Analysis and Prediction of E-commerce User Behavior Based on Collaborative Filtering and Clustering Algorithm

Xiao Zhang, Mingyang Song

***Abstract*—The rapid development of e-commerce has led to exponential growth in user behavior data. Against this backdrop, how to refine personalized recommendations and predict user preferences has become an important requirement for e-commerce platforms. Therefore, a hybrid model based on collaborative filtering and clustering algorithms is proposed to analyze and predict e-commerce user behavior. Firstly, the K-means clustering algorithm is optimized. Secondly, the data processing efficiency is improved on the basis of the genetic algorithm. Finally, the improved Slope one algorithm is introduced for personalized recommendation. From the results, the proposed hybrid model had better iteration efficiency, reaching a stable state in the training and testing sets after 48 and 49 iterations, respectively. The recommendation precision, recall, and F1 were 98.03%, 97.63%, and 97.85%, respectively. In practical applications, the recommendation accuracy on five different e-commerce platforms exceeded 95%, and the required time was as low as 0.6s. The proposed model can further provide effective user behavior analysis and personalized recommendation services for e-commerce platforms.**

***Index Terms*—Slope one, K-means, User behavior, E-commerce, Personalized recommendation**

I. Introduction

With the rapid progress of e-commerce, online shopping has become an indispensable part in daily lives. According to the China E-commerce Report, the transaction volume of e-commerce in China has exceeded 30 trillion RMB in 2023, with a year-on-year increase of 14.5% [1-2]. In this context, e-commerce platforms not only need to enhance the quality of goods and service levels, but also deeply analyze user behavior patterns to provide personalized recommendations and optimize user experience. When meeting the personalized recommendation needs of users, user behavior analysis and prediction have gradually become one of the key technologies to improve the competitiveness of e-commerce platforms [3-4]. In e-commerce, some methods for analyzing and predicting household behavior have been proposed. Traditional user behavior analysis methods mainly include rule-based methods and statistical model-based methods, which have achieved certain results in the early stage. However, with the rapid growth of user numbers and behavioral data, traditional methods have shown certain limitations in dealing with large-scale data and complex behavioral patterns [5]. Machine learning and data mining methods have been widely applied to analyze user behavior in e-commerce. Collaborative filtering algorithm and clustering algorithm are the two most commonly used methods among them. Collaborative filtering algorithm recommends products by analyzing users' historical behavior and similar user behavior, with high accuracy and interpretability. However, traditional collaborative filtering algorithms have poor performance when facing sparse data and cold start problems. Clustering algorithm discovers user groups with similar behavioral characteristics by grouping them, which are used for market segmentation and personalized recommendations. However, the interpretability and real-time performance still need to be improved [6]. Therefore, the research aims to combine the advantages of collaborative filtering and clustering algorithm to propose a new e-commerce user preference recommendation and prediction model to deal with the data sparsity and cold start in existing methods. This study innovatively incorporates the Genetic Algorithm (GA) to enhance the randomly selected cluster centers in the K-means Clustering Algorithm (K-means). Then, a weighted method is used to optimize the Slope one collaborative filtering algorithm, thereby improving the data clustering and recommendation performance. The final recommendation prediction model is expected to provide a more effective user behavior analysis and prediction method for e-commerce platforms, while also helping to enhance user experience and platform competitiveness. The unique contribution is to propose a new hybrid recommendation model, GA-K-means-slope One, which systematically solves the limitations of traditional recommendation methods in dealing with data sparsity and cold start problems by combining K-means and improved collaborative filtering recommendation algorithm. GA is innovatively introduced to optimize the center selection of K-means, improve the clustering effect and algorithm stability. Meanwhile, weighted Slope one algorithm is adopted to reflect user preferences more accurately. In addition, the model has shown superior recommendation accuracy, recall rate and recommendation speed in practical applications, which provides efficient and personalized user behavior analysis and recommendation services for e-commerce platforms, enhances user stickiness and platform competitiveness, and makes it have important business value and wide application prospects.

Manuscript received September 9, 2024; revised March 27, 2025.

Xiao Zhang is a lecturer of City University of Zhengzhou, Zhengzhou 452370, China. (corresponding author; e-mail: 17513206575@163.com).

Mingyang Song is a product manager of Pingan One Account Intelligent Technology Co., Ltd, Shenzhen 518000, China. (e-mail: 17600118605@163.com).

II. Related Works

Collaborative filtering is a classic recommendation technique that has been extensively practically taken. Given its simplicity, high precision, and wide applicability to various data types, it has been extensively applied in e-commerce, news, video, etc. Zhang S et al. designed a multi-neural collaborative filtering attraction recommendation strategy to recommend tourist attractions. Various tourism backgrounds and trajectory backgrounds were modeled to obtain the feature representation of tourists. A neural network was taken to project the attractions into the feature space. This architecture was used to recommend different attractions [7]. To identify and predict unobserved target genes, Lim H et al. built a weighted pulse neighborhood regularization three-factor classification collaborative filtering method. The model outperformed other matrix factorization approaches, which was taken to tissue-specific data [8]. Zhang J et al. developed a collaborative filtering method for information recommendation in e-commerce. Adjacent sets were decomposed into probability matrices. The algorithm outperformed other common methods [9]. Sun G et al. designed a multi-level music audio database based on big data and collaborative filtering. The search efficiency and accuracy exceeded other databases [10].

Personalized recommendation system is an advanced intelligent platform based on massive data mining. It mainly consists of three modules: user modeling, recommendation object modeling, and recommendation algorithm module. Some scholars have conducted research on optimizing e-commerce personalized recommendation systems. Guo S et al. improved the personalized recommendation algorithm by combining blockchain technology to enhance the personalized recommendation effect of entrepreneurial service information. A corresponding revenue model was established to derive the optimal weights for each incentive mechanism. The results indicated that the blockchain-based personalized recommendation system for entrepreneurial service information provided more reliable information [11]. Pei H et al. proposed a personalized cloud platform service recommendation model that combined long-term preferences and instant interests. The model utilized a dual attention mechanism to assign different weights to long-term preferences, and outputted predictions. The model was more accurate and efficient compared with the other five comparison methods [12]. Lin Y et al. proposed a multi-scale enhanced personalized recommendation model for online learning course recommendation services. An attention-based multi-variable perceptron recommendation model was developed. The model used relevant parameters and data structures to control the loss optimization direction. The model was better than other models on the MOOC platform dataset [13]. Zhen Y et al. considered various preference characteristics of users to design an adaptive preference entity recommendation method. An adaptive dual domain transfer model was proposed for item main domain and social domain that met users' personalized needs. The explicit and implicit features were mined through user feedback migration method. The entity rating matrix was embedded to address the sparse user feedback information [14].

Although personalized recommendation service systems have matured, problems such as data sparsity and cold start still exist. Therefore, an improved collaborative filtering algorithm is proposed to accurately analyze heterogeneous multi-dimensional data, ensure the comprehensiveness of e-commerce customer information and the effectiveness of e-commerce personalized recommendation, and promote the development of the information recommendation field.

III. E-commerce User Behavior Recommendation and Prediction Combining Collaborative Filtering and Clustering Algorithms

Firstly, the traditional K-means is optimized by incorporating the GA to change the selection method of its clustering centers. Then, combined with the weighted Slope One collaborative filtering algorithm, the final e-commerce user behavior prediction and preference recommendation model is constructed.

A. Optimization Design of K-means

K-means clustering is an extensively applied unsupervised learning algorithm that can iteratively partition a set of sampling points into subsets of multiple classes [15]. In the analysis and prediction of e-commerce user behavior, K-means is commonly applied to cluster user behavior data to discover the characteristics and behavior patterns of different user groups. Dividing user data into several subsets can better understand users' needs and preferences, thereby providing support for personalized recommendation systems and marketing strategies. The clustering process of K-means is shown in Fig. 1.

In Fig. 1, first, K-means randomly selects K initial cluster centers. Second, each data point is assigned to the cluster to which the nearest cluster center belongs. Then, the center of each cluster is calculated and updated to the average value of all data points within the cluster. Next, data points are reassigned based on the new center and updated the center again, repeating this process until the center position no longer undergoes significant changes. Finally, the K-means reaches a convergence state and forms stable K clustering results. The error function of K-means is shown in Equation (1) [16].

$$P = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - c_j \right\|^2 \tag{1}$$

In Equation (1), i and j signify the data and clusters. n and k represent the range for i and j, respectively. x_i represents the i-th data. c_j signifies the center point of the j-th cluster class. P represents the error function of K-means. Considering the poor performance of traditional K-means in processing e-commerce big data, long running time and low computational efficiency may exist. Therefore, a new hybrid algorithm called GA-based K-means (GA-K-means) is designed by combining GA and K-means to optimize the efficiency and effectiveness of data mining. The operation process of GA is displayed in Fig. 2.

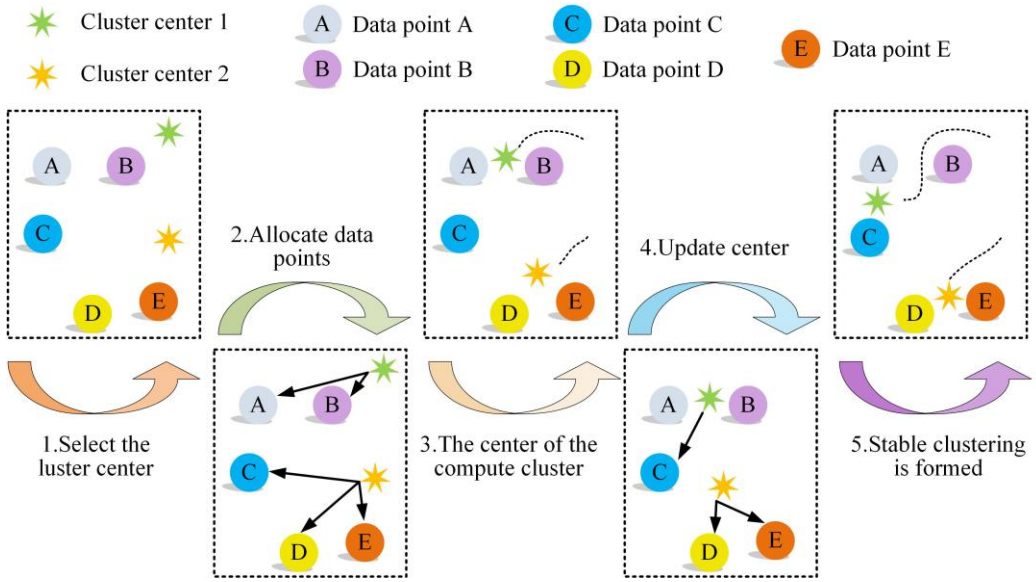


Fig. 1. K-means clustering process

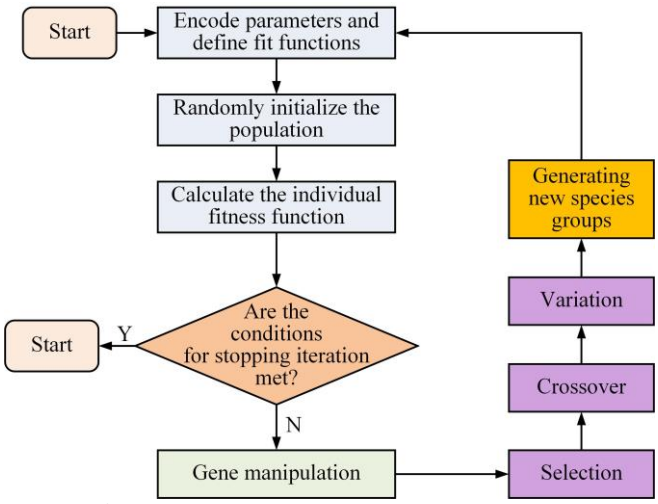


Fig. 2. GA flowchart

The running process of the GA is displayed in Fig. 2. First, an initial population is randomly produced, with each individual representing a possible solution. Second, the individual fitness in the population is assessed and calculated. Then, based on the fitness value, individuals with higher fitness are selected as parents for crossover operation, generating new offspring individuals, and mutating these offspring individuals to introduce diversity. Next, the fitness of the newly generated offspring individuals is evaluated, and then individuals from the old population are replaced according to certain strategies to form a new population. Finally, whether the termination condition is met is determined. Conversely, the optimal solution is output. Otherwise, the process returns to the selection step to continue iterating. To optimize the cluster center selection in the K-means algorithm, GA is introduced to perform floating-point encoding on the center points and optimize them based on the fitness function to improve the stability of the algorithm.

The dataset obtained from the e-commerce system is $D=\left\{d_j,\left|j=1,2,3,\cdots,n\right.\right\}$. The set of cluster centers in D is $N_k=\left\{C_i,\left|i=1,2,3,\cdots,k\right.\right\}$. The sum of standard deviations is used to determine the clustering effect of the data, as shown in Equation (2).

$$f\left(P_k\right)=\sum_{i=1}^{k}\sum_{P\subset C_i}\left|P-m_i\right|^2 \tag{2}$$

In Equation (2), $f(\cdot)$ is the criterion function. P_k represents the sum of errors for all clusters in K-means. C represents the center of the cluster. m represents the average value of C, as shown in Equation (3).

$$m_i=\frac{1}{t_i}\sum_{P\subset C_i}P \tag{3}$$

In Equation (3), t_i signifies the data points in C_i. The similarity between each data point is calculated using Euler's equation, as shown in Equation (4).

$$d\left(x_i,m_j\right)=\sqrt{\sum_{i=1}^{d}\left(x_{it}-m_{jt}\right)^2} \tag{4}$$

In Equation (4), d represents the similarity between data points. m_j represents a multidimensional point at the center of the cluster. t signifies the dimension. x_{it} signifies the value of data x_i in the t dimension. m_{jt} represents the value of cluster center m_j in the t dimension. The Euclidean distance between x_i and m_j is used to measure the similarity between data points and cluster centers in multidimensional space. According to Equations (1) to (4), taking Equation (1) as the criterion function can reduce the efficiency of business data mining.

Meanwhile, due to the high randomness in selecting cluster center points, clustering errors will be significant. To reduce clustering errors, the mutation operation in GA is introduced into the K-means algorithm. Fig. 3 displays the running process of GA-K-means.

In Fig. 3, first, the data of e-commerce users is collected and preprocessed. Second, the K-means is applied to cluster the collected e-commerce data and compute the similarity between each data point. If the similarity is low, the GA is used to re-select the clustering center and perform genetic mutation operation on it. When the similarity between data points is high and the calculated fitness value is less than the set convergence coefficient, the clustering result is output.

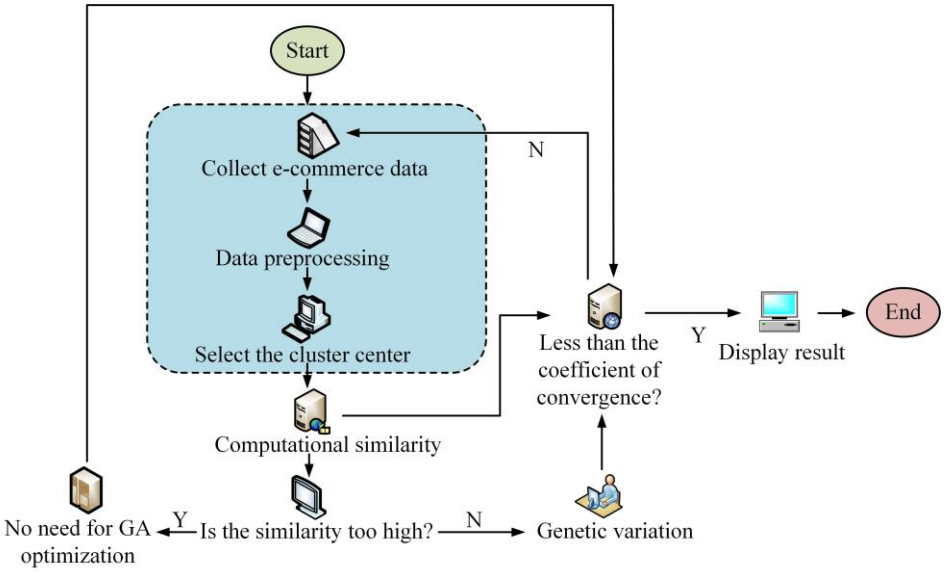


Fig. 3. Running flow diagram of GA-K-means

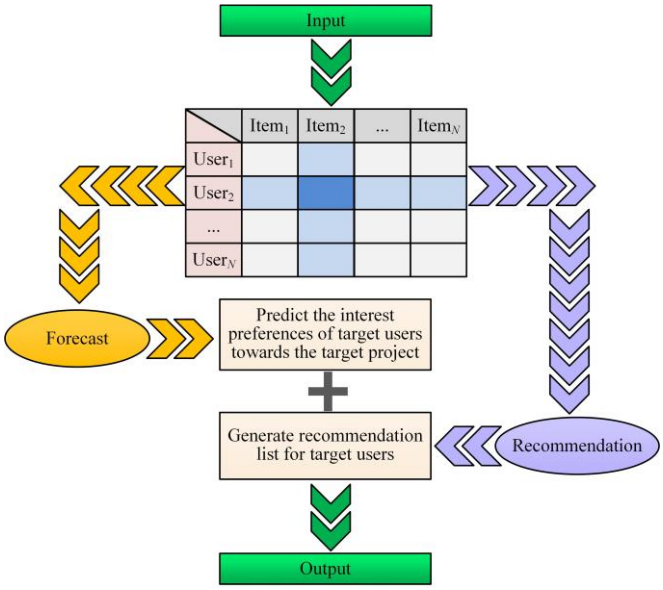


Fig. 4. Collaborative filtering recommendation process

The fitness calculation for GA-K-means is shown in Equation (5).

$$f\left(R_i\right)=E_{\max}-E\left(R_i\right) \tag{5}$$

In Equation (5), R represents the population. $E_{\max}$ represents the maximum chromosome variance value of R. In the GA-K-means iteration process, the roulette wheel algorithm is used to calculate the selection probability of data, as shown in Equation (6).

$$P_i=\frac{f\left(R_i\right)}{\sum_{i=1}^{n} f\left(R_i\right)} \quad \left(i=1,2,\cdots n\right) \tag{6}$$

In Equation (6), P_i represents the probability of data i being selected.

B. Construction of an E-Commerce User Recommendation Model Combining GA-K-means and Improved Slope One

In addition to using GA-K-means to analyze various data of e-commerce users, the study also adopts collaborative filtering algorithm to build an e-commerce user recommendation model, aiming to make more reasonable recommendation predictions based on the preferences and habits of various e-commerce users. Collaborative filtering algorithm is a classic and commonly used recommendation algorithm, which generally determines the user's preference direction through historical data analysis to find other user collections that are similar to the preference, or analyze item features to find the most similar item collection. Fig. 4 displays the collaborative filtering recommendation process.

In Fig. 4, each user's rating for each item is represented in matrix form, which can also represent the preference of different users for different items. The data in the algorithm prediction and recommendation matrix is taken to obtain the recommendation list of the target user. Among various collaborative filtering algorithms, Slope One is a simple and practical collaborative filtering recommendation algorithm for rating matrices. Its process involves two steps. The first is to calculate the average deviation, and the second is recommendation. This study combines GA-K-means and Slope One to build the final recommendation model. The example of the user-item rating matrix in Slope One is displayed in Table I.

In Table I, to predict the rating of *Userl* on *Itemm*, it is first necessary to calculate the average deviation of *Itemj* and *Itemm* from *Userj* and *Itemj*.

According to a series of calculations, the average deviation value in the matrix is 1.

TABLE I
USER-PROJECT RATING MATRIX

/	*Itemi*	*Itemj*	*Itemk*	*Iteml*	*Itemm*
Useri	5	1	1	/	2
Userj	4	2	2	3	3
Userk	3	/	2	5	2
Userl	/	4	3	/	?

Therefore, the prediction score of *Userl* for *Itemm* is 4.5. The average deviation calculation is shown in Equation (7) [17-18].

$$dev(i', j') = \frac{\sum_{u \in U(i') \cap U(j')} \left(S_{ui'} - S_{uj'} \right)}{\left| U(i') \cap U(j') \right|} \quad (7)$$

In Equation (7), S represents the rating value. $S_{ui'}$ signifies the rating of u for item i'. $S_{uj'}$ represents the rating value of u for item j'. $U(i')$ and $U(j')$ signify sets of users who rate i' and j'. *dev* represents the average deviation value. The recommended generation process is shown in Equation (8).

$$pre(u, j') = \frac{\sum_{j' \in S(u) - \{i'\}} (S_{ui'} - dev(i', j'))}{S(u) - \{i'\}} \quad (8)$$

In Equation (8), $S(u)$ represents the item set rated by u. $S(u) - \{i'\}$ represents at least one item set i' that has been rated by u. *pre* represents the predicted score. Due to the fact that the Slope One algorithm does not take into account the impact of item audience size, there may be significant deviations between the recommendation results and the actual situation when there is a large difference in project audience size. Therefore, a weighted value is adopted based on the Slope One algorithm for optimization. The weighted method is used to make the recommendation results more accurate. The weighting process is shown in Equation (9).

$$pre(u, j') = \frac{\sum_{j' \in S(u) - \{i'\}} (S_{ui'} - dev(i', j') * \left| S_{i'j'} \right|)}{\sum_{j' \in S(u) - \{i'\}} \left| S_{i'j'} \right|} \quad (9)$$

In Equation (9), $\left| S_{i'j'} \right|$ signifies the set of users who collectively rate items i' and j'. As the number of items increases, the weighted Slope One algorithm cannot guarantee that every user rates all items, which can affect the recommendation accuracy. To address this issue, a method combining item rating similarity and item semantic similarity is proposed to comprehensively compute the item similarity. When calculating the similarity of item ratings, Pearson correlation coefficient is taken to more accurately reflect the similarity between items, as displayed in Equation (10).

$$RatingSim_{i', j'} = \frac{\sum_{u \in U_{i', j'}} \left(R_{ui'} - \bar{R}_{i'} \right) * (R_{uj'} - \bar{R}_{j'})}{\sqrt{\sum_{u \in U_{i', j'}} \left(R_{ui'} - \bar{R}_{i'} \right)^2} * \sqrt{\sum_{u \in U_{i', j'}} \left(R_{uj'} - \bar{R}_{j'} \right)^2}} \quad (10)$$

In Equation (10), $\bar{R}$ represents the mean score. The value of $RatingSim_{i', j'}$ is positively correlated with the item similarity. The semantic similarity of different items is shown in Equation (11).

$$SemSim(i', j') = \frac{\left| N(i') \cap N(j') \right|}{\left| N(i') \cup N(j') \right|} \quad (11)$$

In Equation (11), $N(i')$ and $N(j')$ represent the semantic description sets of i' and j', respectively. After determining the item similarity ratings and semantic similarity, the linear weighting is applied to calculate the mixed item similarity, as shown in Equation (12).

$$Sim(i', j') = \gamma * RantingSim(i', j') + (1 - \gamma) Sem\mathrm{Sin}(i', j') \quad (12)$$

In Equation (12), the Person correlation coefficient is first used to compute the item rating similarity $RatingSim_{i', j'}$. Then, the item semantic descriptor is used to calculate the item semantic similarity $SemSim(i', j')$. These two are linearly weighted and combined to obtain the mixed item correlation $Sim(i', j')$. After obtaining the method for calculating item similarity, the study integrates the GA-K-means into the weighted Slope one to construct an e-commerce user data analysis and preference recommendation prediction model. The final item rating deviation is shown in Equation (13).

$$dev(i', j') = \frac{\sum_{u \in U(i') \cap U(j')} (S_{ui'} - S_{uj'}) * match(u, v)}{\left| U(i') \cap U(j') \right|} \quad (13)$$

In Equation (13), *match* represents the clustering matching degree between u and v. Adding user matching degree to the rating deviation calculation equation can significantly improve the recommendation effect compared with only calculating rating deviation. Considering the matching degree between users, the algorithm can more accurately reflect users' true preferences. This process is shown in Equation (14).

$$pre(u, j') = \frac{\sum_{j' \in S(u) - \{i'\}} (S_{ui'} - dev(i', j')) * sim(i', j')}{S(u) - \{i'\}} \quad (14)$$

In Equation (14), $pre(u, j')$ represents the predicted rating of user u for item j'. $S(u)$ represents the collection of rating items for u. $S(u) - \{i'\}$ represents item i' and a set of items that are known to have been rated by u. $sim(i', j')$ signifies the similarity between items i' and j'. The e-commerce user data analysis and preference recommendation prediction model combining GA-K-means and weighted Slope one is referred to as GA-K-means-Slope one. The recommendation prediction process of GA-K-means-Slope one is displayed in Fig. 5.

IV. Performance Analysis of E-Commerce User Recommendation Model Based on GA-K-Means-Slope One

To verify the preference recommendation and prediction performance of the GA-K-means-Slope one model for e-commerce users, the study first tests the benchmark performance of the GA-K-means-Slope one. Then, the actual recommendation and prediction performance of the model are compared.

A. Benchmark Performance Test Results of the GA-K-means-Slope One

The data required for the experiment is sourced from the Amazon e-commerce platform, collected through API calls and web crawling technology, including user personal information, purchase records, browsing history, shopping cart information, and other related data. To ensure the experiment accuracy, user data with purchase records and browsing history less than 15 times, as well as user data with no recent activity behavior, are excluded. The final dataset consists of 4650 users, 89,214 purchase records, and 53,246 browsing information, which is separated into training and testing sets in an 8:2.

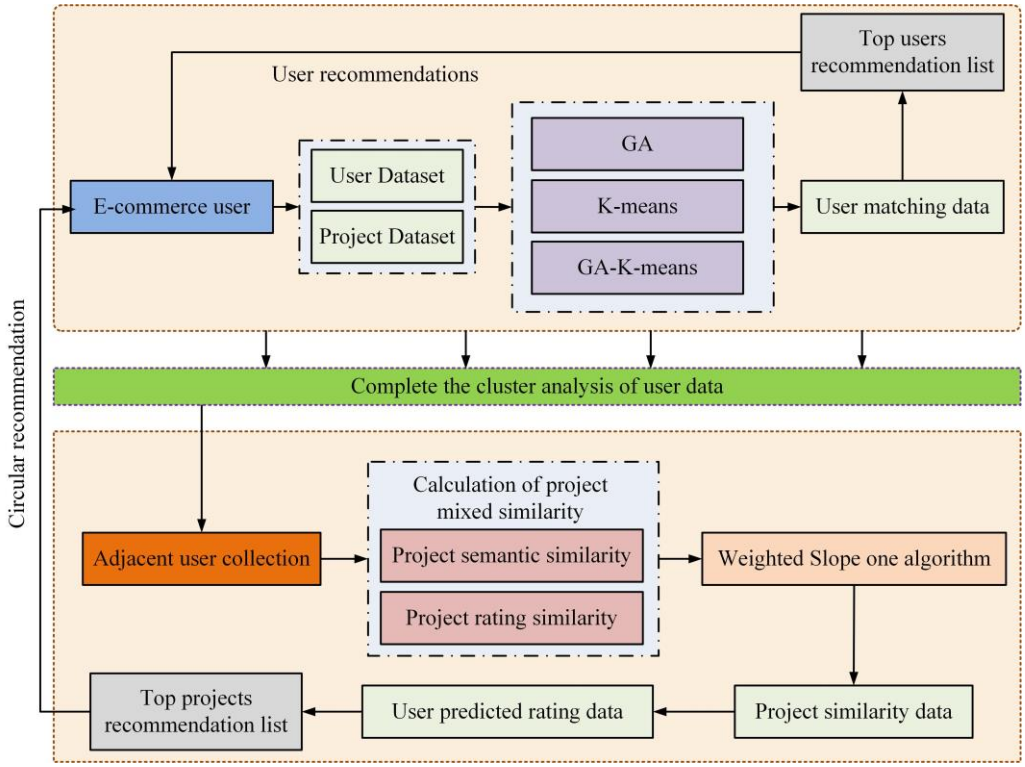


Fig. 5. Recommendation prediction flow chart of GA-K-means-Slope one model

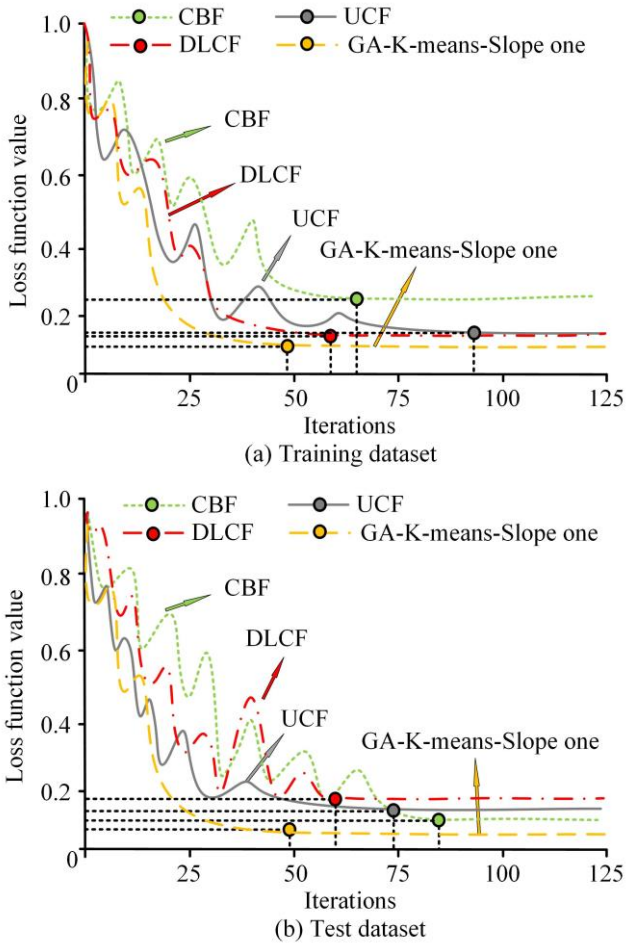


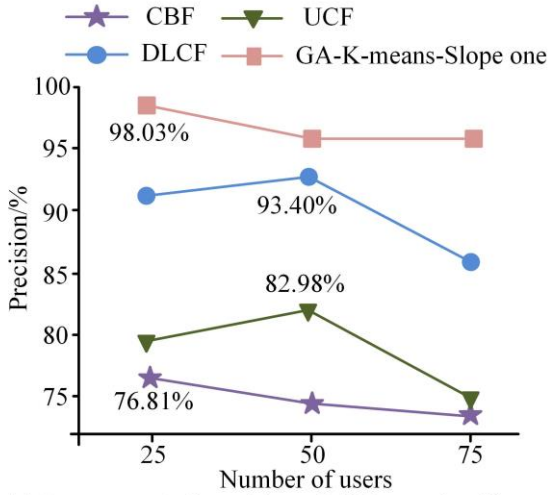
Fig. 6. Loss function curves for different algorithms

Three commonly used recommendation algorithms for comparison are Content-based Recommendation (CBF), User-based Collaborative Filtering (UCF), and Collaborative Filtering Recommendation based on Deep Learning (DLCF). Firstly, the loss function changes of each recommendation algorithm during training and testing are compared, as displayed in Fig. 6.

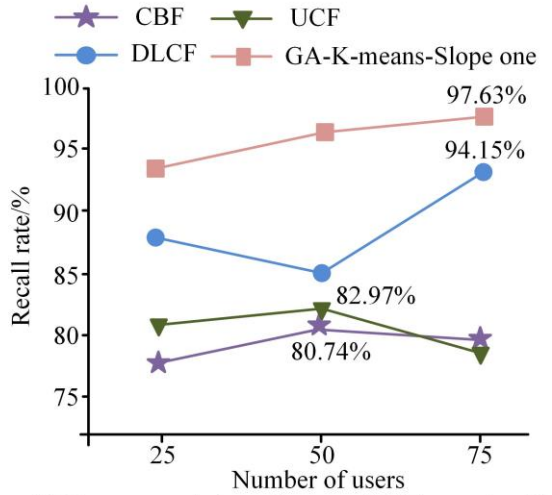
Figs. 6(a) and 6(b) present the changes in loss functions of four algorithms, namely CBF, UCF, DLCF, and GA-K-means-Slope one, in the training and testing sets. From Figs. 6(a) and 6(b), as the iteration increased, all four algorithms experienced varying degrees of fluctuations before reaching a stable state. In Fig. 6(a), CBF, UCF, DLCF, and GA-K-means-Slope one reached a stable state after 63, 92, 57, and 48 iterations, respectively. In Fig. 6(b), CBF, UCF, DLCF, and GA-K-means-Slope one reached a stable state after 81, 74, 60, and 49 iterations, respectively. Combining K-means and the improved Slope One algorithm, the proposed algorithm uses GA to optimize cluster centers, reducing the impact of initial cluster center randomness on the results, and effectively reducing the training complexity and volatility of the model. This technical advantage enables GA-K-means-Slope One to have higher computational efficiency and stability when processing large-scale data. The benchmark recommendation performance of four algorithms is tested. The recommendation precision, recall rate, and F1 of each algorithm are displayed in Fig. 7.

Figs. 7(a), 7(b), and 7(c) show the recommendation precision, recall rate, and F1 of four algorithms under different recommendation numbers. From Fig. 7, the benchmark recommendation performance of different algorithms varied with the number of recommended individuals. The highest recommendation precision values of CBF, UCF, DLCF, and GA-K-means-Slope one were 76.81%, 82.98%, 93.40%, and 98.03%, respectively. The highest recommendation recall rates were 80.74%, 82.97%, 94.15%, and 97.63%, respectively. The highest recommendation F1 values were 78.24%, 85.96%, 93.62%, and 97.85%, respectively. The results indicate that this research method has advantages on personalized

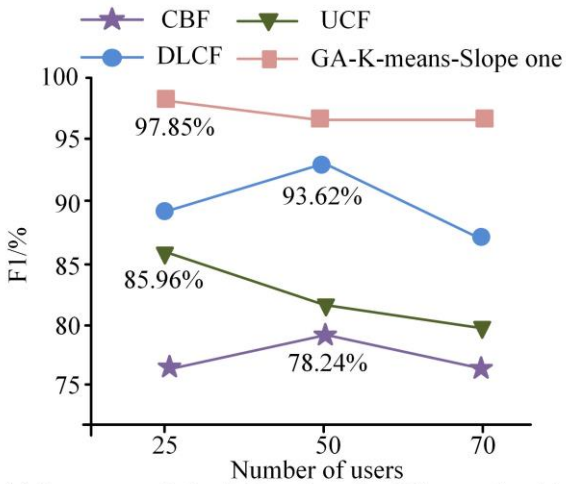
recommendation accuracy. The ability to make highly accurate recommendations stems from the accuracy of the model in clustering user behavior data and the efficiency of collaborative filtering, enabling the algorithm to capture users' different needs and potential preferences.



(a) Recommended precision of different algorithms



(b) Recommended recall rates of different algorithms



(c) Recommended call F1 values for different algorithms

Fig. 7. Recommendation precision, recall, and F1 value of different algorithms

Specifically, the K-means algorithm optimized by GA can effectively divide users into groups with similar behavioral characteristics, thus generating personalized recommendation lists based on more accurate user preferences in the Slope One recommendation process. The PR curve is chosen to more intuitively demonstrate the benchmark performance, as displayed in Fig. 8.

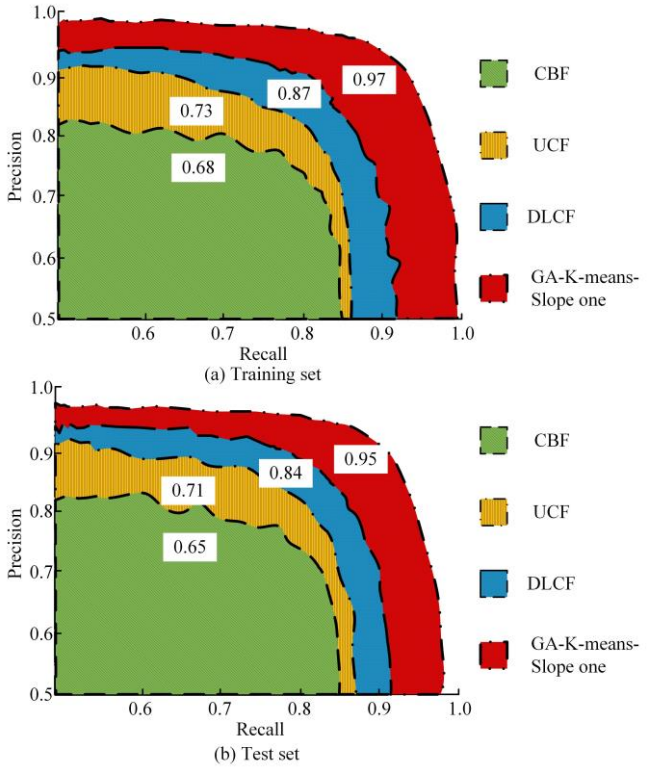


Fig. 8. The PR curves for different algorithms

Figs. 8(a) and 8(b) display the PR curves of CBF, UCF, DLCF, and GA-K-means-Slope one in the training and testing sets. The area value under the PR is denoted as Area Under the Precision Recall Curve (AUC), which can effectively measure the performance of various algorithms under various conditions. In Fig. 8 (a), the AUC of CBF, UCF, DLCF, and GA-K-means-Slope one during the training process was 0.68, 0.73, 0.87, and 0.97, respectively. In Fig. 8 (b), the AUC of CBF, UCF, DLCF, and GA-K-means-Slope one during the testing process was 0.65, 0.71, 0.84, and 0.95, respectively. According to the test results of the PR curve, the GA-K-means-Slope one algorithm performs better in various indicators in benchmark performance testing. The PR curve in Figure 8 further supports the stability performance of GA-K-means-Slope one algorithm under different conditions. Its high AUC value indicates that the model also performs well in accepting recommendations, reflecting high accuracy and low error rates. This not only improves the user's shopping experience, but also enhances the e-commerce platform's sensitivity to changes in user needs and can respond more quickly to market changes.

B. Application Effectiveness Analysis of Recommendation Prediction Models

After verifying the benchmark performance of the GA-K-means-Slope one, this study applies the GA-K-means-Slope one algorithm and comparison algorithms to recommendation prediction models. The

application effects on actual recommendation prediction services are compared. Amazon, Taobao, JD, Pinduoduo, and Tiktok are selected as application objects, respectively. Firstly, the average recommendation error values of the four models on the five e-commerce platforms are compared, as shown in Fig. 9.

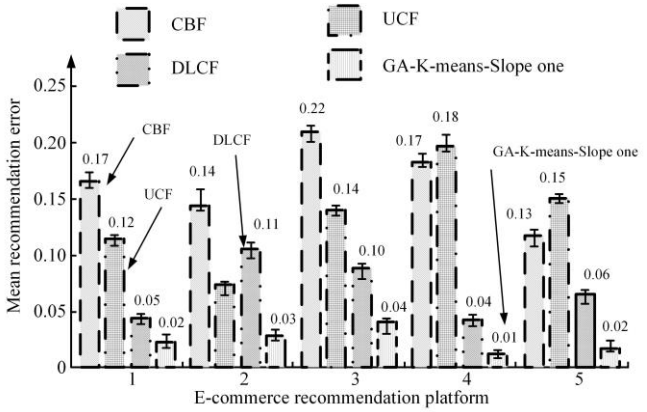
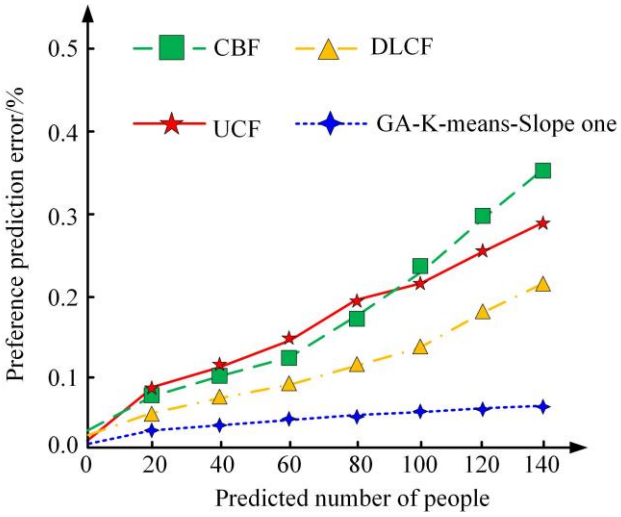


Fig. 9. Mean recommendation error for the different models

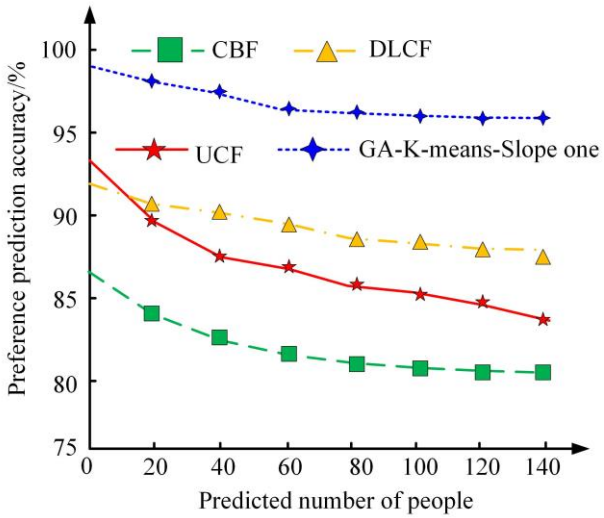
Fig. 9 shows the mean recommendation error values of four models on five e-commerce platforms. The e-commerce platforms 1 to 5 in Fig. 9 are Amazon, Taobao, JD, Pinduoduo, and Tiktok, respectively. The mean recommendation error of CBF, UCF, DLCF, and GA-K-means-Slope one on multiple e-commerce platforms were different, but the overall error performance of GA-K-means-Slope one was the best, with the minimum mean recommendation error of 0.04. The improved Slope One algorithm combined with optimized K-means not only considers the user's rating history, but also comprehensively considers the similarity between items in the recommendation process. This feature reduces reliance on cold start users and sparse data, ensuring high-quality recommendations can be generated even in situations where information is scarce. The recommendation accuracy and recommendation time are further compared, as displayed in Table II.

TABLE II
THE RECOMMENDATION ACCURACY AND RECOMMENDATION TIME IN THE FIVE E-COMMERCE PLATFORMS

E-commerce platform	Index	CBF	UCF	DLCF	GA-K-means-Slope one
1	Recommendation accuracy	86.2%	89.7%	93.1%	97.5%
	Recommendation time	5.8s	4.5s	2.6s	1.1s
2	Recommendation accuracy	85.4%	91.2%	94.0%	98.2%
	Recommendation time	6.0s	4.1s	2.8s	0.8s
3	Recommendation accuracy	83.8%	88.7%	90.5%	97.9%
	Recommendation time	6.3s	3.8s	3.1s	1.5s
4	Recommendation accuracy	84.9%	89.3%	92.4%	99.1%
	Recommendation time	5.7s	3.5s	2.5s	0.6s
5	Recommendation accuracy	85.5%	88.5%	91.6%	98.6%
	Recommendation time	5.3s	4.6s	2.2s	1.3s



(a) The preference of different models to predict error



(b) The preferences of different models predict accuracy

Fig. 10. Prediction results of different models for e-commerce user preferences

In Table II, CBF, UCF, DLCF, and GA-K-means-Slope one had the highest recommendation accuracy of 86.2%, 91.2%, 94.0%, and 99.1% on the five e-commerce platforms, respectively, with the shortest recommendation time of 5.3s, 3.5s, 2.2s, and 0.6s. The results in Table II indicate that users can quickly and accurately obtain recommendations for products that interest them, greatly improving the timeliness and convenience of information acquisition for users. At the same time, the fast response ability of the algorithm also provides good support for the operation of the e-commerce platform in high concurrency situations, ensuring that the recommendation system can maintain stable performance during peak user traffic periods. In addition to testing the recommendation performance, the predictive ability of various models for e-commerce user preferences is tested, as shown in Fig. 10.

Figs. 10(a) and 10(a) present the preference prediction errors and accuracy of four models for different e-commerce users. In Fig. 10 (a), as the number of e-commerce users continued to increase, the preference prediction errors of CBF, UCF, DLCF, and GA-K-means-Slope one models also gradually increased, with the highest reaching 0.36%, 0.29%, 0.23%, and 0.06%, respectively. Similarly, the preference

prediction accuracy of CBF, UCF, DLCF, and GA-K-means-Slope one also changed with the increase of the users. After reaching stability, the prediction accuracy was 82.6%, 84.7%, 87.9%, and 97.3%, respectively. By considering the similarity of scores and semantic similarity of items comprehensively, the model uses linear weighting method to refine the correlation between items, so that the recommendation is more in line with the real preferences of users. This multi-level similarity calculation enables GA-K-means-Slope one to reflect users' individual needs more comprehensively when generating recommendations.

The performance of GA-K-means-Slope one hybrid recommendation model is verified by experiments. The proposed method achieved a recommendation precision of 98.03%, which was significantly higher than other comparison algorithms. This result shows that the model can capture user preferences more accurately. In terms of recall rate, the research model reached 97.63% and the F1 value was 97.85%, both of which were superior to the other three algorithms. The high recall rate and F1 value indicate that the model can not only accurately recommend, but also cover more potential user preferences. In practical applications, the average time to complete the recommended task was 0.6 seconds. This speed was much faster than other algorithms, which proved the efficiency and practicability of the model. When the number of users exceeded 100, the preference prediction error was as low as 0.06%, and the prediction accuracy was as high as 97.3%. This indicates that the model can maintain high prediction ability when facing large-scale user data. GA-K-means-Slope one hybrid recommendation model has outstanding performance on multiple indicators, which is due to the combination of GA and K-means. The optimized GA is introduced into the model when selecting clustering centers, which significantly reduces the clustering error. The weighted Slope One algorithm takes into account the scores and semantic similarity between items, so as to more accurately reflect user preferences. In the same type of research, Feng L proposed an e-commerce data analysis and prediction method based on GBDT deep learning model. The method classified the purchase behavior into a binary classification problem, and extracted 107 features that reflected the user behavior. The GBDT model was constructed. The results showed that the proposed GBDT model was more suitable for e-commerce data analysis and prediction [19]. Yadav V et al. combined the Long Short-Term Memory Network (LSTM) with the continuous word bag model to deal with the complexity of Hindi sentiment analysis. The optimal word size of the input vector was calculated to enhance the accuracy. The proposed method achieved an average accuracy of more than 87%. Compared with the current mainstream methods, its performance is superior to the selected method [20]. Compared with similar studies, the results are presented in Table III.

Table III compares the performance of GA-K-means-Slope one with GBDT and the improved LSTM algorithm. The model outperformed other models in several indicators. This not only demonstrates the rationality of the GA-K-means-Slope one algorithm design, but also demonstrates its strong competitiveness in personalized recommendation in e-commerce. Through the effective combination of clustering and collaborative filtering, GA-K-means-Slope one can perform well in problems such as sparse data and cold start, which significantly improves user retention rate and purchase conversion rate in practical applications. Its excellent performance not only proves the effectiveness of combining clustering and collaborative filtering, but also provides a new research idea for e-commerce user preference recommendation and analysis.

TABLE III
Performance comparison results of research method

Model	Recommended precision (%)	Recall rate (%)	F1 value (%)	Average recommendation time (s)	Prediction accuracy (%)
GA-K-means-Slope one	98.03	97.63	97.85	0.6	97.3
GBDT	93.6	87.8	89.32	0.9	92.6
Improved LSTM	86.4	83.5	84.94	1.1	87.4

V. CONCLUSION

To deeply explore various data of e-commerce users and further improve the preference recommendation and prediction performance of current e-commerce users, a new hybrid recommendation model GA-K-means-Slope one was developed by combining clustering algorithm and collaborative filtering recommendation algorithm. The experimental results showed that GA-K-means-Slope one iterated to a stable state faster than the three comparison algorithms, including CBF, UCF, and DLCF, which maintained stability with a minimum of 48 iterations. In addition, GA-K-means-Slope one had better precision, recall, and F1 value in benchmark tests, reaching 98.03%, 97.63%, and 97.85%, respectively. The AUC in the PR curve was also as high as 0.97. In practical applications, the GA-K-means-Slope one model was used to complete recommendation tasks on five different e-commerce platforms. The average recommendation error was as low as 0.04, the recommendation accuracy reached up to 99.1%, and the recommendation time was as low as 0.6 seconds. In the prediction task, when the predicted individual exceeded 100, the preference prediction error of GA-K-means-Slope one was as low as 0.06%, and the prediction accuracy was as high as 97.3%. Overall, the GA-K-means-Slope one has good recommendation and prediction performance. The main contribution of this research is to introduce a hybrid model to fill the gap in the existing research on user behavior analysis. The GA is used to optimize the K-means clustering process. The clustering effect and stability are improved, and the randomness problem of the traditional K-means in selecting the clustering center is solved. It has strong applicability and scalability in multiple e-commerce platforms, providing practical solutions for the e-commerce industry. It has promoted the customer stickiness and sales growth of the e-commerce platform, which has great commercial value for the development of the e-commerce platform. Future research can explore real-time performance changes and adaptability to dynamically changing user preferences.

REFERENCES

[1] Mu R. A novel neural collaborative filtering recommendation based on side information fusion. Comptes Rendus de l Academie Bulgare

des Sciences: Sciences Mathematiques et Naturelles, 2023, 76(1): 84-95.
[2] Tian Z, Liu Y, Sun J, Jiang Y, Zhu M. Exploiting group information for personalized recommendation with graph neural networks. ACM Transactions on Information Systems (TOIS), 2021, 40(2): 2-23.
[3] He M, Wang J, Ding T, Shen T. Conversation and recommendation: knowledge-enhanced personalized dialog system. Knowledge and Information Systems, 2023, 65(1):261-279.
[4] Hebbi C, Mamatha H. Comprehensive dataset building and recognition of isolated handwritten kannada characters using machine learning models. Artificial Intelligence and Applications, 2023, 1(3):179-190.
[5] Chen X, Wang J. Accurate medical information recommendation system based on big data analysis. International Journal of Industrial and Systems Engineering, 2022, 41(2): 237-253.
[6] Ye L, Yang Y, Zeng J. An interpretable mechanism for personalized recommendation based on cross feature. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 2021, 40(5): 9787-9798.
[7] Zhang S, Wang L, Fei R, Xu X, Wei Li. Attraction recommendation based on tourism context modeling and multi-neural collaborative filtering algorithm. IEEJ Transactions on Electrical and Electronic Engineering, 2023,18(8): 1280-1295.
[8] Lim H, Xie L. A New Weighted imputed neighborhood-regularized tri-factorization one-class collaborative filtering algorithm: application to target gene prediction of transcription factors. IEEE/ACM transactions on computational biology and bioinformatics, 2021, 18(1): 126-137.
[9] Zhang J, Yang J, Wang L, Jiang Y, Qian P, Liu Y. A novel collaborative filtering algorithm and its application for recommendations in e-commerce. Computer Modeling in Engineering and Science, 2021, 1(3): 1275-1291.
[10] Sun G, Zheng J. Construction and application of music audio database based on collaborative filtering algorithm. Discrete Dynamics in Nature and Society, 2022, 2022(8): 1026-1035.
[11] Guo S, Zhu X, Liu Y, Han J. Personalized recommendation method of entrepreneurial service information based on blockchain. Journal of Interconnection Networks, 2022, 22(3): 2-21.
[12] Pei H, Liu X, Huang X, Wu M, Wen Z, Zhao F. A personalized recommendation method under the cloud platform based on users? long-term preferences and instant interests. Advanced Engineering Informatics, 2022, 3(54): 63-78.
[13] Lin Y, Feng S, Lin F, Xiahou J, Zeng W. Multi-scale reinforced profile for personalized recommendation with deep neural networks in MOOCs. Applied Soft Computing, 2023:148(2): 2-13.
[14] Zhen Y, Liu H, Sun M, Yang B, Zhang P. Adaptive preference transfer for personalized IoT entity recommendation. Pattern recognition letters, 2022, 162(10): 40-46.
[15] Rani A, Taneja K, Taneja H. Life Insurance-Based Recommendation System for Effective Information Computing. International Journal of Information Retrieval Research, 2021, 11(2): 2-14.
[16] Jaffke L, Lima P T. On the maximum number of edges in planar graphs of bounded degree and matching number. Discrete Mathematics, 2023, 346(8): 431-438.
[17] Kamoji S, Kalla M. Effective Flood prediction model based on Twitter Text and Image analysis using BMLP and SDAE-HHNN. Engineering Applications of Artificial Intelligence, 2023, 128(2): 2-15.
[18] Frehill L M, Leung M A. Twitter gone wrong: how constructive dialog and collaboration enable innovation. Computing in Science & Engineering, 2021, 23(1): 97-101.
[19] Feng L. Data analysis and prediction modeling based on deep learning in e‐commerce. Scientific Programming, 2022, 2022(1): 1041741.
[20] Yadav V, Verma P, Katiyar V. Long short term memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages. International Journal of Information Technology, 2023, 15(2): 759-772.