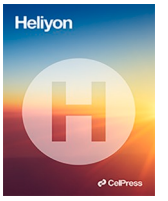
Heliyon 10 (2024) e29045

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

E-commerce recommender system based on improved K-means commodity information management model

Wei Zhang *, Zonghua Wu

School of mathematics, South China University of Technology, Guangzhou, 510641, Guangdong, China

ARTICLE INFO

Keywords:
E-commerce
Recommender system
K-mean clustering algorithm
Fuzzy C-Mean algorithm
Genetic algorithm

ABSTRACT

Since the start of the 21st century, there has been a rapid development of internet technology, causing electronic computers and smartphones to become increasingly popular. The e-commerce industry also experiences quick development. However, the recommendation technology of e-commerce progresses slowly, hindering it from keeping up with the changing times. To enhance the efficiency and accuracy of e-commerce recommender systems, this research introduces an e-commerce recommender system that utilizes an enhanced K-means clustering algorithm to manage commodity information. This method combines the K-means algorithm with a genetic algorithm by encoding the genetic algorithm, setting the initial population, defining the fitness function, and configuring other parameters. The results of the test indicated that the K-mean clustering algorithm and fuzzy C-mean algorithm had a recommendation accuracy of 87.9 % and 84.8 % respectively under the test dataset. The highest recommendation accuracy was observed from the improved K-mean clustering algorithm, which was 91.1 %. The convergence rate of the improved K-mean clustering algorithm was faster by 44 % compared to the traditional K-mean clustering algorithm and 73 % quicker than the fuzzy C-mean algorithm. The study's findings demonstrate that the refined K-means clustering algorithm greatly enhances the recommendation proficiency and precision of the e-commerce recommendation system, in comparison to other comparable algorithms. This research can potentially advance the e-commerce industry and stimulate its growth.

1. Introduction

As Internet technology continues to advance and gain traction, e-commerce has become increasingly popular globally and has become a major force behind the modern commercial sector [1,2]. E-commerce platforms produce a significant quantity of product data daily, making the efficiency and precision of product information management and recommender systems (RS) a vital concern [3]. Commodity information management encompasses various aspects including classification, labeling, and pricing of commodities. Meanwhile, RS aims to enhance user experience and boost sales by providing users with personalized recommendations tailored to their interests and needs. K-means clustering algorithm (KMCA) is a widely adopted approach for commodity information management, as it enables merchants to organize and manage commodity information more efficiently by grouping similar products [4]. However, the conventional K-means algorithm (CKMA) has certain drawbacks, including reduced mining ability when faced with vast amounts of data and reduced mining efficiency, which limits its effectiveness in e-commerce applications. The purpose of this study is

* Corresponding author.
E-mail addresses: zw2020@scut.edu.cn (W. Zhang), Zonghuawu202130321523@gmail.com (Z. Wu).

https://doi.org/10.1016/j.heliyon.2024.e29045
Received 29 November 2023; Received in revised form 23 March 2024; Accepted 28 March 2024
Available online 16 April 2024

to enhance KMA to better assist e-commerce platforms in managing product information and apply it in the E-commerce recommender system (ECRS). The enhanced CKMA will consider additional factors, including the product description, purchasing history, and visual attributes of goods, to enhance clustering effectiveness. This study offers novel concepts and techniques for product information management and personalized recommendations in the e-commerce domain, advancing the growth of this sector. The study comprises five parts. The initial part provides a general introduction, followed by a comprehensive literature review of domestic and international studies. The third section thoroughly examines the technique of the algorithm, while the fourth section assesses its performance through a battery of tests. Finally, the fifth section summarizes the study's findings and limitations. The K-means algorithm has been improved by first solving the problem of cluster center selection using a genetic algorithm (GA). This algorithm can find a more suitable initial centroid set through global search, increasing the stability of clustering results. Secondly, the introduction of the GA solves the problem of local optimal solutions in the traditional K-means algorithm. The GA gradually approaches the global optimal solution through continuous iteration, effectively avoiding local optimal solution problems and increasing classification accuracy. Additionally, the GA enhances the data mining performance of the K-means algorithm.

2. Related works

The KMCA finds wide application in diverse industries owing to its exceptional performance. Researchers have exploited this algorithm to address various issues. Abellana D P M and his team put forth a novel decision-making experiment that hybridizes the KMCA to propose a decision-making framework for analyzing the obstacles to the adoption of green computing. The methodology was proposed by the researchers after summarizing a vast body of literature. They experimented with the framework, and the results of the experiment revealed that the model played an essential part in analyzing the hindrances to adopting green computing [5]. In order to more effectively segment Pap pictures, Hadianti S. et al. suggested a method for segmentation and corresponding analysis of cellular images of Pap smear using KMA. The experimental results showed that the segmentation method had good performance in terms of accuracy as well as sensitivity and the method has potential for clinical application [6]. Jiaquan Huang et al. found that the traditional collaborative filtering algorithm suffers from data scarcity problem, which reduces the recommendation efficiency of RS. To solve this problem, the researchers proposed an improved collaborative filtering personalization algorithm. Experimental results showed that the algorithm had significant improvement in the accuracy and precision of recommendation [7]. The firefly algorithm, developed by Zhang F et al. and improved by KMA, can help maintain the power batteries in electric vehicles by addressing the issue of power battery voltage platform load overload. Experimental findings show that the algorithm improves convergence speed and prevents the power battery's internal model from slipping into a local minima. The updated algorithm has improved the stability of electric car power battery management systems [8]. Fang Z et al. utilized the K-clustering algorithm to enhance the stock prediction and recommendation algorithm, providing better prediction references for stock investors. The trial results showed that the algorithm can effectively assist stock investors in identifying high-quality and low-quality companies among equities [9]. Jawad T M and his team improved the existing energy saving protocol by using K-mean clustering algorithm in order to extend the life cycle of the network more efficiently. This algorithm was capable of detecting the Euclidean distance of each sensor node to the base station and its remaining energy, which efficiently helps the staff to detect the nodes. Simulation results showed that the algorithm had significant improvement compared to the original algorithm [10].

To solve combinatorial problems in personalized recommendation, the use of GAs, a modern heuristic algorithm, is well suited. Liu Z et al. proposed a hybrid algorithm that considers both objective QoS attributes and customer preference attributes to solve the personalized recommendation problem of manufacturing service combination (MSC). The algorithm used the collaborative clustering filter (CCF) to quantify customer preferences and the third-generation non dominant sorting genetic algorithm (PoNSGA-III) to optimize MSC's multiple attributes. According to the findings, this method can suggest the most appropriate solution for the target customers [11]. Wang CL et al. proposed an improved GA to solve the optimization scheduling problem of stores. The algorithm used double-layer encoding for optimization and the roulette wheel method for individual screening. The results indicated that the GA is highly stable and can recommend better picking orders and vehicle quantities for various types of picking problems [12]. Wu L, Ye X et al. utilized a GA as the primary recommendation method in their framework to explore various collaborative drug recommendation combinations in different cancer cell lines. The framework included imbalanced data processing and a search for global optimal solutions. The results indicated that this method has a more accurate recommendation ability compared to the other 11 algorithms [13].

Due to the overwhelming volume of information available on the Internet, users sometimes struggle to make decisions. To help users locate the stuff they need fast, academics have extensively studied RS. To improve the performance and interpretability of recommendations, the construction method and effectiveness of the recommendation model are shown in Table 1.

Table 1
Methods and effects of constructing several recommendation models.

Literature	Means	Accuracy
[14]	A Probability Propagation Framework for Non dimensional Time Graph Based on Multidimensional Metapath	76.3 %
[15]	Introduced an attention flow network to model user purchase records and construct a recommendation model	79.3 %
[16]	Machine learning classifiers recognize travel related tweets, which are then used to obtain personalized travel recommendations	75.23 %
[17]	Applying K-Means clustering method to possible clustering recognition of similar product groups and constructing an image-based recommendation system	81.23 %
[18]	Using k-means clustering algorithm to recommend journals	76.36 %

In summary, CKMA has been applied to various fields by researchers, but it has been less frequently applied to ECRS. The current common RS is not flawless. This study consequently enhanced K-means to better suit ECRS and suggested an ECRS grounded on the refined K-means model for managing commodity information.

3. ECRS based on improved K-means commodity information management models

CKMA is widely used in several industries, fields because of its high efficiency in mining data and simple algorithmic logic. However, researchers have gradually found that CKMA's mining ability has decreased in the current era when faced with a large amount of data, and the mining efficiency is reduced. This study used GA to enhance CKMA's data mining effect and boost the efficiency and accuracy of ECRS recommendations based on commodity information management.

3.1. Application of CKMA and GA in RS

The CKMA is an unsupervised learning algorithm. The key to the algorithm is the ability to calculate the similarity between samples. The samples in the algorithm are classified without repetition. The distance between the individual samples of K-Means is determined using the square of the Euclidean distance [19,20]. The Euclidean distance expression is shown in equation (1).

$$d(x_i, x_j) = \sum_{k=1}^{m} (x_{ki} - x_{kj})^2 \tag{1}$$

In equation (1), x_i and x_j denote two d dimensional data objects respectively. After calculating the distance between the samples using the Euclidean distance equation the loss function is defined as the sum of the distances between the samples and the centre they belong to. The loss function expression is shown in equation (2) [21].

$$W(C) = \sum_{m=1}^{k} \sum_{C(i)=m} \|Xi - \overline{X_m}\|^2 \tag{2}$$

In equation (2), i is the number of samples in each class, m is the number of classes, $\overline{X_m}$ denotes the mean value of class m. $W(C)$ denotes the similarity of each sample in the same class [22,23]. The equation for minimizing the loss function is shown in equation (3).

$$\min_{C} \sum_{m=1}^{k} \sum_{C(i)=m} \|X_i - \overline{X_m}\|^2 \tag{3}$$

In equation (3), X_i denotes the i th sample and $\overline{X_m}$ denotes the mean of the m th class. Subsequently the centre of each class is found, i.e., $(e_1, e_2, \ldots, e_k)$ is found and the loss function is minimized. The expression for the minimized loss function in this step is shown in equation (4) [24].

$$\min_{e_1, e_2, \ldots, e_k} \sum_{m=1}^{k} \sum_{C(i)=m} \|X_i - \overline{X_m}\|^2 \tag{4}$$

In equation (4), $(e_1, e_2, \ldots, e_k)$ represents each initial cluster centre. After finding the centre of each class it is necessary to update the

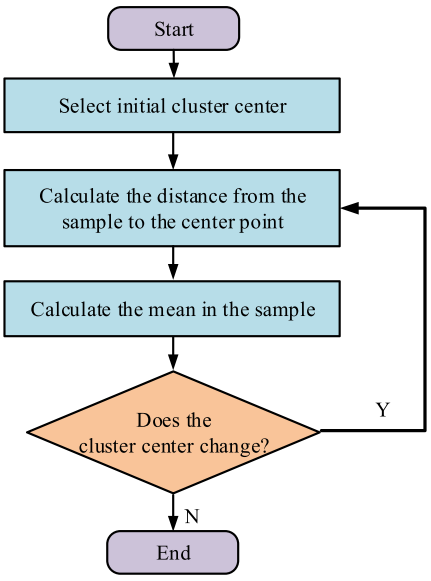


Fig. 1. K-means algorithm flowchart.

mean $\overline{X_m}$. the updated expression is shown in equation (5).

$$\overline{X_m} = \frac{1}{n_m} \sum_{C(i)=m} X_i \tag{5}$$

In equation (5), n is samples per class and m is the classes. The CKMA flowchart is shown in Fig. 1 [25].

First, K objects in the data space are chosen in Fig. 1 as the initial clustering centre, where each object represents the centre of a grouping. Next, the distance between the samples and the centre is calculated, and the samples are assigned based on the minimum distance. The third step calculates the mean in the samples. And the fourth step determines whether the position of the clustering centre has been changed, and if it has been changed, then it is necessary to return to the second step until the position remains unchanged, and finally outputs the result, the process ends [26]. GA is an algorithm designed and proposed according to the law of biological evolution in nature. The flowchart of GA is shown in Fig. 2.

The generic GA in Fig. 2 is made up of biological people, biological groups, biological gene forms, biological fitness calculations, biological hybridization, reproduction, and mutation techniques. The individual organisms in it will use digital simulation of the chromosomes of the organisms. The chromosomes are then encoded using a binary coding method. The coding method is shown in equation (6).

$$g\left(x_l^t, k\right) = u_k + \frac{u_k - v_k}{2l - 1} \left(\sum_{j=1}^{t} x_t^{i(kl+j)} \times 2^{j-1} \right) \tag{6}$$

In equation (6), u_k denotes the upper limit of the real number k and v_k denotes the lower limit of the real number k. The floating point encoding in GA is shown in equation (7).

$$X_0 = \left({x_0}^1 {x_0}^2 \ldots {x_0}^n \right) \tag{7}$$

In equation (7), ${x_0}^n$ denotes the n th chromosome of generation 0. The RS designed for this research uses CKMA for the classification work of e-commerce related data, where each set of data will be considered as a cluster class. A group of data with similar similarity will be formed. k cluster classes will be obtained by CKMA from the corresponding e-commerce databases and subsequently the results will be calculated using an iterative method until the error value meets the requirements [27,28]. The current CKMA are challenging to extract valuable information from a large amount of e-commerce data, and traditional CKMA are significantly impacted by the selection of cluster center points. Therefore, this study integrated a GA into the K-means algorithm to improve it. The improved CKMA encodes the center point using floating-point coding and selects candidates based on the GA's fitness function, which may enhance the algorithm's stability. The flowchart of the improved CKMA is shown in Fig. 3.

In Fig. 3, it is important to note that when the similarity of different data is low, genetic variation is required to determine whether it is less than the convergence factor. There are four key points in the improved CKMA. First, chromosomes and components are the focus in the improved CKMA. The chromosome definition method constructed in this study uses $Si(j,k)$ representation. Second, the

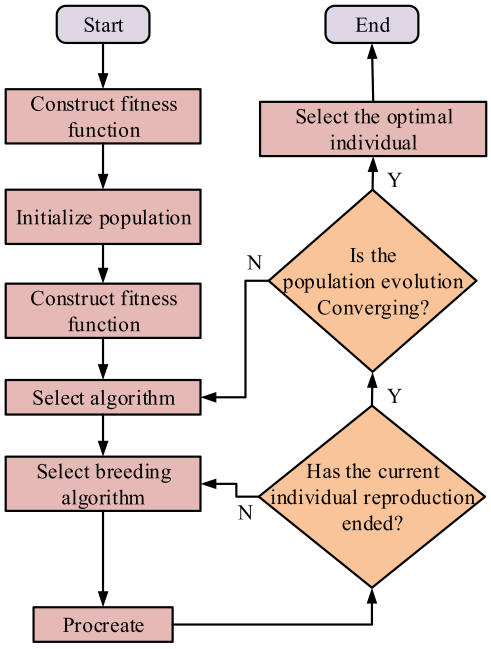


Fig. 2. Flowchart of genetic algorithm.

algorithm adaptation calculation. The adaptation calculation requires the construction of an error function. The specific representation of the error function is shown in equation (8).

$$P = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - c_j \right\|^2 \tag{8}$$

In Equation (8), x_i denotes the i th data in the j th cluster class and c_j denotes the j th cluster class centroid. Equation (9), which shows the calculation of the value of the sum of standard deviations in the data set, is used.

$$f(P_k) = \sum_{i=1}^{k} \sum_{P \subset C_i} |P - m_i|^2 \tag{9}$$

In equation (9), $f(P_k)$ is also called the criterion function, which is a function to determine whether the cluster centre meets the criteria, and it represents the sum of the errors of each cluster. c_i denotes the i th cluster; and m_i denotes the average of c_i clusters, and equation (10) displays its particular calculating equation.

$$m_i = \frac{1}{t_i} \sum_{p \subset C_i} p \tag{10}$$

In equation (10), c_i denotes the i th cluster. According to equation (10) it can be obtained that the clustering effect is best when the function achieves the minimum value, so the equation for the degree of adaptation is specifically shown in equation (11).

$$f(R_i) = E_{\max} - E(R_i) \tag{11}$$

In equation (11), $E_{\max}$ denotes the maximum chromosomal variance of population R. The third key point is the selection of data. The improved CKMA uses the roulette wheel method to obtain the selection probability of the centroids. Equation (12) displays the specific equation.

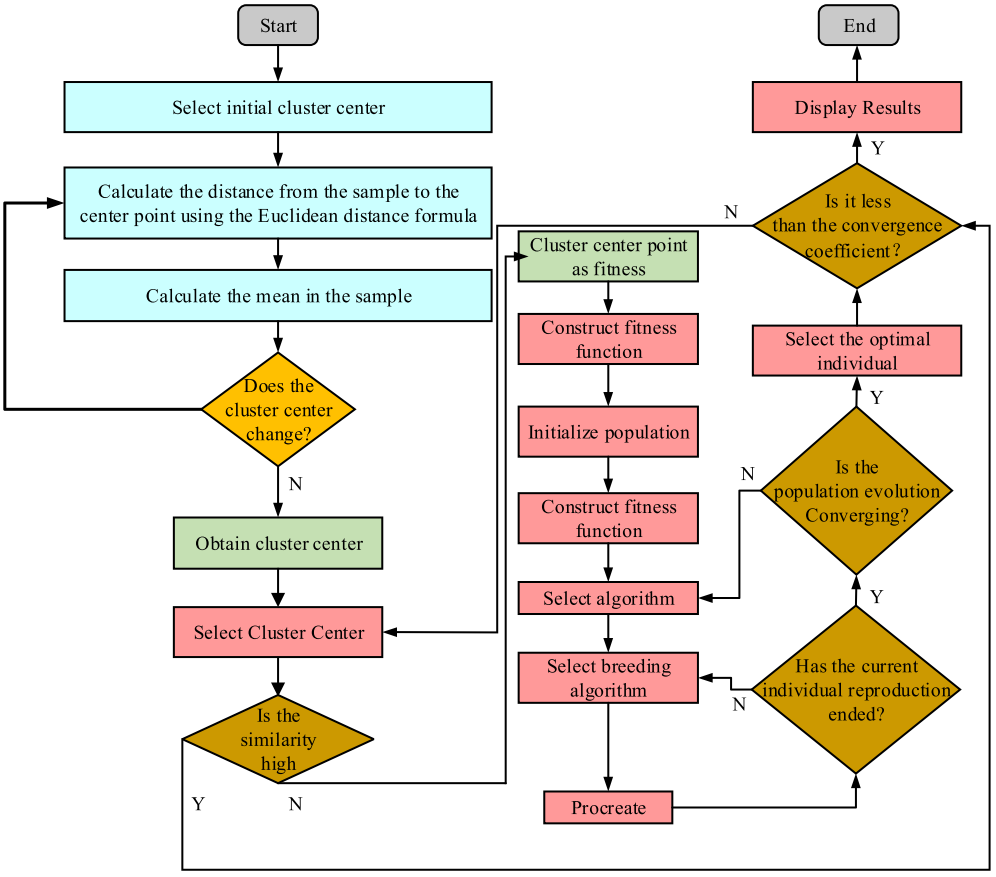


Fig. 3. Improved K-means algorithm flowchart.

$$P_i = \frac{f(R_i)}{\sum_{i=1}^{n} f(R_i)} (i = 1, 2, ..., n) \tag{12}$$

In equation (12), $f(R_i)$ denotes the fitness function. The fourth key point is to set the mutation rate $P = 0.03$ for the mutation operation. The improved K-means algorithm first solves the problem of cluster center selection using GA. GA can find a more suitable initial centroid set through global search, increasing the stability of clustering results. Secondly, the introduction of GA solves the problem of local optimal solution in traditional K-means algorithm. Through continuous iteration, GA gradually approaches the global optimal solution, effectively avoiding local optimal solution problems and increasing classification accuracy. In addition, the GA also enhances the data mining performance of the K-means algorithm.

3.2. ECRS based on commodity information management model

RS must select products that match the user's purchasing preferences and interest in making a purchase from a wide range of options. Therefore, the management of commodity information becomes very important. Each commodity in e-commerce has a corresponding category. The categories of commodities are clearly categorized in e-commerce platforms [29,30]. However, in RS, commodity information should be managed in a refined way. The commodity information management system includes the classification of commodity information, the management of user shopping cart, the management of commodity information and the management of customer complaints. Fig. 4 displays the commodity information management system's schematic diagram.

In Fig. 4, the system administrator and users are the key users in the commodity information management system. The system administrator has multiple privileges and can operate all functions in the commodity information management system. For example, commodity information classification management, e-commerce commodity information management and so on. Users can view, add, or delete commodity categories and shopping cart items, as well as view commodity information and file complaints. Complaints can be added or replied to independently. E-commerce product recommendation management is an important function of the RS. Fig. 5 displays the schematic diagram for commodity suggestion management.

In Fig. 5, the recommendation management includes search term recommendation, search result interface recommendation, and purchase product information management. The search term recommendation function in this system includes the intelligent search function, which can record the user's search history, so that the customer can pop up the relevant search records by entering some keywords in the subsequent searches. The search result interface recommendation refers to recommending product information to customers based on the search information entered by users, and recommending related products based on the similarity information mined by the improved CKMA. In this RS, the system administrator can manage search terms, search result interfaces, and other related functions. The information that the user can receive includes the display of recommended results for search terms, the display of recommended pages for search results, and other recommended displays of related results. The management of the system is also one of the important parts of the system. The RS designed for the study also contains a system management module. The functions of this module include the management of e-commerce users, the management of e-commerce roles and the management of sub-commerce data. E-commerce user management can be able to add e-commerce user information, e-commerce user information view, cancellation and modification of the function; user role management including e-commerce user role addition, e-commerce user information view and e-commerce user information modification. Data management includes e-commerce data backup, e-commerce data restore management; ECRS needs to meet the needs of both e-commerce platforms and users. Therefore, the ECRS is designed according to four main requirements. First, easy operation: ECRS has a wide audience, users of all ages, industries and educational levels may use the system, so ECRS should be designed according to the principle of simple and convenient operation. Secondly, security, there is a large amount of user data in the RS, including gender, age and other basic information, but also includes the relevant user's financial account information. The system architecture of RS prioritizes the security of user data to prevent any potential leaks. System stability is also

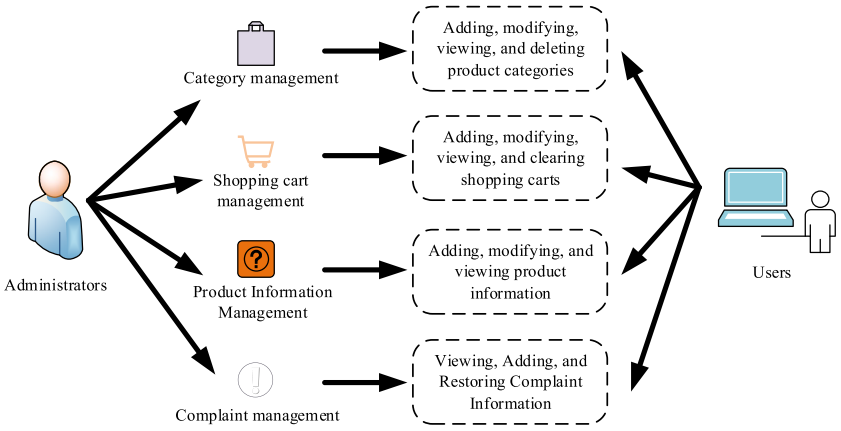


Fig. 4. Product information management system.

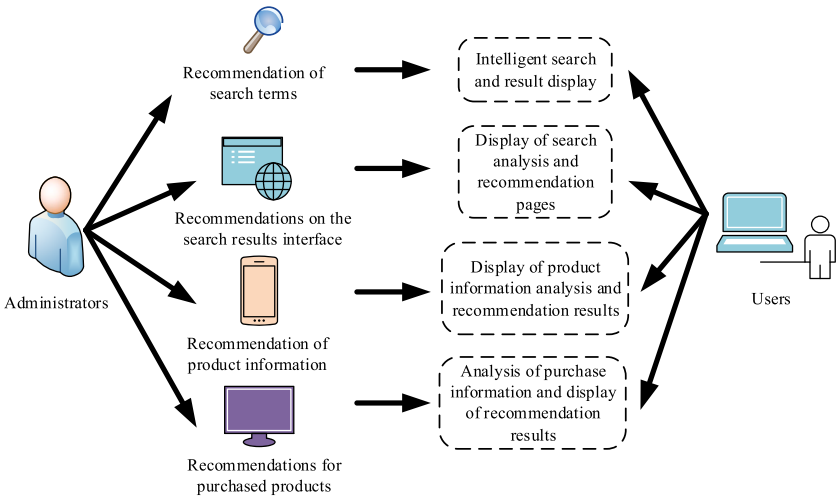


Fig. 5. Schematic diagram of product recommendation management.

important, but comes in third. It is crucial that both the amount of data processed by RS and the amount of data received by users are substantial. Users who browse through large amounts of data may experience unresponsiveness, leading to a negative user experience and potentially driving away users. This can have a negative impact on both the growth of the e-commerce platform and the activities of its users. Lastly, scalability is also a consideration. The e-commerce platform RS collects a variety of functions, but at the same time it can also lead to a variety of problems. When problems occur, the system needs to be improved or new tools need to be developed to help solve the problem. Therefore, the e-commerce platform RS needs a certain degree of scalability, in the design of the system needs to be reserved API interface to help subsequent development tools access.

4. Interpretation results of e-commerce recommendation system

4.1. Improvement of CKMA's performance tests

The performance test of this research on the improved CKMA was run on Windows 11 Professional, 13th Gen Intel(R) Core(TM) i5-13400F 2.50 GHz 32.0 GB RAM. In this study, the performance of KMCA, improved KMCA and FCM clustering algorithm were tested for comparison using contour coefficients. The comparison results of contour coefficients are shown in Fig. 6.

In Fig. 6, the vertical axis represents the contour coefficient, while the horizontal axis represents the clusters. The contour coefficient of the K-Means enhanced by this study gradually decreases with an increase in the number of clusters. However, it remains within the range of 0–1, indicating that the clustering outcomes of the improved CKMA are more accurate. When the clusters exceeds 7, the profile coefficient of the conventional CKMA is between −1 and 0, indicating that its clustering results are subpar. The profile coefficient of the traditional CKMA decreases as the number of clusters increases. The profile coefficient of the FCM clustering algorithm decreases rapidly as the clusters in the data increases. For four clusters, the coefficient falls between −1 and 0, indicating poor performance of the algorithm when processing data with additional clusters. In summary, the improved CKMA has better performance than the traditional K-Means and FCM clustering algorithms, and its clustering results are better. The convergence performance of the three algorithms with the number of iterations is shown in Fig. 7.

In Fig. 7, when the iterations reach 207 times, the improved CKMA starts to converge and the convergence performance is stable; when the iterations reach 469 times, the CKMA starts to converge, but as the number of iterations increases, its convergence performance fluctuates slightly. And when the iterations reach 983 times, the FCM algorithm starts to converge, but as the number of

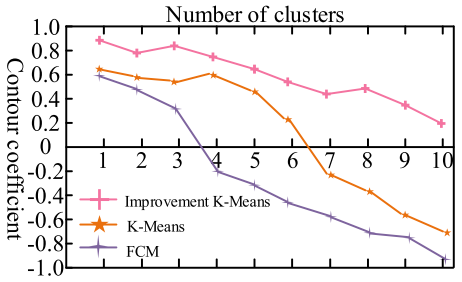


Fig. 6. Profile coefficient comparison chart.

iterations increases, the convergence performance fluctuates significantly. The convergence rate of the improved CKMA is 44 % faster than that of the CKMA and 73 % faster than that of the FCM algorithm. This demonstrates that the enhanced CKMA's convergence performance outperforms both the CKMA and FCM algorithms by a substantial margin.

4.2. Quality analysis of recommended results

The study randomly selected 2000 consumers from an e-commerce platform and divided their information into two datasets: a test dataset and a training dataset, in a 4:6 ratio. The experiment compared the accuracy of various recommendation numbers across different datasets. Fig. 8 presents the results.

Fig. 8(a) shows the accuracy results under the training dataset, when the number of recommenders is 100, all three algorithms reach the highest accuracy, 89.2 % for FCM, 92.3 % for K-means, and 94.9 % for the improved CKMA. The accuracy of the three algorithms rapidly declines as the number of recommenders increases. The accuracy results under the test dataset are displayed in Fig. 8(b). The accuracy of CKMA and FCM reaches its peak at 500 recommenders, with values of 91.1 % and 84.8 %, respectively. Meanwhile, K-means reaches its peak accuracy at 700 recommenders, with 87.9 %. The experimental results demonstrate that the modified CKMA outperforms the other two algorithms in terms of accuracy in both the training and test datasets. The recall of the three recommendation algorithms for different recommendation numbers under different datasets is shown in Fig. 9.

Fig. 9(a) represents the recall results in the training dataset, when the number of recommenders is 100, all three algorithms have the highest recall, 41.6 % for FCM, 42.8 % for K-means, and 44.2 % for the improved K-means. When the number of recommenders increases, all three algorithms show an overall decreasing trend. Fig. 9(b) shows the recall results in the test dataset, when the number of recommenders is 500, all three algorithms have the highest recall, 42.8 % for the improved K-means, 41.9 % for K-means and 40.5 % for CBF. The graph of recall results indicates that the recall of all three algorithms across various datasets has a diminishing tendency as the number of recommenders rises. On many datasets, the enhanced K-means method has a higher recall than the other two. To ensure the accuracy of the study, the MAE metrics and RMSE metrics of the three algorithms are compared.The experimental results of the MAE metrics are shown in Fig. 10.

Fig. 10(a) shows the experimental results of MAE metrics under the training dataset, FCM and CKMA reach the lowest MAE value at 300 referrals, FCM is 0.723; K-means is 0.718. improved CKMA reaches the lowest at 400 referrals, 0.706. Fig. 10(b) shows the experimental results of MAE metrics under the test dataset, when the number of referrals is 600, the MAE value of FCM reaches the lowest value of 0.723. When the referrals is 700, the MAE value of the improved CKMA reaches the lowest value of 0.72. The MAE of the improved CKMA is lower than that of the CKMA and FCM algorithms under both the training and test datasets, indicating better algorithmic performance. To ensure the comprehensiveness of the performance test results, the study compares the RMSE values of the different algorithms on MovieLens (1 M) and Epinions datasets. MovieLens (1 M) dataset consists of 8635 user information, 4568 items and 31248 ratings. Epinions dataset consists of 6482 user information, 2157 items and 23488 ratings. Table 2 displays the comparison findings of the RMSE values of various techniques on the MovieLens (1 M) dataset.

In Table 2, sparsity refers to the ratio of elements in the dataset with no scoring data to the entire dataset. The RMSE values of different algorithms decrease with the decrease of data sparsity. CF algorithm and Spark algorithm both have RMSE values above 1 when the sparsity is more than 70 %, and their performance is poor. Under all sparsity settings, the modified CKMA's RMSE values are much smaller than those of the other algorithms, suggesting that it performs better in the MovieLens (1 M) dataset. Table 3 displays the comparison findings of the RMSE values of various algorithms on the Epinions dataset.

In Table 3, sparsity refers to the ratio of elements in a dataset with no scoring data to the entire dataset.CF algorithm and Spark algorithm have RMSE values greater than 1 in all sparsity conditions, and perform poorly on this dataset; FCM algorithm has RMSE values less than 1 only at 60 % sparsity, and its performance is also poor; the improved CKMA in the Epinions dataset, the improved CKMA performs well on the Epinions dataset as the RMSE value is significantly smaller than the other algorithms under all sparsity conditions.

5. Conclusion

Aiming at the problem that traditional KMCAs are difficult to effectively mine information when facing large amounts of data, a e-

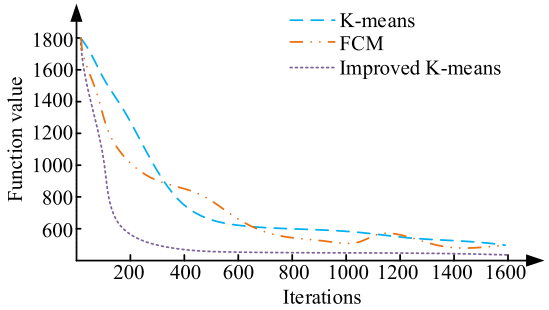


Fig. 7. Comparison of convergence performance of three algorithms.

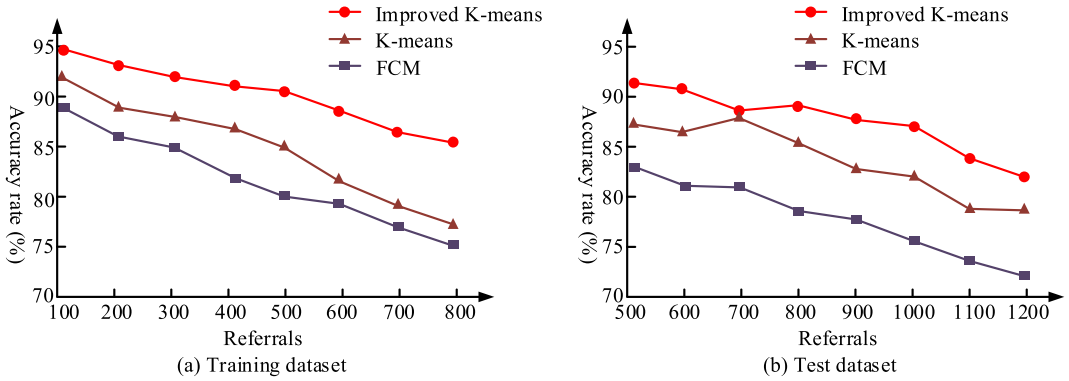


Fig. 8. The accuracy of the three algorithms under different recommendation numbers.

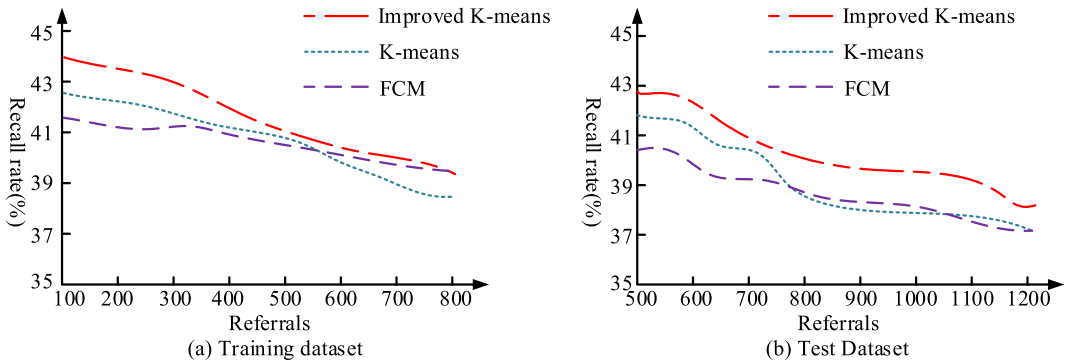


Fig. 9. Recall rate of three algorithms under different recommendation number.

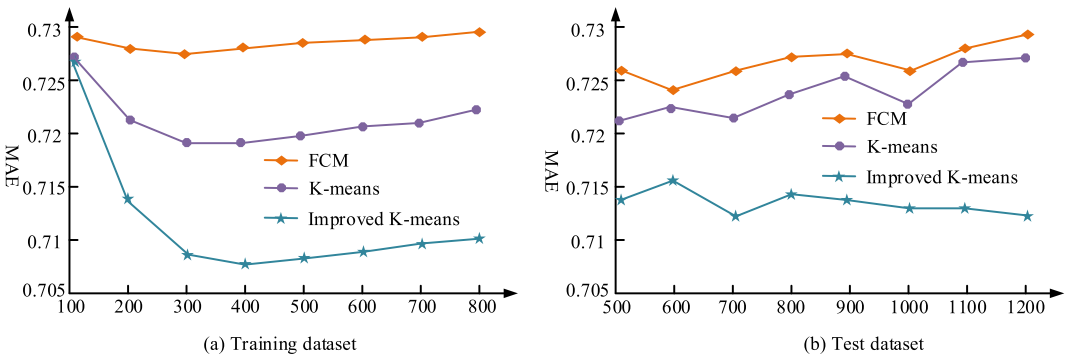


Fig. 10. Comparison results of MAE index under three different algorithms.

commerce recommendation system based on an improved K-means product information management model is proposed. This method utilizes a GA to improve the search capability of the KMCA. It combines personalized recommendation technology with e-commerce to efficiently process information in e-commerce systems. The results showed that the improved KMCA converges quickly after 207 iterations, with a recall rate of 42.8 % and a high recall rate, indicating that the method has higher accuracy and speed. In the MovieLens (1 M) dataset, the improved KMCA showed the lowest RMSE value of 0.68 at 60 % data sparsity, which is smaller than other algorithms. In the Epinions dataset, the improved KMCA showed the lowest RMSE value of 0.71 at 60 % data sparsity, which is also smaller than other algorithms. The KMCA has been improved to reduce prediction errors for user ratings and provide more accurate recommendation results. It now has superior data mining capabilities and can achieve personalized recommendations in e-commerce systems. However, the testing of this method is mostly focused on algorithm performance testing, and the testing of practical application functions still needs to be strengthened. Subsequent research will further improve the practical application ability of the research method.

Table 2
Comparison results of RMSE values on the movie lens (1 M) dataset.

Sparsity	Algorithm				
	CF	Spark	FCM	K-means	Improved K-means
95 %	1.50	1.50	1.40	1.20	0.96
90 %	1.40	1.36	1.25	1.13	0.91
85 %	1.38	1.33	1.19	1.06	0.88
80 %	1.30	1.28	1.13	0.97	0.82
75 %	1.25	1.21	1.06	0.89	0.79
70 %	1.16	1.07	0.95	0.86	0.75
65 %	0.98	0.92	0.93	0.83	0.71
60 %	0.96	0.85	0.84	0.80	0.68

Table 3
Comparison results of RMSE values on the epinions dataset.

Sparsity	Algorithm				
	CF	Spark	FCM	K-means	Improved K-means
95 %	1.83	1.79	1.56	1.36	1.25
90 %	1.79	1.77	1.51	1.28	1.13
85 %	1.71	1.68	1.43	1.21	0.95
80 %	1.66	1.56	1.35	1.13	0.88
75 %	1.51	1.49	1.29	0.96	0.83
70 %	1.42	1.41	1.22	0.86	0.79
65 %	1.34	1.32	1.16	0.78	0.76
60 %	1.26	1.13	0.96	0.73	0.71

Data availability

All data generated or analysed during this study are included in this published article.

CRediT authorship contribution statement

Wei Zhang: Investigation. **Zonghua Wu:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Iwendi, E. Ibeke, H. Eggoni, S. Velagala, G. Srivastava, Pointer-based item-to-item collaborative filtering recommendation system using a Machine learning model, Int. J. Inf. Technol. Decis. Making 21 (1) (2022) 463–484.
[2] Adriano Mourthé, C.E. Mello, Less is more: improving neural-based collaborative filtering by using landmark modeling, Inf. Sci. 590 (11) (2022) 217–233.
[3] H.C. Yan, Z.R. Wang, J.Y. Niu, T. Xue, Application of covering rough granular computing model in collaborative filtering recommendation algorithm optimization, Adv. Eng. Inf. 51 (8) (2022) 101485–101495.
[4] Z. Liu, L. Wang, X. Li, A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm, J. Manuf. Syst. 58 (12) (2021) 348–364.
[5] D.P.M. Abellana, P.E. Mayol, A novel hybrid DEMATEL-K-means clustering algorithm for modeling the barriers of green computing adoption in the Philippines, J. Model. Manag. 17 (2) (2022) 486–517.
[6] S. Hadianti, D. Riana, Segmentation and analysis of Pap smear microscopic images using the K-means and J48 algorithms, Jurnal Teknologi dan Sistem Komputer 9 (2) (2021) 113–119.
[7] Jiaquan Huang, Zhen Jia, Zuo Peng, Improved collaborative filtering personalized recommendation algorithm based on k-means clustering and weighted similarity on the reduced item space, Mathematical Modelling and Control 3 (1) (2023) 39–49.
[8] F. Zhang, W. Ye, G. Lei, Y. Liu, X. Wang, SOH estimation of Li-ion battery based on FA-BPNN-K-means optimization algorithm, Comput. Methods Sci. Eng. 22 (4) (2022) 1209–1222.
[9] Z. Fang, C. Chiao, Research on prediction and recommendation of financial stocks based on K-means clustering algorithm optimization, Journal of Computational Methods in Sciences and Engineering 21 (5) (2021) 1081–1089.
[10] T.M. Jawad, N.A. Ali, Using K-means clustering algorithm with Power 1, International Journal of Computer Science Engineering and Information Technology 6 (1) (2021) 9–13.
[11] Z. Liu, L. Wang, X. Li, S. Pang, A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm, J. Manuf. Syst. 58 (2021) 348–364.
[12] C.L. Wang, Y. Wang, Z.Y. Zeng, C.Y. Lin, Q.L. Yu, Research on Logistics Distribution vehicle scheduling based on heuristic genetic algorithm, Complexity 2021 (11) (2021) 1–8.
[13] L. Wu, X. Ye, Y. Zhang, J. Gao, Z. Lin, B. Sui, X. Bo, A genetic algorithm-based Ensemble learning framework for drug combination prediction, J. Chem. Inf. Model. 63 (12) (2023) 3941–3954.

[14] Y. Wang, L. Han, Q. Qian, J. Xia, J. Li, Personalized recommendation via multi-dimensional Meta-paths Temporal graph Probabilistic Spreading.Information processing management: Libraries and information Retrieval systems and Communication networks, Int. J. 59 (1) (2022) 100–125.
[15] Y. Chen, Y. Dai, X. Han, Y. Ge, P. Li, Dig users' intentions via attention flow network for personalized recommendation, Inf. Sci. 547 (32) (2021) 1122–1135.
[16] Paromita Nitu, Joseph Coelho, Praveen Madiraju, Improvising personalized Travel recommendation system with Recency effects, Big Data Mining and Analytics 4 (3) (2021) 139–154.
[17] Addagarla Ssvr Kumar, Amalanathan Anthoniraj, Probabilistic unsupervised Machine learning approach for a similar image recommender system for E-commerce, Symmetry 12 (11) (2020) 1783, 1783.
[18] N. Vara, M. Mirzabeigi, H. Sotudeh, S.M. Fakhrahmad, Application of k-means clustering algorithm to improve effectiveness of the results recommended by journal recommender system, Scientometrics 127 (6) (2022) 3237–3252.
[19] H.R. Kazemi, Kaveh Khalili ɓ amghani, Soheil Sadikmezhad, Tuning structural parameters of neural networks using genetic algorithm: a credit scoring application, Expet Syst. 38 (7) (2021) 1–24.
[20] J. Zhang, Y. Yao, W. Sun, L. Tang, X. Li, H. Lin, Application of the non-dominated sorting genetic algorithm II in multi-objective optimization of Orally Disintegrating Tablet Formulation, AAPS PharmSciTech 23 (6) (2022) 1–7.
[21] X. Zhao, S. Zou, Z. Ma, Decentralized Resilient H∞ load frequency Control for Cyber-Physical power systems under DoS Attacks, IEEE/CAA Journal of, Automatica Sinica 8 (11) (2022) 1737–1751.
[22] N. Elgharably, S. Easa, A. Nassef, A.E. Damatty, Stochastic multi-objective vehicle Routing model in green Environment with customer Satisfaction, IEEE Trans. Intell. Transport. Syst. 24 (1) (2023) 1337–1355.
[23] X. Chen, H. Deng, Research on personalized recommendation methods for online Video learning Resources, Appl. Sci. 11 (2) (2021) 804–815.
[24] M. Al-Ghamdi, H. Elazhary, A. Mojahed, Evaluation of collaborative filtering for recommender systems, Int. J. Adv. Comput. Sci. Appl. 12 (3) (2021) 102–113.
[25] S. Choudhuri, S. Adeniye, A. Sen, Distribution Alignment using Complement Entropy objective and adaptive Consensus-based Label Refinement for partial domain adaptation, Artificial Intelligence and Applications 1 (1) (2023) 43–51.
[26] B. Pérez-Canedo, J.L. Verdegay, On the application of a lexicographic method to fuzzy linear programming problems, Journal of Computational and Cognitive Engineering 2 (1) (2023) 47–56.
[27] Y. Liu, S.L. Wang, J.F. Zhang, A neural collaborative filtering method for identifying miRNA-disease associations, Neurocomputing 422 (5) (2021) 176–185.
[28] R. Ravanifard, A. Mirzaei, W. Buntine, Content-Aware Listwise collaborative filtering, Neurocomputing 461 (23) (2021) 479–493.
[29] Z. Cai, G. Yuan, S. Qiao, FG-CF: Friends-aware graph collaborative filtering for POI recommendation, Neurocomputing 488 (6) (2022) 107–119.
[30] M.K. Najafabadi, A. Mohamed, M.A.B. Nair, An effective collaborative user model using hybrid clustering recommendation methods, Ingénierie Des. Systèmes Inf. 26 (2) (2021) 151–158.