International Core Journal of Engineering Volume 8 Issue 9, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

Research on Collaborative Filtering Recommendation Algorithm based on Item Feature Clustering

Jie Fang, Qingchun Fan

School of Computer Science and Technology, Hefei Normal University Hefei, 230601, China

*E-mail: fangjie@hfnu.edu.cn

Abstract

In order to solve the problems of poor scalability and low recommendation accuracy of traditional item similarity measurement methods, an improved collaborative filtering recommendation algorithm based on item feature clustering was proposed. Based on full consideration of the similarity of item scores, clustering methods are used to cluster similar items into sets to more effectively discover similar neighbor items of the target item. Based on the MovieLens data set, the experimental results show that the recommendation results of the algorithm are significantly better than the current commonly used collaborative filtering methods in terms of prediction effectiveness, prediction error and recommendation accuracy. The algorithm only considers the ratio of item rating values and does not fully utilize the absolute rating values of items. Comprehensive comparison, the algorithm has obvious improvement in scalability and accuracy compared with the traditional method.

Keywords

Recommendation System; Project based; Collaborative Filtering; Similarity.

1. Introduction

The world wide web is a mass of information transmitted and shared by human beings all over the world. Massive information leads to the phenomenon of "information load", which makes it more and more difficult for Internet users to accurately obtain the information they are interested in. The traditional artificial intelligence system can not meet the growing demand for personalized consumption recommendation in the era of big data. Collaborative filtering recommendation algorithm has been adopted in related fields. This algorithm realizes the mining and recommendation of related information retrieval by mining readers with similar preferences or similar project information. This paper proposes an improved collaborative filtering recommendation algorithm based on project. Through in-depth mining of data set information, the collaborative filtering recommendation algorithm based on user project score and project similarity is proposed to recommend more accurate matching commodity information for users.

2. Based on the Proposal of Collaborative Filtering Algorithm

With the increasing popularity of information technology, the development of computer information retrieval algorithm, especially data mining and recommendation system, has also entered the golden stage. In the face of massive commodity information, service providers need to search for commodities that meet their individual needs for Internet users quickly, accurately, efficiently and conveniently, and provide users with reasonable suggestions. In order to overcome the phenomenon of "information overload" and realize the effective utilization of information, it has become a key topic in the field of Internet and e-commerce industry. The essence of personalized recommendation

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

is to replace the user to judge and agree on the objects he has not contacted. Driven by the thinking mode of developing a personalized recommendation system that can meet the uncertain needs of users, a number of personalized recommendation systems have been born. According to the different speed of responding to user behavior and considering the characteristic data, the recommendation system based on offline training, online training and collaborative filtering are derived. Collaborative filtering (CF) is the most classical and commonly used algorithm in recommendation system CF algorithm originated in 1992 and was used by Xerox to customize the personalized email system. Xerox users need to select three to five topics from dozens of topics. Xerox filters emails according to different topics, so as to achieve the purpose of personalized recommendation, marking the birth of the recommendation system.

Collaborative filtering algorithm has been widely used in the field of e-commerce sales. The recommendation algorithm collects and analyzes consumers' accurate personalized consumption score information, and realizes personalized service recommendation through algorithm filtering. The world's first collaborative filtering recommendation system for e-commerce is Grundy [2]. Its interest model is based on the user's historical information to generate the user's interested book directory and recommend it to the user. Systems with similar principles include Amazon, eBay, Dangdang, Douban, etc.

Based on the measurement from the perspective of user similarity and project preference similarity, the collaborative filtering algorithm derives and develops user based collaborative filtering (ubcf) and item based collaborative filtering (ibcf). Ubcf calculates and measures the similarity between users by analyzing the historical access data of users, and then provides personalized recommendation services by using the nearest neighbor user association data. Ibcf measures the similarity between items by analyzing the behavior data information of users originated from items, and makes personalized recommendations for users according to the similarity of items and users' historical interests.

Cosine similarity, modified cosine similarity and Pearson correlation coefficient are classical similarity calculation methods. It can be seen that the calculation of similarity is the core of recommendation system. How to mine similar users or similar items has always been the focus of research in this algorithm field. In the context of Internet big data application, user based collaborative filtering, due to the large number of users can not effectively match the small number of items, the cost of storing user similarity matrix increases sharply, the initial user behavior matrix is too sparse, and the accuracy of the algorithm to find similar users will be reduced. In view of the technical defects and cost considerations of ubcf algorithm, Amazon and Netflix adopt ibcf algorithm when implementing basic recommendation system.

3. Principles and Concepts Related to Collaborative Filtering Recommendation

3.1 Definition of Collaborative Filtering

Collaborative filtering refers to the use of shared interests, common personal hobbies and group preferences to actively recommend personalized information of user preferences. Individuals give information a considerable degree of response (such as evaluation) through a cooperative system and record it, and then assist others in filtering information to achieve the purpose of filtering. Responses are not limited to those of particular interest, and records of information that are not of interest are also quite critical, so records of information that are not of interest should not be ignored and ignored.

From the perspective of behavioral characteristics, collaborative filtering can be divided into two types: rating and social filtering. Collaborative filtering algorithm has received extensive attention and research in the field of Internet research all over the world due to its excellent linear rate and robustness.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

3.2 Clustering Concepts and Algorithms

Clustering refers to the process of grouping a collection of entities or abstract data elements into multiple classes or clusters of similar elements. It can be divided into the following detailed steps: data preparation, feature selection and abstraction, concrete feature extraction, and cluster grouping. Hierarchy-based, partition-based, data density-based, data grid-based and model-based clustering algorithms are widely used.

Item-based clustering is to select m representative item attributes to form an m*n two-dimensional feature matrix for all items (commodities), and then cluster the items according to the similarity measurement method. cosine similarity, modified cosine similarity, or Pearson correlation coefficient. The purpose of clustering is to group data objects (items) into multiple classes (clusters), to achieve high similarity between data elements in the same cluster, while objects in different clusters are quite different.

Suppose the entire project collection $i=\{i_1,i_2,...,i_m\}$, The generated s clusters are represented by set $S=\{s_1,s_2\cdots s_m\}$, Then the item characteristics contained in the same cluster are as similar as possible.The steps of clustering the project based on the clustering algorithm are as follows:

Input: item characteristic initial matrix of $D=m\times n$, s is the number of expected clusters.

Output: s clustering matrices.

1) Based on the project data set D, the algorithm randomly selects m project samples from the project,m vector sets $i=\{i_1,i_2,...,i_m\}$as initialization parameters;

2) Select any s items Take their attribute characteristic data as the initial clustering center Record as a set $SS=\{ss_1,ss_2,\cdots ss_m\}$;

3) Initialize s clusters $s_1,s_2\cdots s_m$ to null Denoted as set $S=\{s_1,s_2\cdots s_m\}$;

4) For the remaining items after clustering, calculate the similarity $ss_j(j=1,2...,s)\ s(i_i,ss_j)$ with the cluster center, and merge each item into the cluster with the highest similarity;

5) For the newly created cluster, calculate the similarity mean of all items in it to generate a new cluster center;

6) Repeat steps 4) and 5) until new clusters are no longer generated.

After the above clustering operations, the items with high feature similarity are merged and clustered into the same class to form their own feature groups. When performing preliminary pre-scoring of specific items, it is beneficial to improve the speed of finding similar item groups. Clustering items is time-consuming, but can be done in an offline cycle.

3.3 Classification Concepts and Algorithms

Classification technology and taxonomy is a method of establishing a class model based on an input sample set, and labeling the unknown sample class number according to the class model. The data classification process is mainly composed of two processes: the first step is to build a model, and then use the model to classify. The algorithms that are used more and are relatively mature include: K nearest neighbors; decision tree;Bayesian classification;Support vector machine.

4. Similarity Metrics

Euclidean distance is the most familiar way to describe distance, and it is generally considered to define the basic distance of similarity. If in n-dimensional space, each point can be regarded as an n-dimensional real vector. It is defined as follows:

$$d(x_i,x_j)=\sqrt{|\sum_{k=1}^{p}(x_{ik}-x_{jk})^2|}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

Manhattan distance is based on the fact that traveling between two points must follow the grid line, which is similar to the street line of the city. The specific definitions are as follows:

$$d(x_i, x_i) = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$$

Euclidean distance and Manhattan distance are special cases of Minkowski distance. Minkowski distance is defined as follows:

$$d(x_i, x_j) = \sqrt[r]{\sum_{k=1}^{p} (x_{ik} - x_{jk})^r}$$

The calculation method of cosine similarity coefficient is defined as follows:

$$s(u,v) = \cos(u,v) = \frac{r_u * r_v}{\| r_u \|^2 * \| r_v \|^2} = \frac{\sum_{i=1}^{n} r_{u,i} * r_{v,i}}{\sqrt{\sum_{i=1}^{n} r_{u,i}^2} * \sqrt{\sum_{i=1}^{n} r_{v,i}^2}}$$

5. Description of Project-based Collaborative Filtering Recommendation Process

5.1 Algorithm Principle

The core idea of the IBCF algorithm is to recommend similar items to the user based on the similarity of items. Item similarity is distinguished from similarity in content or item attributes, but because they are commonly liked by multiple users. Therefore, the IBCF algorithm is used to calculate the similarity between the two items by measuring and comparing the historical behavior of visiting users. It is assumed that the more users who like this type of item in common, the higher the item similarity. The principle of item-based collaborative filtering algorithm is shown in Figure 1.

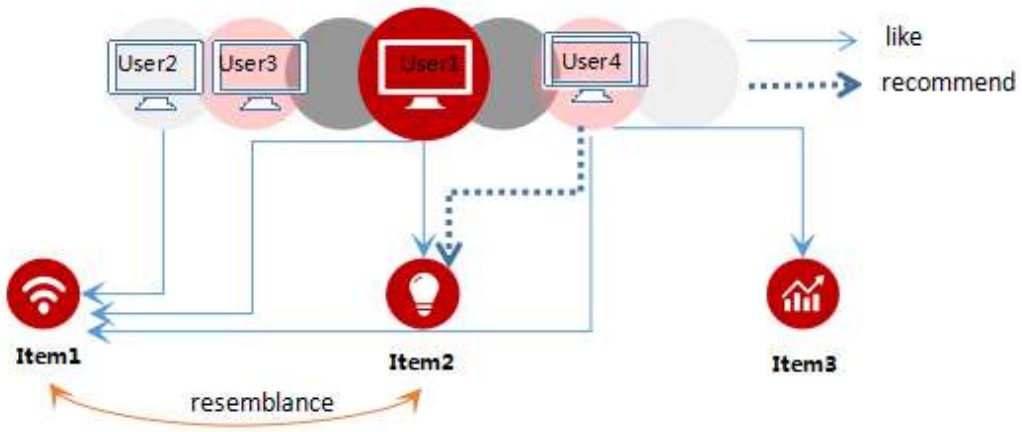

Figure 1. Schematic diagram of project-based collaborative filtering algorithm

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

Step1 Build a project scoring matrix.

Based on the initial access data (taking Table 1 as an example, the subsequent user is referred to by u, and Item1 is referred to by I), build an m * n-order matrix $R = \left[r_{(u,i)}\right]^{m\times n}$ with the access user (set the total number of users as m) as the row coordinate and the item score (set the total number of items as n) as the column coordinate. $r_{(u,i)}$ represents the score of the access user u on item i, which is usually measured in the 1-5 score system. When y, it indicates that the item does not generate a score.

Table 1. detailed user item scoring table

	Item1	Item2	Item3	Item4	...	Item-n
User1	i_{11}	i_{12}	i_{13}	i_{14}	...	i_{1n}
User2	i_{21}	i_{22}	i_{23}	i_{24}	...	i_{2n}
User3	i_{31}	i_{32}	i_{33}	i_{34}	...	i_{3n}
...	...	...	...	...	...	...
User-m	i_{m1}	i_{m2}	i_{m3}	i_{m4}	...	i_{mn}

Step 2 Based on Table 1, a matrix of user and item evaluations is formed. As shown in table 2.

Table 2. User-item rating matrix

	i1	*I2*	*I3*	*I4*	*I5*	*I6*
u1	3	5	0	0	4	0
u2	0	2	5	4	0	4
u3	4	5	0	5	4	5
u4	4	3	6	0	5	4
u5	5	0	5	0	0	5

Further, the m * n matrix can be further decomposed into m * P matrix and P * n matrix, which describes the potential interaction between users and projects. The decomposition principle is that the potential characteristics reflect the way users evaluate the project [19]. Calculate the similarity between two columns of vectors in matrix R and construct an item similarity matrix of order n * n [5]. At the same time, the clustering algorithm is used to cluster the items and generate a new cluster item set.

Step 3 obtains the target user u historical behavior data and forms a positive feedback item list to construct set H.

Step 4 search and retrieve set H, find the k most similar nearest neighbor elements of each item, take the union set and delete the items already contained in H to generate candidate item set H *.

Step 5 predicts the score $p_{u,i}$ of target user u on item i for candidate item set $\forall_i \in H^*$.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

Through the similarity calculation, the item set of the candidate target user u is formed, and then the score is predicted. The most commonly used prediction scoring method is the weighted average strategy based on the project mean, and its calculation formula is as follows:

$$p(u,i) = \bar{r}_i + \frac{\sum_{j \in H} s(i,j) \times \left(r_{u,j} - \bar{r}_j \right)}{\sum_{j \in H} s(i,j)}$$

Where H is the element set of positive feedback items of target user u, and $\bar{r}_i$ and $\bar{r}_j$ represent the average scores of all users on items I and j respectively. After the prediction score based on the item set, the maximum sum with high similarity to the user's favorite items is retrieved as the similarity recommendation list.

5.2 Recommendation Strategies

Take u_1 as the target user to recommend similarity. Then, the positive feedback item element set H of user u_1 is obtained, where $H = \{i_1, i_2\}$. Based on the user item scoring matrix, the vector elements of i_1, i_2, i_3, i_4, i_5 five items are extracted, and the pairwise similarity between items is calculated by Pearson correlation coefficient to obtain the similarity matrix of items.

Suppose when $k = 2$, find the 2 most similar neighbors of each item in the set H.Therefore, the two nearest neighbors most similar to are i_3 and i_5.The two nearest neighbors most similar to i_2 are i_4 and i_5,Generate candidate itemset by taking union set $H^* = \{i_3, i_4, i_5\}$.Next, the weighted average strategy based on the item mean is used to predict the score of user u_1 on the item in h *, and the values of $p_{1,3}$, $p_{1,4}$ and $p_{1,5}$ can be obtained.Use sum recommendation, take 2 as the value of N, and finally generate the recommendation list as $\{i_5, i_4\}$.

6. System Experiment

6.1 Test Dataset

The data set for testing and validating the algorithm in this paper adopts the free public data set MovieLens provided by the research experiment Grouplens of the University of Minnesota, including the ratings of 8,570 movies by 706 users, and a total of 100,023 comment records. We select MovieLens100k to verify the experimental results. The dataset contains 100,000 ratings of 1,682 movies by 943 users, each of whom rated at least 20 movies. The user-item rating matrix of the dataset has a sparsity of 100000/(943×1682)=6.3%. In order to test the performance of the recommendation algorithm, one dataset is divided into training set and the other as testing set.

6.2 Prediction and Evaluation Indicators

We can calculate the accuracy of rating predictions by means of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Accuracy.

MAE measures the accuracy of predictions by calculating the deviation between the actual score and the predicted score. The MAE value is positively correlated with the prediction accuracy, and the MAE calculation is shown in the formula.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

$$MAE = \frac{\sum_{u,i \in T} \left| \left(p_{(u,i)} - \hat{p}_{(u,i)} \right) \right|}{|T|}$$

where $p_{u,i}$ is the actual rating of item i by user u on the test set, and $\hat{p}_{(u,i)}$ is the predicted rating given by the recommendation algorithm.

RMSE increases the strict filtering of inaccurately predicted item scores, and the evaluation of the recommendation system is more stringent. The RMSE value is negatively correlated with the recommendation accuracy, and the RMSE calculation is shown in the formula.

$$RMSE = \frac{\sqrt{\sum_{u,i \in T} \left(p_{(u,i)} - \hat{p}_{(u,i)} \right)}}{|T|}$$

The accuracy rate (Precision) is calculated as shown in the formula:

$$precision = \frac{\sum_{u \in user} |p_u \cap| t_u}{\sum_{u \in user} |p_u|}$$

Among them, p_u is the recommendation list provided to the user by the recommender system based on the user's behavior on the data training set; and t_u is the user's behavior list on the data test set.

6.3 Analysis of Results

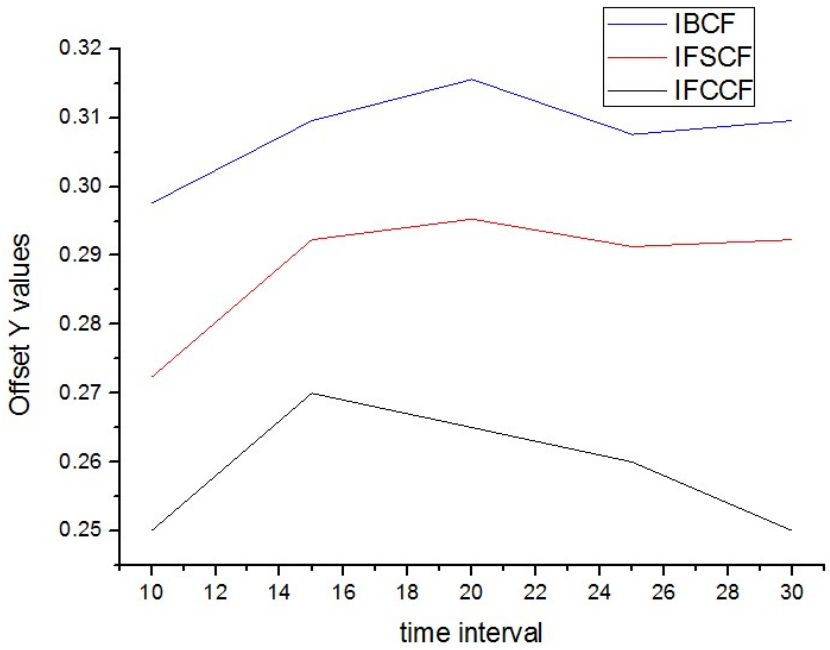

Figure 2. Comparison of the experimental results of the accuracy rate

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

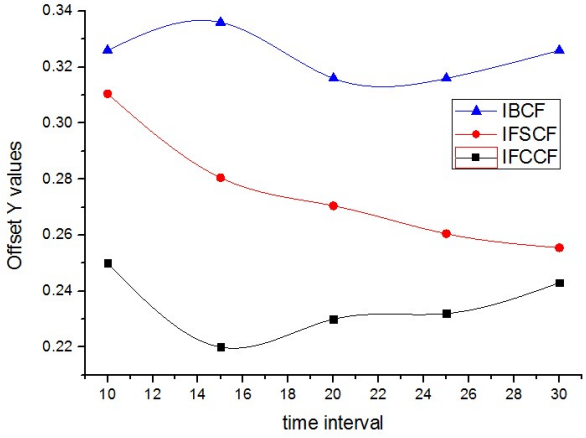

Figure 3. Comparison of recall experiment results

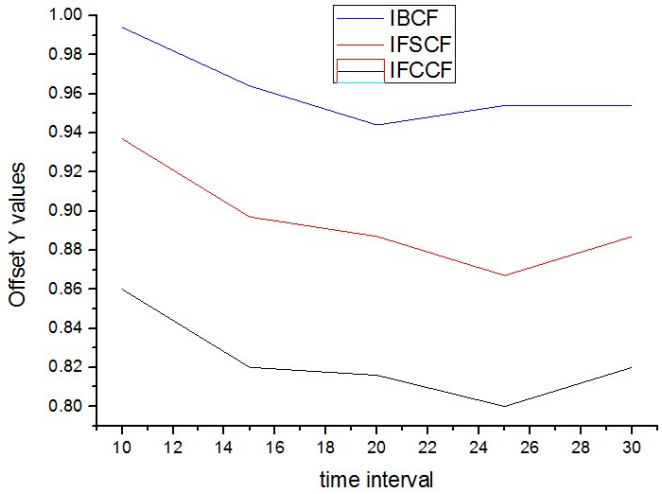

Figure 4. Comparison of MAE experimental results

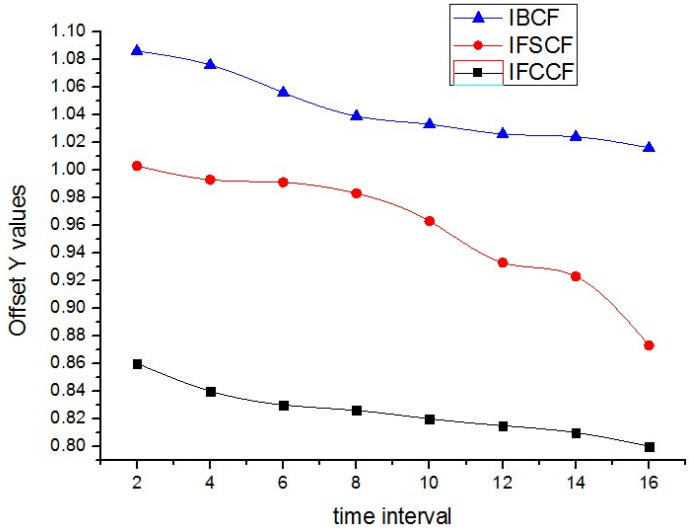

Figure 5. Comparison of recommended algorithm accuracy

In the experiment, the proposed algorithm based on item feature clustering collaborative filtering (IFCCF) and the recommendation algorithm based on item based collaborative filtering (IBCF) and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202209_8(9).0041

the recommendation algorithm based on fuzzy similarity of items collaborative filtering (IFSCF)for comparison, the recommended list length is from 10 to 30 with a step interval of 5. Run the precision, recall and MAE values of each recommendation algorithm, and generate a line graph for comparison. The experimental results are shown in Figures 2-5.

By analyzing the experimental results of accuracy, RMSE and MAE values, it is found that compared with other item-based collaborative filtering algorithms, the method using item feature clustering similarity calculation improves the accuracy and recall rate of recommendation results, and the MAE value lower performance.

7. Epilogue

This paper briefly introduces the project-based collaborative filtering algorithm from the basic idea and its implementation steps, studies and analyzes the calculation of the similarity of key steps in the algorithm and the commonly used recommendation strategies, and verifies the effectiveness of the algorithm by combining example analysis. Evaluation metrics for recommendation algorithms. The item-based collaborative filtering recommendation system can effectively solve the problem of "information overload", but the problems of data sparseness and low scalability in item-based collaborative filtering have not been solved. With the continuous extension and expansion of recommender systems in new fields, new problems will continue to emerge, and the research and solutions for such problems will become a hot spot in the field of recommender system research in the future.

References

[1] Xu Xiangyu, Liu Jianming Collaborative filtering recommendation algorithm based on multi-level item similarity [J] Computer science, 2016.10, Vol.43 No.10.

[2] Ding Heng, Huang Quanzhou Research on personalized tourism recommendation algorithm based on attribute features [J] Smart computers, January 2020, No.

[3] Su Xiaoyun, Zhu Yongzhi Research on hybrid recommendation algorithm based on feature and item nearest neighbor [J] Computer technology and development, September 2019, Vol. 29, No. 9.

[4] Ma Ruimin Research on item based collaborative filtering recommendation algorithm [J] Journal of Jinzhong University, 2021.6, Vol.38, No.3.

[5] Li Zhuan, sun Cuimin Collaborative filtering recommendation algorithm based on item attribute weight [J] Journal of Xinxiang University, March 2019, Vol.36, No.3.

[6] Ma Ruimin Research on item based collaborative filtering recommendation algorithm [J] Journal of Jinzhong University, vol38 no32021.

[7] Wang Yonggui, Shang Geng Deep collaborative filtering recommendation algorithm integrating attention mechanism [J] Computer engineering and application, April 2019, vol.55 No.13.

[8] Cheng Lei, Gao Maoting Hybrid recommendation algorithm combining time weighting and LDA clustering [J] Computer engineering and application, 2019, 55 (11): 160-166.

[9] Zhuang Fuzhen, Luo Dan, he Qing Recommendation algorithm based on integrated local feature learning [J] Computer science and exploration, 2018, 12 (6): 851-858.

[10] Guo Ningning, Wang Baoliang, Hou Yonghong, et al Collaborative filtering recommendation algorithm integrating social network features [J] Computer science and exploration, 2018, 12 (2): 208-217.

[11] Ding Shaoheng, Ji Donghong, Wang Lu Collaborative filtering recommendation algorithm based on user attributes and scores [J] Computer engineering and design, 2015 (2): 487-491.

[12] Cao Junhao, Li zehe, Jiang long, et al A hybrid recommendation algorithm combining collaborative filtering and user attribute filtering [J] Electronic design engineering, 2018, 26 (9): 60-63.