ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

Product Personalized Recommendation Engine Based on Collaborative Filtering Algorithm

Zhong Zichao

Guangzhou Institute of Science and Technology, Guangdong, Guangzhou, China, 510540
zhongzichao202@163.com

Abstract: *Nowadays, information plus Internet technology is continuously developing, and people have entered the era of information overload from the era of lack of information access. In this context, it is difficult to find one's own "tailor-made" information from the huge amount of information. In this context, the recommendation system came into being. This article uses the order data under an e-commerce platform, uses user-based and item-based collaborative filtering algorithms to personalize products, and builds a product recommendation system. Through the feasibility of actual application scenarios and the prediction score of the model, a series of evaluations were made on the algorithm, and reliable and effective prediction results were obtained from the three indicators of accuracy, recall and F1 score. The final recommendation effect of the TopN recommendation list obtained was significant, and the recommendation efficiency could reach 85%.Provide some ideas for the research and development direction of modern recommendation algorithms.*

Keywords: *Recommended system; Collaborative filtering; Personalize; Algorithm model*

1. Introduction

Information retrieval technology is based on the subjective initiative of users. The main way is that users input the keywords they want to search, and then retrieve all relevant information based on the keywords and finally present it to users, who make their own judgments and choices. The representative products of information retrieval technology are search engines, such as Google, Bing, etc. According to the nature of information retrieval technology, if users cannot provide keywords accurately, the search results will be reduced and good recommendations will not be achieved. And the search results are based on the keywords retrieved by users, and recommend relevant information from all aspects, which reduces the accuracy of recommendations.

Personalized recommendation system is more personalized based on the product recommendation results. In many areas where recommendation systems are used, e-commerce is particularly prominent. For example, if the search engine is used to search for the keyword "toothpaste", the results are generally various recommendations of toothpaste category related information, but the recommendation system may recommend you to better match toothbrushes, which are some essential differences. The recommendation system can create more specific user profiles and actual application scenarios. In order to tap the potential needs of users and achieve the purpose of effective recommendation greatly.

This paper starts from the actual needs of personalized recommendation in the actual application scenarios, and describes how to specifically implement personalized recommendation, how to completely establish, compare, and evaluate the entire process of collaborative filtering algorithm models based on users and commodities, and how to improve the models, such as SVD matrix decomposition and hybrid recommendation algorithms. At the same time, it also comprehensively introduces the understanding and mathematical basis of the model, algorithm ideas, etc. On the other hand, there is also a clear idea on how to build a model, from data acquisition to data cleaning to algorithm ideas. At the same time, it also describes some trends of the recommendation system algorithm and the future development direction [1].

2. E-commerce Commodity Recommendation System and Application Scenarios

The research on e-commerce product recommendation system can be traced back to 1992. It was originally used to deal with spam, and "tags" began to be the main form of recommendation/filtering. Later, it was extended to the film field, and collaborative filtering algorithm also came into the

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

researchers' sight. Netflix's recommendation algorithm competition set off a wave of recommendation algorithm waves at that time. Later, the recommendation algorithm penetrated into all walks of life [2], involving a wide range of fields. The recommendation system can be classified into three categories: application field based, design idea based and usage data based [3]. The first category includes the recommendation of friends on social software, the recommendation of products including e-commerce platforms, the recommendation of news and other content related information, and the recommendation of various search engines; The second is algorithms, such as content-based and knowledge-based collaborative filtering, hybrid recommendation algorithms, etc; The third type is data, including user portrait, user historical data, text information of information related context, etc.

This paper adopts the similarity based recommendation: Collaborative Filtering algorithm, which is the earliest and more famous recommendation algorithm, used to find similar users and similar products for recommendation. Collaborative filtering is a typical method to use collective wisdom. Collaboration refers to group behavior, and filtering refers to individual behavior. Collaborative filtering is based on the following assumption: If A is the same as B in terms of one product or view, then B is more likely to like the same product or hold the same view as A for another product or view than for a new C [4].

Similarity is mainly calculated by calculating the distance similarity between users/items. There are three common methods, namely, cosine similarity, Pearson correlation coefficient and Jackard similarity [5]. The calculation formulas of the three similarity degrees are relatively simple, and x and y refer to users and items respectively. The cosine similarity is the most common one.

Cosine similarity calculation is shown in Formula (1), where Cosine (x, y) represents the similarity between users/products, and represents the scored user set/product set.

$$Cosine(x,y) = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (1)$$

Pearson correlation coefficient is shown in Formula (2), where Person (x, y) represents the similarity between users/products, and represents the average score of users x and y on products.

$$Pearson(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \quad (2)$$

Jaccard is shown in Formula (3), where Jaccard (x, y) represents the similarity between user and user/commodity and commodity, the numerator is the element of the intersection of two sets and the numerator is divided into the proportion of each element.

$$Jaccard(x,y) = \frac{x \cap y}{x \cup y} \quad (3)$$

The core idea of the whole algorithm is to first generate user/item ratings based on the user's historical behavior data, that is, to obtain a user information matrix, as shown in Table 1, then determine the similarity matrix according to the similarity, as shown in Table 2, and finally conduct corresponding scoring to recommend. The basic algorithm used in this paper is user based collaborative filtering algorithm, and the specific process is divided into the following three parts:

(1) The user information matrix is obtained by scoring the products.

(2) Calculate the similarity between users to obtain the similarity matrix.

(3) Personalized recommendation based on the similarity of users.

Table 1: User Information Matrix

item / user	I_1	I_2	I_3	I_4	…	I_n
U_1	R_{11}	R_{12}	R_{13}	R_{14}	…	R_{1n}
U_2	R_{21}	R_{22}	R_{23}	R_{24}	…	R_{2n}
U_3	R_{31}	R_{32}	R_{33}	R_{34}	…	R_{3n}
U_4	R_{41}	R_{42}	R_{43}	R_{44}	…	R_{4n}
…	…	…	…	…	…	…
U_j	R_{j1}	R_{j2}	R_{j3}	R_{j4}	…	R_{jn}

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

Table 2: User similarity matrix

user / user	U_1	U_2	U_3	U_4	⋯	U_n
U_1	R_{11}	R_{12}	R_{13}	R_{14}	⋯	R_{1n}
U_2	R_{21}	R_{22}	R_{23}	R_{24}	⋯	R_{2n}
U_3	R_{31}	R_{32}	R_{33}	R_{34}	⋯	R_{3n}
U_4	R_{41}	R_{42}	R_{43}	R_{44}	⋯	R_{4n}
⋯	⋯	⋯	⋯	⋯	⋯	⋯
U_n	R_{n1}	R_{n2}	R_{n3}	R_{n4}	⋯	R_{nn}

The model evaluation criteria are divided into confusion matrix secondary indicators and SVD matrix decomposition.

First, before introducing the indicators of each evaluation, briefly introduce the secondary indicators of the confusion matrix. For the classification model of general prediction class, the positive result is that the more accurate the better. Therefore, the corresponding table of this confusion matrix must hope to make no/fewer mistakes. The TP/TP should be as large as possible, while the FP/FN should be as small as possible. However, the number of statistics in the confusion matrix is a number. In some large data sets, it is not enough to just count the number, and it is difficult to evaluate the quality of the measurement model. Therefore, four indicators are extended from the results of the basic statistical results framework of the confusion matrix, which are called secondary indicators. I summarize the definition, calculation and personal understanding of the four secondary indicators into a table for clearer presentation. The accuracy rate is specific to the whole model, as shown in Table 3.

Table 3: Confusion matrix secondary indicators

	Formula	Significance
Accuracy (ACC)	$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$	Proportion of all correctly judged results of the classification model in the total observation value
Precision (PPV)	$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$	The proportion of all results predicted by the model as positive
Sensitivity(TPR)	$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$	In all results where the real value is positive, the proportion of the predicted pair of the model
Specificity(TNR)	$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$	In all results where the true value is negative, the proportion of the predicted pair of the model

Through the four secondary indicators, we can well convert the quantitative results in the confusion matrix into a 0-1 ratio, which is better for the standardization of measurement. However, if we expand the four indicators, we will finally get a tertiary indicator, namely, F1 Score, with the formula (4). It is the result of combining the precision and recall outputs. The value range of F1 score is 0-1. 1 represents the best model output, and 0 represents the worst model output.

$$F1 - Score = 2\left(\frac{Precision * Recall}{Precision + Recall}\right) \quad (4)$$

Second, because the data set used in this paper is relatively sparse and the time of model operation is compared with the field, the improved model uses the matrix decomposition method (SVD). The core idea of SVD is to do the eigenvalue decomposition, and then obtain the left singular matrix according to the eigenvalue decomposition, and then indirectly obtain part of the right singular matrix. The model matrix is too large and very sparse, so the matrix can be decomposed to reduce dimensions, and the key information can be less missing, and most of the information of the original matrix can be retained. Formula and comparison efficiency are shown in Formula (5):

$$A = U\begin{pmatrix}\Sigma_r & 0\\ 0 & 0\end{pmatrix}V^T \quad (5)$$

Where U is the orthogonal matrix of m x m and V is the orthogonal matrix of n x n, Σ R is the singular value arranged from large to small along the diagonal, and the final user information matrix can be represented by a matrix with lower dimensions. It is mainly used to reduce the dimension and improve the matrix operation speed.

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

3. Data Preprocessing

Data size (rows): There are 102304 valid data, with a total size of (102304 x 7).

Data source: background order data of a T-mall flagship store (I obtained data from the database through desensitization processing with the approval of the leader during my internship in Internet e-commerce).

Data variable (column): the selected fields are customer ID(User ID), sub order ID, category ID, commodity ID (item ID), commodity name, payment status, order frequency (freq). See Table 4 for details.

Table 4: Data collection and data types

Data Sheet	Data Type
Behavior Table	User ID, Order ID, User Purchase Record, Timestamp Information, etc.
User Table	User ID, Gender, Age, Occupation, Telephone, Address, etc.
Item Table	Commodity ID, commodity name, category ID, commodity price information, etc.
Data table used after the last linked table	Customer ID (user_id), sub order ID, category ID, commodity ID (item_id), commodity name, payment status, order frequency (freq)

In the actual scenario, some of the data extracted from the system database are "dirty data". For this dataset, the following problems may exist:

(1) In the process of data embedding, that is, in the process of collection and transmission, errors will occur more or less, resulting in the lack or loss of order data dimensions.

(2) The payment dimension is stored as 0-1 data, but there may be errors in uploading non 0-1 data.

(3) This article uses the order data of Taobao background, so there may be multiple sub order information in one order information, so there may be complete duplicate data.

(4) Because the order data may contain more sensitive information, it is necessary to desensitize the data and replace sensitive information such as user ID and phone address with some numerical symbols.

Here you can view all the characteristic fields to observe the data in each column, including statistics (maximum, minimum, mean, standard deviation, quartile). Observing these statistics will help us quickly and intuitively observe the general range and situation of the data, and be able to judge the outliers. We found that the maximum and minimum values of ID differ greatly, but because this is a desensitized number, user ID only represents the ID and does not represent the size, so it is ignored. We found that the same is true for freq, which may be because some customers place orders for many times, leading to outliers. We observe outliers, as shown in the figure, and it is true that the scatter plot is biased to the right. However, it is normal and within the logical range. It will not have a great impact on the final results.

We find that the data is power-law distribution (long tail distribution), because most users only buy one product, and a few will buy several times. The proportion of users who buy more than five times is very small. In fact, this is real data. The real situation is that there are few people who repurchase across categories, which requires us to predict and recommend these potential possibilities through the recommendation algorithm. We then logarithmize the long tail distribution, and finally get the results as shown in Figure 1 - Figure 4.

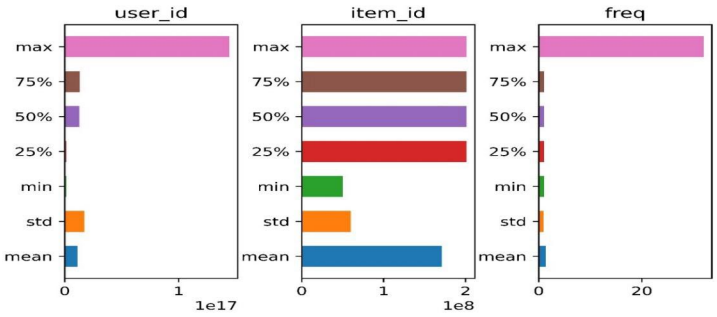

Figure 1: Description statistics visualization

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

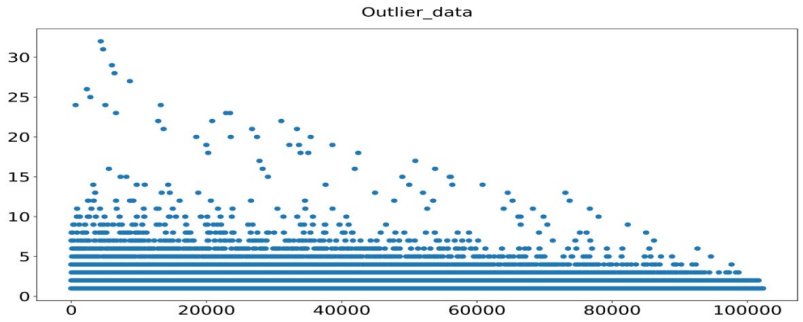

Figure 2: Outlier data

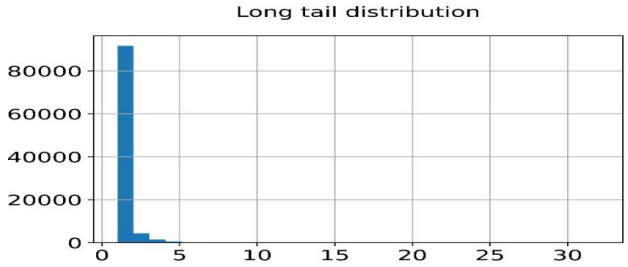

Figure 3: Data power law distribution

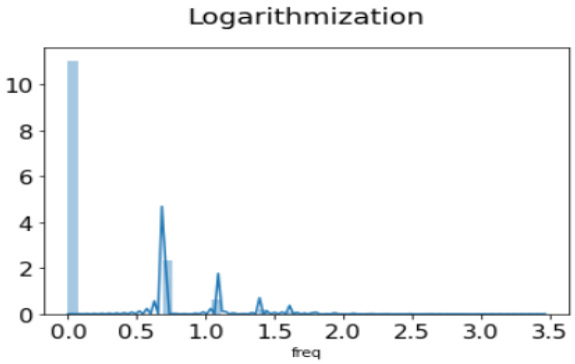

Figure 4: Data logarithmization

4. Establishment of Collaborative Filtering Algorithm Model

The three core fields of the user based collaborative filtering recommendation algorithm: (1) User; (2) Commodity; (4) Evaluation (scoring). Users in the data set ID, because it is real data, the user in the order information. The ID is desensitized, and the user is presented as a number ID; Goods are items purchased by customers ID, each customer may purchase goods for many times, and then the intersection of lines and columns is the customer order frequency, abbreviated as " freq " .The result table is as shown in the table (because the amount of data is too large, 10 x 5 data are taken for display to facilitate viewing).

Table 5: User information matrix

Item ID User ID	50009032	50009047	50010406	50011556	50012907
1244531471995514	NA	1	2	NA	NA
13049148252164926	3	NA	NA	1	3
13099350138880928	NA	NA	1	NA	NA
13125237547008923	1	NA	7	NA	2
12958610333715889	NA	NA	NA	2	NA
1283257280921676	NA	NA	NA	NA	NA
1245908714486880	NA	NA	NA	NA	NA
1246019801163080	NA	3	1	6	5
12958590607381855	3	1	NA	NA	NA
1246019430872482	NA	NA	1	NA	NA

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

From Table 5, we can see that the matrix is still very sparse, which contains a large number of null values NA. Because it is impossible for every user to buy every product, and the corresponding product will not have a score, so there will be a large number of null values in the middle. At this time, we need to convert all null values to 0, and finally get a user information matrix of the training set (69283 x 14).

Next, we need to use Equation (2-1) to calculate the similarity between users, and then get a similarity matrix of the test set (69283 x 69283), as shown in Table 6 (because the amount of data is too large, 6 x 6 data are taken out for display to facilitate viewing).

Table 6: User similarity matrix

User ID User ID	124453165	124601975	124602153	124601982	124601979	124453498
124453165	1.000	1.000	0.742	1.000	0.823	1.000
124601975	1.000	1.000	0.567	0.423	0.331	1.000
124602153	0.742	0.567	1.000	1.000	1.000	1.000
124601982	1.000	0.423	1.000	1.000	1.000	0.334
124601979	0.823	0.331	1.000	1.000	1.000	1.000
124453498	1.000	1.000	1.000	0.334	1.000	1.000

The row and column indexes of the matrix are both User ID, the cross position is the similarity. Because the user is exactly the same as himself, the diagonal value is 1, while the other values are 1. In most cases, the products purchased by the two are identical. The similarity is calculated by the cosine similarity of equation (2-1). The range is 0-1. 1 represents almost the same, and 0 represents not the same at all.

(4) Now we have got two matrices, one is the user information matrix, and the other is the user similarity matrix. Now we need to cross the two matrices, that is, find the item in the user information matrix. All scores of ID, and then find the current user in the similarity matrix. For the first K users with the most similar IDs, the score of the item is finally extracted. It should be noted here that there will be some similar users who have scored 0 on the goods. Because this user has not used/placed an order for/scored the goods, this user cannot recommend things that he has not used. Therefore, when calculating the recommendation score, the denominator cannot be added with the user whose score is 0, which is also easy to understand logically. Because of the data volume limitation of this dataset, not all users can match the most similar users, nor can all products of similar users be recommended for completion. Because most people only buy one product, there will still be a lot of 0 values in the matrix.

In this way, after completing all the user information tables, we can easily get a structure of User ID, see in Table 7. There are two parameters to be determined, one is the user information table, and the other is the N of the first N products to be recommended. Not all users will get the recommended list, because some users may not match similar users at all. The N in this paper is taken as 3 (the top three products recommended by a user are obtained), and the Top 3 recommendation list is obtained.

Table 7: Final result

	User ID	**Item ID**	**Recommend Score**
1527	1244531508145710	50018971	1.0
3626	1244531560574310	201328302	1.0
12874	1244531757117110	50012907	1.0
19514	1244531868502410	201292210	0.5
20955	1244531894869610	201328302	1.0
21184	1244531904606310	50018971	1.0
		201292210	0.8
		50009032	0.5
28363	1244532107690310	201292210	1.0
44764	1244532753217110	50018971	0.7
46185	1244532779471910	50009032	1.0

In fact, the article based collaborative filtering algorithm, according to the principle of the algorithm, is the same as the user based idea. User based collaborative filtering is to find similar users and then recommend the items used by similar users, while article based collaborative filtering is to find items similar to the items that "target users" like, and then recommend the similar items to "target users" [6]. The algorithm based on items is also a two step process. First, calculate the similarity of items, and then make recommendations based on the similarity of items and the historical items used by the target users. Therefore, building an article based algorithm actually makes a transposed transformation of the user

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

similarity matrix, and then filters according to the same process to obtain an article based collaborative filtering algorithm. As shown in Table 8 (the results are too long to show only partial data).

Table 8: Final result

	Item ID	User ID	Recommend Score
1	201292210	1244531730997513	0.3
		13157945384963909	1.0
13	50023725	13166566580224908	1.0
2	124988005	1246020309187989	0.6
		1246019666784882	0.7
6	201328302	13131139850240905	1.0
		13129828343808899	0.6
		1244531472033318	0.9
3	201275777	13170129051648897	1.0
5	201313101	1244532160148114	0.8
12	50018971	1246020115676386	1.0

The final TopN recommendation list is compared with the test set, that is, the products that the user is recommended to buy are predicted correctly if the user does buy, and the products that the user does not buy are predicted incorrectly, which does not achieve a better recommendation effect. Take the user based collaborative filtering as an example to show the final results. The article based collaborative filtering form is also the same (only part of the data is displayed).

Table 9: Final result

Item ID / User ID	50009032	50009047	50010406	50011556	50012907	
1244531471995510	0	1	2	1	3	Forecast Error
13049148252164900	3	2	0	1	3	Original Information
13099350138880900	0	3	1	1	1	Correct Prediction
13125237547008900	1	0	7	0	2	
12958610333715800	1	0	1	2	0	
1283257280921670	0	0	0	4	0	
1245908714486880	3	0	1	0	1	
1246019801163080	0	3	1	6	5	
12958590607381800	3	1	0	0	1	
1246019430872480	0	0	1	1	0	

It is obvious from Table 9 that the prediction error probability is not high for almost one user, and most errors are caused by the low similarity of users due to the sparse matrix. There are indeed many spaces that can be mined or even predicted in the real shopping scene, and the collaborative filtering algorithm is relatively dependent on prior data, so this may be because the data volume is large enough, so the effect of the model fitting is considerable. It has certain reference value for the recommendation system to deploy personalized recommendation of goods.

5. Model Evaluation

This paper uses Formula (2-4) F1 score to show the prediction results of the model. Based on the prediction results of the training set and the test set, the final accuracy rate of the model operation is 0.854, the recall rate is 0.875, and the final F1 score is 0.864, as shown in Table 10.

Table 10: Model score

Index	0-1 Probability Score
Accuracy	0.854
Recall	0.875
F1 Score	0.864

It can be seen that even in the real purchase scenario, the collaborative filtering algorithm has a certain degree of controllability, excavatability and enforceability for the promotion of commodity repurchasing, commodity joint sales and other aspects of mass commodities on the e-commerce platform.

ISSN 2616-5775 Vol. 6, Issue 2: 36-43, DOI: 10.25236/AJCIS.2023.060206

6. Conclusion

First of all, this paper expounds the significance of the research on the recommendation system and specifically introduces the development history of the recommendation system, and analyzes the current research status of the recommendation system and the development of the algorithm; Secondly, what role does the recommendation system play in our daily life. Then it introduces the theoretical basis of collaborative filtering algorithm and recommended evaluation indicators; Then, according to the data set captured by itself, the code compiled step by step analyzes the specific application of collaborative filtering algorithm to personalized recommendation of goods in Python. Finally, the model is scored and some mainstream optimization schemes are proposed to provide research direction and ideas for readers.

In recent years, papers on information filtering, recommendation engine robustness, user privacy and other aspects in RecSys conference have been included more and more vigorously, which will be an important research direction of future recommendation system development [7].

References

[1] Sun Jixiang. Research and Application of User Portrait in Recommendation System [D]. Beijing: North University of Technology, Master's Thesis, 2020.

[2] Zhao Liang, Hu Naijing, Zhang Shouzhi,et al. Design of Personalized Recommendation Algorithm [J]. Computer Research and Development, 2002, 39(8):986-991.

[3] Chang Hao, Yang Shengquan, et al. Research on Product Recommendation Algorithm Based on Collaborative Filtering Decision tree [J]. Value Engineering, 2020,9(52):127-128.

[4] Fan Zezhou. Research on the Online Shopping Guide Product Recommendation System of Company A [D]. Beijing: Beijing Jiaotong University, Master's Thesis,2019.

[5] Gao Jian. Research and Implementation of Personalized Recommendation Algorithm on WeChat E-commerce Platform [D]. Lanzhou: Chang'an University, Master's Thesis,2019.

[6] Gao Yukai, Wang Xinhua, Guo Lei, et al. User Cold Start Recommendation Algorithm Based on Collaborative Matrix Decomposition[J]. Computer Research and Development , 2017,8: 188-198.

[7] LindenG, SmithBR, YorkJC, et al. Amazon. Comrecommendations: Item-to-item Collcaborative Filtering[J]. IEEE Internet Computing, 2003,7(1):76-80.