigenvector $\theta_{ij}$, randomly initializing the user's preference degree $\theta_{ij}$ to the product feature to form the vector $\theta_{ij}$; (3) For the correlation loss function, the gradient descent method is used to solve the $x_{ij}$ values of $\theta_{ij}$ and respectively; (4) Predicting the user's rating of the unrated article using the values of $\theta_{ij}$ and $x_{ij}$.

According to the above steps and the basic loss function shown in Equation (1), and combining with the example of this study, the relevant loss function according to this example is constructed. The optimization objective loss function is shown in Equation (5).

$$ J(x^{(1)}, \ldots, x^{(u)}, \theta^{(1)}, \ldots, \theta^{(u)}) = \frac{1}{2} \sum_{(j,i) \in T} ((\theta^{(j)})^T x^{(i)} - y_{ij})^2 + \frac{\lambda}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij})^2 + \frac{\lambda}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} (\theta_{ij})^2 $$

The gradient descent function is shown in Equation (6).

$$ \theta_{ij} := \theta_{ij} - \alpha ( \sum_{j,r(i,j) = 1} ((\theta^{(j)})^T x^{(i)} - y_{ij}) x_{ij} + \lambda \theta_{ij}) $n_{1} \alpha \sum_{j,r(i,j) = 1} ((\theta^{(j)})^T x^{(i)} - y_{ij}) \theta_{ij} + \lambda x_{ij}) $$

$$ x_{ij} := x_{ij} - \alpha ( \sum_{j,r(i,j) = 1} ((\theta^{(j)})^T x^{(i)} - y_{ij}) \theta_{ij} + \lambda x_{ij}) $$

The scoring prediction formula is shown in Equation (7).

$$ \text{pre}(j, i) = (\theta^{(j)})^T (x^{(i)}) $$

Where, $y_{ij}$ represents the vector formed by the user $j$ score, $x_{ij}$ is a $5 \times 1$ vector, and the first 1-behavior intercept 1, 2-behavior batik article eigenvalues; $\theta_{ij}$ is also a $5 \times 1$ vector, the first 1 behavior 0, 2 ~ 5 behavior of the user's rating of the eigenvalues of batik products.

3.4. Model Construction of Improved Algorithm

According to the above analysis, the model of the improved collaborative filtering algorithm is shown in Figure 1.
In accordance with figure 1, the model is divided into two modules: ① A preprocessing module of a user-product scoring matrix, User-based collaborative filtering algorithm is used to improve the sparsity of scoring matrices: First of all, the users' scoring (1 ~ 5 grades) is obtained through the established scoring system of batik products, establishing scoring matrix. Then, the set of users who have scored the products \( i, j \) is constructed, and the users who have not scored the products \( i \) are searched in the set, and the set is formed. Then, the similarity between the users in the set and the users who have scored the products \( i \) is calculated by formula (2). According to the similarity, the score of the users to the products \( i \) in the set \( D_{ij} \) is predicted by formula (3), and the predicted value calculated by UB-CF algorithm is possible from 1 to 5. The predicted scores are then backfilled into the original sparse scoring matrix. ② Recommendation module of collaborative filtering algorithm based on gradient descent: Based on the filled user-item scoring matrix, the random initialization values of the four features (geometric pattern, animal-plant pattern, circular symmetry and square symmetry) of the batik product are added, and then the product eigenvalue vector \( \mathbf{x}^{(p)} \) and the user preference vector \( \theta^{(u)} \) of the product feature are optimized continuously through the formula (8) until convergence; Finally, the similar products are determined by calculating the similarity of the eigenvalues of different products, and then the user is graded and predicted by the formula (10).

For a better illustration of how this algorithm has been improved, an example is given below, as shown in Example 1.

Example 1. Let dataset data1 be the score of 5 product patterns by 5 users, as shown in Table 1.

|   | \( i_1 \) | \( i_2 \) | \( i_3 \) | \( i_4 \) | \( i_5 \) |
|---|---|---|---|---|---|
| \( u_1 \) | 5 | 2 | 4 | 0 | 2 |
| \( u_2 \) | 2 | 4 | 0 | 3 | 3 |
| \( u_3 \) | 0 | 3 | 4 | 1 | 2 |
| \( u_4 \) | 1 | 0 | 2 | 5 | 0 |
| \( u_5 \) | 3 | 0 | 0 | 3 | 1 |

Where, 0 indicates that the article is not rated by the user. According to the improved methods mentioned in Section 3.1 and 3.2, combined with Table 1, the specific method of reducing matrix sparsity in this algorithm is as follows: according to Table 1, \( I_1 = \{ u_1, u_2, u_4, u_5 \} \), \( I_2 = \{ u_3, u_4, u_5 \} \), then \( D_{ij} = \{ u_1 \} \), indicating that \( u_1 \) does not grade product \( i_1 \); Then, the similarity between \( u_1 \) and \( u_2 \), \( u_3 \), \( u_4 \), \( u_5 \) is calculated according to the formula (2), and the score of the article \( i_1 \) with \( u_1 \) is predicted according to the formula (3), and backfilled.
4. Example Verification and Result Analysis

4.1. Experimental data acquisition

Through the developed Miao batik products scoring system to obtain the original user scoring, as shown in Figure 2, the system is programmed with MATLAB. There are 100 patterns of batik products in the system, 6 patterns are randomly generated in each group, 50 university students are selected as test users (marked as users 01-50), and the product patterns are scored in the range of 1-5 grades. The higher the score, the greater the user's preference for the batik products, and each student is scored in at least two groups of items.

![Figure 2. Batik product grade system](image)

Figure 2 (a) shows the initial page of the scoring system. Enter the user number in the "User ID" column, and click the "Random Generate Product Pattern" button to generate six product patterns. As shown in Figure 2 (b), the user scores the patterns one by one, and the scores and user numbers are stored in a specified Excel file in the form of user ID-product ID-score.

4.2. Evaluation index

In this experiment, the average absolute error (MAE) and accuracy (PRECISION) were used to compare the performance of the proposed algorithm.

(1) MAE evaluates the accuracy of the recommendation results by calculating the deviation between the actual score and the predicted score [11]. The formula is shown in Equation (8).

\[
MAE = \frac{1}{N} \sum_{u \in U} \sum_{i \in T} |r_{ui} - \hat{r}_{ui}|
\]

Where, \( r_{ui} \) is the actual rating of the article \( i \) by the user \( u \), \( \hat{r}_{ui} \) is the rating predicted by the algorithm, and \( N \) is the size of the set of items. The smaller the MAE value, the higher the prediction quality.

(2) When providing recommendations, users are generally given a personalized list of recommendations called TOP-N recommendations, and the prediction accuracy of TOP-N recommendations is generally measured by accuracy measures [12]. Accuracy is defined as Equation (9).

\[
\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}
\]
Where, $R_n$ is a TOP-N recommendation set consisting of items with the largest $N$ prediction value selected from the prediction scores generated for it by the algorithm, $T_u$ represents the set of scoring items for user $u$ in the test set, and $U$ represents the set of users with recommendation results.

4.3. Experimental results and analysis
The experiment uses MATLAB R2017a software as a simulation tool to realize the programming and testing of the algorithm. The experiment verifies the effectiveness of the proposed algorithm through cross-validation. In this paper, the population size of the algorithm is $N = 100$, the total number of users is $m = 50$, the number of iterations is $G = 10000$, the initialization is $\lambda = 0.1$, $\lambda = 0.1$.

4.3.1. Recommendation Result Analysis. In order to verify the effectiveness of the proposed algorithm, comparative experiments are carried out. By setting the same parameter environment, the MAE values and accuracy of the proposed algorithm are compared with the traditional gradient descent-based collaborative filtering (GD-CF) algorithm and the traditional user-based collaborative filtering (UB-CF) algorithm. The experimental results are shown in Figures 3 and 4.

![Comparison of MAE line chart of three algorithms](image1)

![Comparison of the precision of three algorithms](image2)

**Figure 3.** Comparison of three algorithms

As can be seen from Figure 3 (a), the MAE value decreases with the increase of the number of nearest neighbors and tends to be stable. When the number of nearest neighbors is 16, the MAE value of this algorithm is about 0.436, while the MAE value of the traditional GD-CF algorithm is about 0.506. When the number of nearest neighbors is 12, the MAE value of the traditional UB-CF algorithm is about 0.625. For this example, the algorithm in this paper is always superior to the traditional algorithm.

As can be seen from Figure 3(b), the accuracy of the proposed algorithm increases with the increase of the number of nearest neighbors, and gradually tends to be stable. When the number of nearest neighbors is 16, the accuracy is the highest, about 0.212, while the accuracy of traditional GD-CF algorithm is about 0.189. When the number of nearest neighbors is 12, the accuracy of traditional UB-CF algorithm is the highest, about 0.165. Therefore, for this example, the accuracy of the proposed algorithm is superior to that of the traditional algorithm.

According to the above analysis, the proposed algorithm is superior to the traditional collaborative filtering algorithm for the batik products.

4.3.2. Analysis of user preference eigenvalue results. According to the improved algorithm, the total score of the users on the batik product feature score is shown in Figure 4, including geometric pattern, animal and plant pattern, circular symmetry and square symmetry.
As can be seen from Figure 4, circular symmetry has the highest score, about 40, indicating that most users prefer circular symmetry patterned batik products; Geometrical pattern has the lowest score, about 7, indicating that the geometrical pattern elements of the artifacts conform to the preferences of a small number of users.

The acquisition of user preference eigenvalues can clarify the current user's preference behavior, and recommend the products with this feature to the current user, to provide users with more accurate personalized customization and recommendation services for batik products.

5. Conclusion
A collaborative filtering algorithm based on UB-GD is proposed to alleviate the sparseness of the scoring data for Miao batik products. The initial scoring matrix was obtained by batik product scoring system, and the scoring was predicted by UB-CF algorithm, and the data was backfilled. Then, using GD-CF algorithm, the recommendation list and prediction value are given under the populated user-item scoring matrix. Compared with the traditional algorithm, the results show that the proposed algorithm improves the sparsity of data and the quality of recommendation effectively. And the GD-CF recommendation algorithm can be used to calculate the user’s preference for the characteristics of batik products, as well as the eigenvalues of each product picture, can be customized for the following national handicrafts personalized data acquisition basic work. The development of recommendation system for other ethnic handicrafts is the next step to study.

Acknowledgments
This work was financially supported by the Provincial Project Foundation of Guizhou, China (Nos. [2018]1049, [2016]7467) fund.

References
[1] Isinkaye F O, Folajimi Y O, Ojokoh B A. Recommendation systems: Principles, methods and evaluation [J]. Egyptian Informatics Journal, 16(2015),  pp. 261-273.
[2] L. Xiang. Recommendation System Practice, M. Beijing: People's Post and Telecommunications Publishing House, 2012.
[3] Y. J. Leng, Q. Lu, C. Y. Liang. Survey of Recommendation Based on Collaborative Filtering, J. RP&AI, 27(2014), pp. 720-734.
[4] M. Yu, W. Yang, W. Wang, et al. Co-clustering of Multi-entities Sparse Relational Data in Microblogging, J. Journal on Communications, 37(2016), pp. 151-159.
[5] L. Li, Y. X. Dong, et al. Collaborative Filtering Recommendation Algorithm Combined with User Trust, J. Journal of Chinese Computer Systems, 38 (2017), pp. 951-955.
[6] Y. Lu, Y. A. Liu. A Collaborative Filtering Algorithm Based on Predicting and Filling Missing-data by Iterated [J]. Computer & Digital Engineering, 2016, pp. 13-17.

[7] Y. Y. Mao, J. X. Liu, et al. Collaborative Filtering Algorithm Based on Logistic Function and User Clustering, J. Journal of Zhejiang University: Engineering Science. 51(2017), pp. 1252–1258.

[8] X. L. Weng, WANG Z. J. Research Process of Collaborative Filtering Recommendation Algorithm. Computer Engineering and Applications, J. Computer Engineering and Applications. 54(2018), pp. 25-31.

[9] B. M. Tao, S. X. Lu. Non-equilibrium Data Classification Based on Adaptive Stochastic Gradient Descent Method, J. Computer Science, 45(2018), pp. 487-492.

[10] S. Li, S. Q. Li, B. Liu. Improved Collaborative Filtering Algorithm and Its Parallel Implement, J. Computer Engineering and Design. 39(2018), pp. 3853-3859.

[11] X. W. Meng, S. D. Liu, et al. Research on Social Recommendation System, J. Journal of Software. 26 (2015), pp. 1356-1372.

[12] X. Y. Huang, S. Y. Long, J. Xie. Collaborative Filtering Recommendation Algorithm Combined with User Interest Degree Clustering, J/OL. Application Research of Computers. 2019, 36(9). [2018-05-24]. http://www.arocmag.com/article/02-2019-09-016.html.