American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-12, Issue-9, pp: 121-127
www.ajer.org
Research Paper Open Access

Research on Recommendation Techniques Based on Alternating Least Squares in Agricultural Products E-commerce

*1 Jin-Ying Wu,
2 Shu-Hao Li
(Zhujiang College, South China Agricultural University, Guangzhou 510900)

Abstract
The importance of personalized recommendation system in the field of agricultural e-commerce cannot be ignored, which provides a more intelligent and convenient platform for farmers and consumers, and helps to optimize the sales and purchase experience of agricultural products. With the rise of agricultural e-commerce, the online trading of agricultural products has become more and more prosperous, and consumers often have difficulty in finding suitable choices among many commodities, which is where personalized recommendation systems can come in handy. By analyzing users' purchase history, preferences and behaviors, the system can accurately recommend agricultural products that meet their tastes and improve shopping efficiency. In this paper, based on the Spark platform using the alternating least squares (ALS) algorithm, the implicit feature vectors of users and items are obtained by decomposing the user-item scoring matrix, so as to realize accurate and efficient item recommendation.

Keywords: *spark platform; recommender system; ALS algorithm*

Date of Submission: 06-09-2023 Date of acceptance: 18-09-2023

I. INTRODUCTORY

In recent years, as the economy continues to develop rapidly, e-commerce platforms are increasingly catering to people's consumption needs, and online sales have become the most popular way to sell agricultural products. And in 2022 the growth of agricultural e-tailing is remarkable. In the first three quarters, the national rural online retail sales amounted to 1497.85 billion yuan, an increase of 3.6% year-on-year. Among them, rural physical goods e-tailing sales amounted to 136.423 billion yuan, up 4.8% year-on-year. The national e-tailing sales of agricultural products amounted to 374.51 billion yuan, up 8.8% year-on-year, and the growth rate was 7.3 percentage points higher than that of the same period last year, and the scale of e-commerce of agricultural products gradually increased[1].

It is obvious that with the rapid development of information technology, the agricultural sector is also gradually ushering in the transformation of digitalization and informationization. The analysis of agricultural e-commerce enterprises indicated that most enterprises only rely on the traditional B2C model for commercial marketing, without targeted marketing to user needs. However, the production, circulation and consumption of agricultural products will generate a large amount of data, such as the user's purchase records, agricultural attributes and characteristics. These data contain valuable information, and how to utilize these data to provide better recommendation of agricultural products has become a challenging and practical issue. Personalized recommendation system, as one of the applications of information technology in the field of agriculture, is able to provide users with personalized and accurate recommendations of agricultural products by analyzing users' historical behaviors and preferences, which not only helps to improve users' shopping experience, but also promotes the sales and promotion of agricultural products.

Currently recommender systems are divided into three main categories[2].

1) Content-based recommendation: content-based recommendation, relying on the similarity of the content itself,

such as bag-of-words characterization of the text into a k-dimensional vector, you can calculate the similarity of the items. Because of the text or image features based on the items themselves, there is no cold start problem, but the general effect is poor, because it is difficult to extract the user preference level of content similarity in the content features, in practice, you will find that you feel that the calculated similarity of the items is very good, but the online effect is very poor.

2) Recommendation based on collaborative filtering: Based on collaborative filtering, using the user behavior recorded by the system, it characterizes the items with users who have consumed the items and calculates the similarity. Compared to content-based generally works better, because the essence is based on co-occurrence, which can uncover a certain level of similarity relationship of items, which is difficult to portray from the content dimension; however, there is a cold-start problem, because if an item has not been consumed by a user, it is impossible to characterize it, and it will not be recommended.

3) Hybrid Recommendation Algorithm[3] : Hybrid recommendation algorithm refers to the mixing of multiple recommendation techniques to compensate for each other's shortcomings. Hybrid methods include simple weighted fusion, switching and mixing of recommendation results, combination of features from different data sources, complex multi-model cascading, feature increment and meta-level mixing. Among them, the most common hybrid recommender system is to combine collaborative filtering recommender methods with other recommender methods, so as to solve the problems of cold start and sparsity. In addition, the advantage of hybrid recommender systems is that they can be customized for specific recommendation scenarios, so as to make reasonable and effective use of additional data information. For example, Konstas et al. utilize the social network information between users of the music website Last.FM to build an effective hybrid recommender system, and Wang et al. introduce knowledge information from knowledge graphs to learn potential knowledge associations in news content to build an efficient hybrid recommender system in news recommendation scenarios.

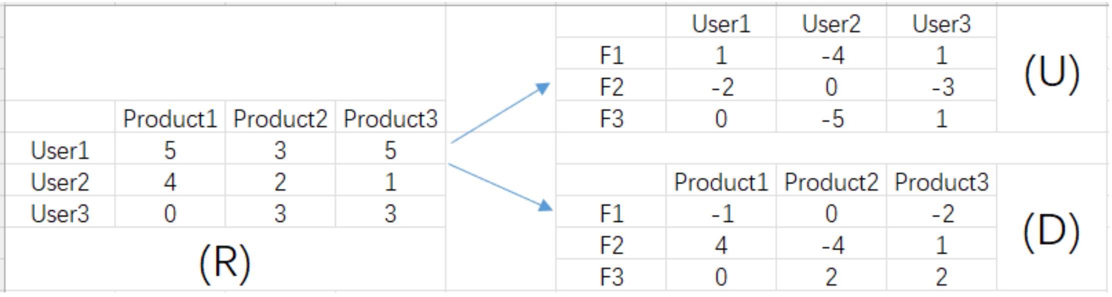
Alternating least squares is a method of collaborative filtering, which is able to learn the implicit feature vectors of users and items by decomposing the user-item rating matrix, so as to realize more accurate item recommendation. Meanwhile, Spark, as a distributed computing framework, provides an efficient platform for processing large-scale data, which can accelerate the computational process of recommendation algorithms and improve the real-time and performance of the system. In recommendation systems, personalized recommendation requires real-time analysis and processing of large-scale user behavior data to provide timely and accurate recommendation results.The distributed nature and memory computing capability of the Spark platform make it an ideal choice for building high-performance personalized recommendation systems. By combining the alternating least squares algorithm with Spark, it is enough to take full advantage of its parallel computing and distributed processing to accelerate the model training and recommendation process.

In the next sections, the principle and implementation steps of the alternating least squares algorithm, the design and implementation of the experiments, and the analysis of the performance and effectiveness of the proposed recommended algorithm through the experimental results will be presented in detail.

II. Collaborative filtering based on ALS algorithm

Alternating Least Squares (ALS) is a collaborative filtering recommendation algorithm. Its goal is to construct an implicit feature matrix using information about the user's ratings of items for recommendation purposes. In this algorithm, the user and the item are mapped into multidimensional factor matrices, respectively, and by multiplying these factor matrices, the user-item rating matrix can be reconstructed to predict the user's preference for the item[4].

The core idea of the ALS algorithm is to process the rating information between users and items through a specific mathematical decomposition method in collaborative filtering recommendation. The goal is to fill in the missing entries in the collaborative filtering matrix and predict more accurately the user's ratings of items. This approach is based on the decomposition of the rating matrix into two smaller matrices representing the implicit features of the user and the item, respectively, in order to represent the user's preferences and the item's characteristics in a low-dimensional space. Suppose there is a rating matrix R with dimensions $m \times n$. By approximate decomposition, this matrix is represented as two smaller matrices U and matrix D, whose latitudes are f*u and f*p, respectively. matrix U and matrix D represent the matrix of the user eigenvectors and the matrix of the item eigenvectors. Figures 1 and 2 show the matrix decomposition:



	Product1	Product2	Product3
User1	5	3	5
User2	4	2	1
User3	0	3	3

(R)

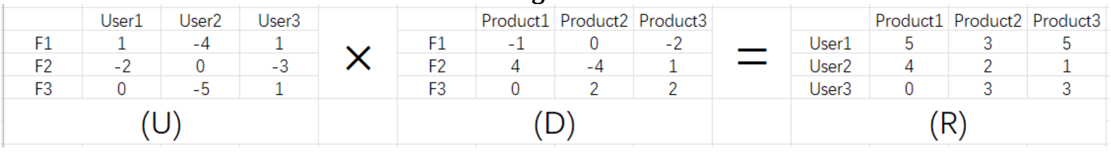
	User1	User2	User3
F1	1	-4	1
F2	-2	0	-3
F3	0	-5	1

(U)

	Product1	Product2	Product3
F1	-1	0	-2
F2	4	-4	1
F3	0	2	2

(D)

Figure 1



	User1	User2	User3
F1	1	-4	1
F2	-2	0	-3
F3	0	-5	1

(U)

×

	Product1	Product2	Product3
F1	-1	0	-2
F2	4	-4	1
F3	0	2	2

(D)

=

	Product1	Product2	Product3
User1	5	3	5
User2	4	2	1
User3	0	3	3

(R)

Figure 2

Specifically, consider a collection of triples consisting of users, items and ratings. This rating matrix is decomposed into the product of the user preference feature matrix (U) and the item feature matrix (D) by means of alternating least squares, where U denotes the implicit preference features of the user and D denotes the implicit features of the item. [5]This decomposition process involves multiple iterations, where the ALS algorithm is performed by the following steps:

1) Considering U as a constant, the D matrix required to approximate the original scoring matrix R is computed by minimizing the loss function.

2) Considering D as a constant, the U matrix required to approximate the original scoring matrix R is computed by minimizing the loss function.

3) Perform steps 1 and 2 alternately until the model converges.

The loss function includes the difference between the predicted score and the actual score, as well as a regularization term to prevent overfitting. The goal of the iterative process is to continuously optimize the U and D matrices so that their product approximates the original scoring matrix. The loss function can be expressed as:

$$J(U,D)=\sum_{i}^{m}\sum_{j}^{n}\left[\left(r_{ij}\_d_j u_i^T\right)^2+\lambda\left(\|u_i\|+\|d_j\|^2\right)\right]$$

where the$\sum_i^m \sum_j^n[]$ Partially denotes for the non-empty terms already in the matrix, i.e., known user-item pairs. These terms are used to compute the error.r_{ij} denotes the true value of a known item in the matrix, i.e., the rating of item j by user i or some other similar value.u_i andd_j denote the implied feature vectors of user i and item j, respectively. These feature vectors are used to represent the representations of users and items in a low-dimensional space, where user feature vectors characterize users and item feature vectors characterize items.$u_i^T d_j$ The inner product of the feature vectors representing user i and item j is used to estimate user i's rating of item j. λ: a regularization parameter that controls the complexity of the model. It introduces a regularization term in the loss function, which helps to prevent overfitting.

This model-based collaborative filtering approach helps to transform sparse user-item rating matrices into dense matrices that can more accurately predict user ratings of items. By alternately optimizing the user and item feature matrices in iterations, the ALS algorithm can approximate these feature matrices to the original rating matrices under certain conditions, leading to more accurate recommendations.

III. Experimental results and analysis

3.1 Experimental cluster environment

In building a distributed computing environment, we chose to use Spark as the computing platform, while integrating the Hadoop distributed system, the operating system using CentOS 6.5 version. The Spark HA high availability cluster deployment based on zookeeper is used, and the hardware running environment of Master node and Slave node are both quad-core CPU and 4GB memory.

Table 1 Experimental cluster environment

software and hardware environment	releases
virtual machine	15.x
operating system	CentOS 6.5
Hadoop	3.0.0
JDK	1.8.0
ZooKeeper	3.5
Master node	Quad-core CPU, 4GB RAM, Qty 1
Slave node	Quad-core CPU, 4GB RAM, Qty 3

3.2 Experimental data set

In this section, the concrete implementation and effectiveness of the ALS algorithm will be demonstrated by means of experiments. In this experiment the dataset used is the Grocery and Gourmet Food sub-dataset of Amazon review data (Amazon e-commerce review dataset), which is commonly used in recommender system research and evaluation, it contains some information such as description, category, price, brand and image features of the product. It contains key data fields such as the reviewer's rating of the product, the reviewer's id, the product's id, the content of the review, the time stamp of the review, etc., and includes a total of 1,143,860 ratings. During the experiment the data is divided into three parts: 60% for training, 20% for calibration and 20% for testing the model.

3.3 Experimental evaluation criteria

RMSE is a commonly used recommender system performance metric to assess the prediction accuracy of a recommendation model. The metric measures the difference between the model's predicted values and the actual observed values, thus reflecting the model's level of prediction error. The tendency in recommender systems is to find the optimal parameter that minimizes the RMSE value, as this represents the best predictive performance of the model under different parameter settings. By comparing the RMSE values under different parameters, it is possible to determine which parameters are effective in improving the prediction accuracy of the model, thus helping to select the optimal combination of parameters to further optimize the recommendation model construction process. The formula is expressed as follows :

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - r_i)^2}{N}}$$

3.4 Experimental analysis

In this paper, experiments are conducted using the Grocery and Gourmet Food dataset, where some parameters need to be set to control the algorithm behavior and performance when using the ALS algorithm for data modeling. Among them, the Rank parameter (also known as the number of latent factors or the number of implied features) specifies the implied feature dimensions of users and items. A lower Rank value reduces the complexity of the model and may lead to inaccurate predictions. Higher Rank values may improve prediction accuracy and may require more computational resources.The Iterations parameter (number of iterations) specifies the number of iteration rounds for the ALS algorithm to perform optimization. A higher number of iterations may improve the fit of the model and also increase the computational time. The appropriate number of iterations depends on the size and complexity of the dataset. The Lambda parameter (regularization parameter) is used to control the complexity of the model and prevent overfitting. Larger Lambda values reduce the complexity of the model and may lead to underfitting. Smaller Lambda values may improve the accuracy of the model and may lead to overfitting[6] .

The above parameters need to be adjusted through the analysis of experiments to obtain the optimal model parameters, which are next verified and obtained through experiments.

3.5 Experimental Flow

This experiment uses java language to write the test cases for the experiment and the following is the execution flow of the experiment.

Data reading and processing: JSON dataset files located at the specified path are first read. Then, Spark is utilized to load this data into a distributed data structure called Resilient Distributed Dataset (RDD). Each row of JSON data is parsed and the user ID, item ID, and rating information is extracted and stored as an object called Rating. Data that fails to be parsed is filtered out to ensure the accuracy of the data.

Data set partitioning: Next, the raw data is partitioned into three parts: the training set, the calibration set and the test set. This division is used for performance evaluation during model training and tuning. 60% of the data is used for training, 20% is used for calibrating the model, and the other 20% is used to test the model's generalization ability.

Model Parameter Optimization: this section is the core part of the experiment, where recommendation models are trained by trying different combinations of model parameters and their performance is evaluated on a calibration set. Nested loops traverse different numbers of hidden factors (Rank), regularization parameters (Lambda) and iterations (Iterations). For each set of parameters, the model is trained on the training set using the ALS algorithm, and then the Root Mean Square Error (RMSE) is calculated on the calibration set. This process aims to find the set of parameters that minimizes the RMSE on the calibration set, which will represent the best model configuration.

Selecting the best model: During the parameter search, the smallest check-set RMSE value, and the model parameters associated with it, are continuously tracked. Once the parameter search is complete, the model with the smallest check-set RMSE value is selected as the best model for final testing and recommendation.

Testing and Recommendation: ultimately, the best model found in the calibration phase is used to evaluate on the test set and calculate the RMSE of the test set. the selected item (via hashed item ID) is then passed to the best model for item recommendation. The model returns some user rating predictions for the item, and eventually outputs these predictions as recommendations that are displayed to the user.[7]

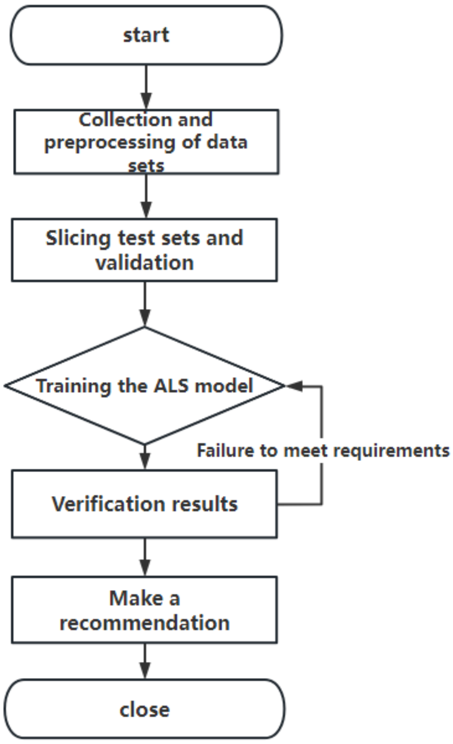
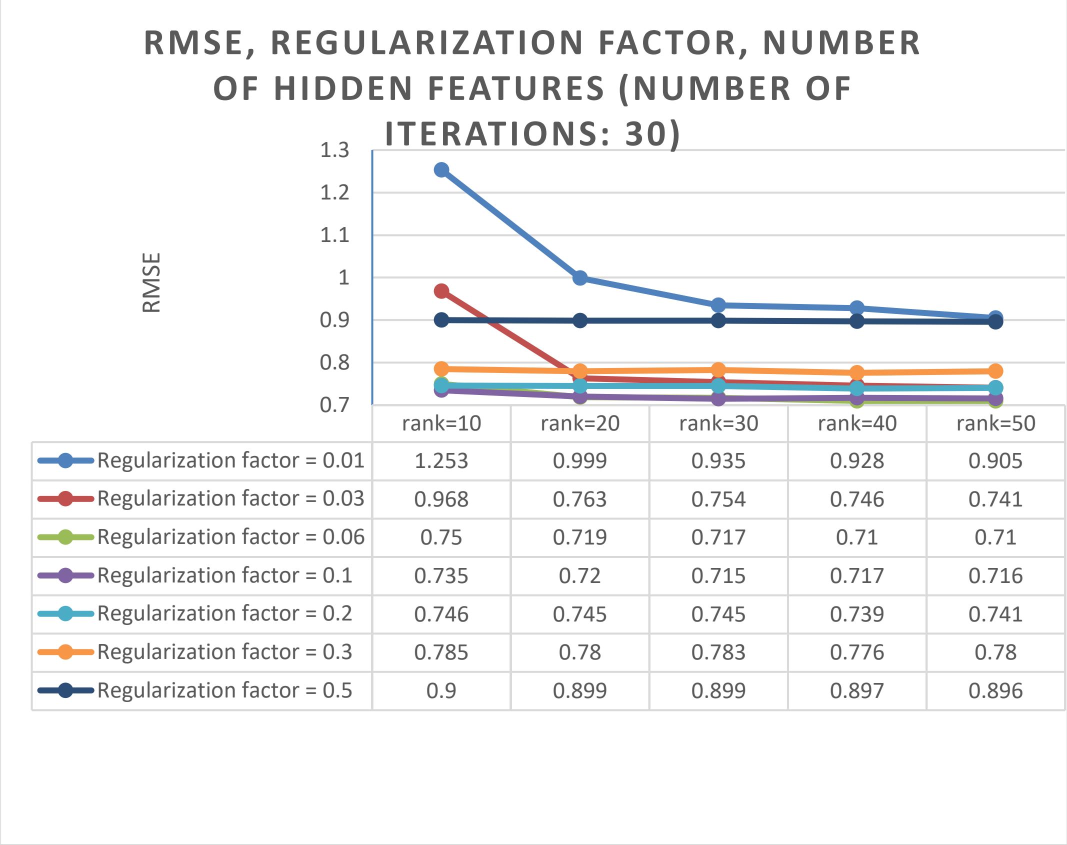

start
Collection and preprocessing of data sets
Slicing test sets and validation
Training the ALS model
Failure to meet requirements
Verification results
Make a recommendation
close

Fig. 3 Flowchart of agricultural products recommendation based on the least squares method

3.6 Experimental results

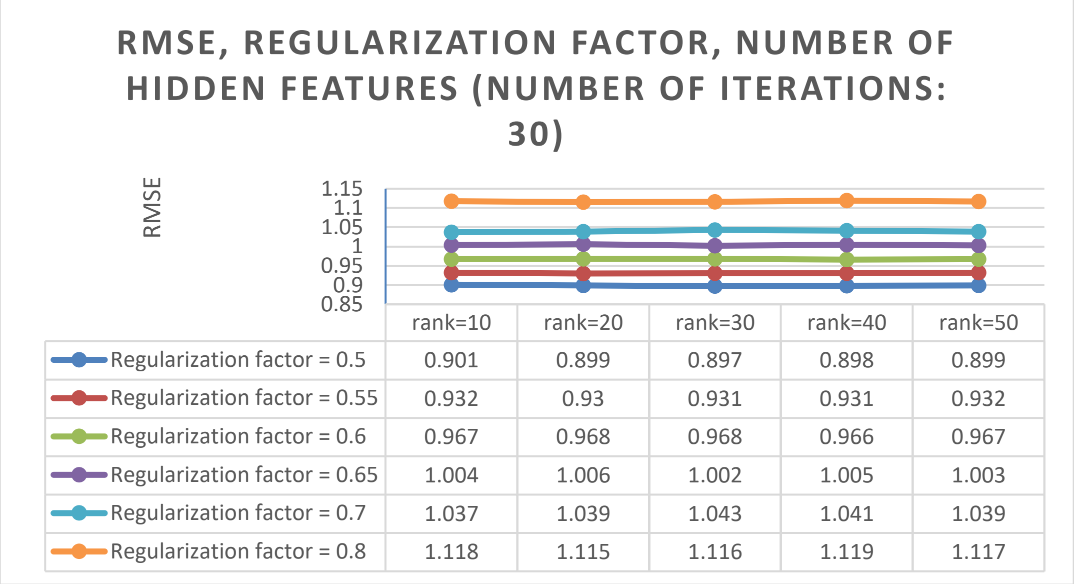
When trying models with different parameter combinations, the evaluation was performed using the calibration set and the root mean square error (RMSE) was calculated as an evaluation metric. After comparison, two sets of experimental results were selected as samples to be analyzed and the best model in Fig. 4 has the parameter combinations of Rank=40, Lambda=0.06, and Iterations=30.This set of parameters gave a good RMSE value of 0.7104519480666466 on the calibration set, whereas on the corresponding test set, the RMSE value was 1.175489738024379, which is a large difference between the two. It shows that although the model performs well on the calibration set, it performs poorly on the unseen test data and overfitting occurs. The model is too well adapted to the training data on the training set and cannot generalize well to the test data.

Figure 5 shows the results of another set of experiments, which has an RMSE value of 0.8977294136340424 on the test set and 0.9928747019456298 on the corresponding calibration set.The performance of the best model on both the calibration set and the test set is relatively close to each other, and it can be assumed that the model has a good generalization ability. Being able to show good prediction performance also on unseen test data without serious overfitting or underfitting problems means that the model has a better prediction performance on the test set and is able to get a more accurate prediction.



	rank=10	rank=20	rank=30	rank=40	rank=50
Regularization factor = 0.01	1.253	0.999	0.935	0.928	0.905
Regularization factor = 0.03	0.968	0.763	0.754	0.746	0.741
Regularization factor = 0.06	0.75	0.719	0.717	0.71	0.71
Regularization factor = 0.1	0.735	0.72	0.715	0.717	0.716
Regularization factor = 0.2	0.746	0.745	0.745	0.739	0.741
Regularization factor = 0.3	0.785	0.78	0.783	0.776	0.78
Regularization factor = 0.5	0.9	0.899	0.899	0.897	0.896

Fig. 4 RMSE values for different parameters



	rank=10	rank=20	rank=30	rank=40	rank=50
Regularization factor = 0.5	0.901	0.899	0.897	0.898	0.899
Regularization factor = 0.55	0.932	0.93	0.931	0.931	0.932
Regularization factor = 0.6	0.967	0.968	0.968	0.966	0.967
Regularization factor = 0.65	1.004	1.006	1.002	1.005	1.003
Regularization factor = 0.7	1.037	1.039	1.043	1.041	1.039
Regularization factor = 0.8	1.118	1.115	1.116	1.119	1.117

Fig. 5 RMSE values for different parameters

IV. Concluding Remarks

The purpose of this paper is to explore and analyze the recommendation model based on alternating least squares, by applying collaborative filtering algorithm, combined with a large-scale rating dataset, to construct a model that can more accurately predict users' recommendations for items. The recommendation model is finally applied to personalized produce recommendation through several experiments of parameter tuning and performance evaluation. There are still some limitations in the recommendation system proposed in this paper, such as the problem of insufficient data in the cold-start phase, which leads to poor accuracy of the model recommendation results, and this problem requires the use of collaborative filtering variants in subsequent practical development, such as item-based collaborative filtering, which utilizes the similarity information between agricultural products, so as to carry out the recommendation. On the basis of this paper, the recommendation in the cold start phase is combined with the subsequent personalized recommendation to achieve a comprehensive and effective recommendation strategy.

REFERENCES

[1]. Department of Commerce Website 2022 Report

[2]. QIN Chuan, ZHU Hengshu, ZHONG Fuzhen, GUO Qingyu, ZHANG Qi, ZHANG Le, WANG Chao, CHEN Enhong, XIONG Hui. A research review on knowledge graph-based recommender system[J]. Science in China:Information Science,2020,50(07):937-956.

[3]. R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Model User-Adap Inter, vol. 12, no. 4, pp. 331 -370, Nov. 2002, doi: 10.1023/A:1021240730564.

[4]. Jiao Jian. Research on collaborative filtering recommendation algorithm based on Spark[J]. Computer Programming Skills and Maintenance,2020,000(3):40-4173

[5]. WANG Cheng,TANG Jianguo.Application of ALS algorithm in dish intelligent recommendation system[J]. Fujian Computer,2023,39(3):78-81

[6]. LI Ning,ZHANG Zilei,WU Gang,ZHENG Tao. Research on user model in personalized movie recommendation system[J]. Computer Application and Software,2010,027(12):51-54

[7]. ZHONG Zhifeng,ZHOU Dongping,ZHANG Yan,XIA Yifan. Research on hybrid recommendation model based on least squares[J]. Modern Electronic Technology,2022,45(17):123-128

Funded Project

2022 National Innovation Training Project: "Design and Implementation of Agricultural Trade Promotion Trading Platform Based on Data Collection and Push" (Project No. 202212623003)

Author Bio

Jin-Ying Wu (2003.07), male Han nationality Qingyuan, Guangdong, undergraduate studying Computer Science and Technology.

Shu-Hao Li (2000.03), male Han nationality Qingyuan, Guangdong, undergraduate studying Computer Science and Technology.