International Journal of Computational Intelligence Systems (2023) 16:142
https://doi.org/10.1007/s44196-023-00321-7

RESEARCH ARTICLE

E-commerce User Recommendation Algorithm Based on Social Relationship Characteristics and Improved K-Means Algorithm

Xia Shen[1]

Received: 16 June 2023 / Accepted: 18 August 2023

Abstract

In the era of the Internet, information data continue to accumulate, and the explosive growth of network information explosion leads to the reduction of the accuracy of users' access to information. To enhance the user experience and purchasing desire of e-commerce users, a e-commerce user recommendation algorithm based on social relationship characteristics and improved K-means algorithm is proposed. It combines the Automatic Time Division Dynamic Topic Model based on adaptive time slice division for building a strength calculation model in view of the characteristics of social relations. Then, it proposes an e-commerce user recommendation algorithm in view of the improved K-means algorithm to improve the accuracy of topic feature extraction and user recommendation. The experiment illustrates that there is no fluctuation in the clustering function of the improved K-means algorithm, and the highest, lowest, and average accuracy remain consistent under the three datasets, with average accuracy of 78.9%, 84.5%, and 95.9%, respectively. The community discovery-based friend recommendation algorithm presented in the study has the highest accuracy, illustrating that improving the K-means algorithm can further improve recommendation accuracy. The accuracy of the feature extraction method in view of alternative cost is 0.63, which improves the accuracy by about 5%. The results indicate that this study can provide technical support for user recommendations on e-commerce platforms.

Keywords Social relations · Characteristic strength · Time division · Improved K-means algorithm · E-commerce users · Friend recommendation

1 Introduction

As the advancement of e-commerce (EC) scale and the continuous improvement of EC technology and services, more and more users are mainly focused on the convenience of EC platforms, and EC users are also rapidly increasing [1]. The diversified advancement of EC platforms has led to a rapid increase in information volume, resulting in information overload issues that seriously affect users' search preferences for products [2]. It costs lots of time for users to choose their preferred products among numerous products, leading to a decrease in their purchase rate and satisfaction with the platform. Under the exponential growth of information data, businesses, users, and EC platforms are facing the problem of information overload, and search engines and recommendation algorithms (RA) have emerged [3]. The search engine utilizes users' clear needs to search and provide them with the necessary information. However, there are also issues with unclear keyword conversion and return lists that do not meet user needs [4]. The RA does not rely on keywords, but only excavates information based on historical data, and achieves personalized recommendations through user preference (UP) judgment. Personalized recommendation systems can improve user satisfaction and purchase rate through UP mining and effective product recommendation, while reducing recommendation time and cost [5]. The recommendation system is also being optimized and upgraded through continuous use, constantly meeting the personalized needs of users. In addition, it achieves precise behavior prediction by mastering UPs, further improving the efficiency of user decision-making. RA are an important part of recommendation systems, which can judge UP information based on the interaction behavior between users and projects, as well as the influence of users' intimate behavior and friends. RAs are widely used for personalized recommendations

✉ Xia Shen
jscjxysxia@163.com

[1] School of Economics and Management, Jiangsu Vocational College of Finance and Economics, Huai'an 223003, China

Published online: 31 August 2023

for users on EC platforms such as Taobao and Tmall. This algorithm mainly manifests as rating prediction and Top-K recommendation, where the basic data for rating prediction are user comment information; Top-K recommendations are based on implicit data such as user browsing, favorites, and purchases [6]. In social networks, users' social relationships are not just a simple form of attention or friendship, but also require calculation of the strength of social relationships. In Granot's statement on the strength of social relationships, it is believed that the transmission of information mainly relies on weak relationships, rather than strong relationships. This indicates that information in unfamiliar weak relationship networks with the same interests spreads quickly and is widely distributed, and can provide users with information that cannot be obtained from strong relationship networks. Therefore, quantitative analysis of the strength of users' social relationships in social networks is very important. For improving the EC recommendations, an EC user RA based on social relationship characteristics and improved K-means algorithm is proposed, aiming to improve the experience and purchase rate of EC users. The research is separated into four parts. The first part is a summary of existing RA research; the second part is the design of EC user RA based on social relationship characteristics and improved K-means algorithm; the third part is the experiment and analysis of EC user RA in view of social relationship characteristics and improved K-means algorithm. The fourth part is a summary of the entire article.

2 Related Works

RA is an algorithm in computer science. This algorithm is mainly applied in various fields of online recommendation, using some historical behavioral data of users to infer the products or services that users may like. N. The Xie research team proposed a user Kansei demand acquisition method based on the double matrix RA, combining the collaborative filtering algorithm and personalized double matrix RA, to achieve the prediction of user perceptual demand. Research has revealed that this method can achieve accurate and timely prediction when there are a large amount of data [7]. X. Ding´s research team presented a privacy protection algorithm in view of group recommendation. This algorithm analyzes the privacy issues of perceptual recommendation systems through questionnaire surveys, combines user privacy concerns classification and group recommendations, and experiments with user data protection. Experimental data show that the recommendation quality of this algorithm is good [8]. Y. Chen's research team proposed a personalized RA in view of UP in mobile EC. This algorithm combines multi-criteria scoring in view of UP for projects to construct a UP model, and utilizes UP clustering for enhancing the personalized recommendations. Simulation data show that this algorithm can effectively improve recommendation quality [9]. D. Shin and his team members proposed a news RA experience model that integrates cognitive, emotional, and behavioral factors into a heuristic process, understanding feature transformation and interaction through the user's cognitive process. Research has revealed that user centered development and algorithmic service evaluation provide higher accuracy [10]. D. Xiang's research team proposes a simple and effective personalized recommendation method for cross-border EC. This method utilizes a hybrid recommendation model in view of complex UP features for mining UP features and achieve personalized product recommendations based on user behavior preferences. The results indicate that the algorithm reduces data sparsity [11]. H. Cong presents a new personalized RA for film and television culture in view of intelligent classification algorithm to address the shortcomings of the film and television culture recommendation system. This algorithm uses the traditional collaborative filtering RA to implement data filtering, and combines the user recommendation set to compensate for the relevant problem of the system. The simulation indicates that this algorithm enhances the accuracy of personalized recommendation by 15% [12].

Relationship strength is a sustained emotional intensity, intimacy, and service exchange function; the strength of relationships can be divided into strong and weak relationships. B. Zhang's research team has proposed a multidimensional comprehensive recommendation method in view of user relationship strength, which generates candidate sets through user relationship strength calculation and calculates users' interest in candidate set entities. The experiment reveals that this method can effectively improve recommendation accuracy [13]. A. Chader research team proposes a social filing method based on weight perception communities. This method utilizes self-friend relationships to construct user contact structures, and combines the power of friend relationships to achieve user neighbor connections, achieving the most realistic description of the network structure. The experiment depicts that this method can accurately predict user interests [14]. Y. Pan research team utilizes trust relationships between users to construct role-based adaptive trust strength. Then, they guide the training process of potential trust intensity by establishing a relationship between potential trust intensity and UP, and integrate role-based trust intensity into the recommendation model. Experimental data show that the function of this method exceeds traditional recommendation models [15]. The K-means algorithm (KMA) is the most notable partition clustering algorithm (CA), making it the most extensively utilized among all CA as its efficiency. G. Shen's research team used sample density and tree crown to optimize the KMA, and combined it with Canopy algorithm to cluster the original sample data. Simulation experiments were conducted on a standard dataset with a self-built dataset. The results indicate that the clustering results of this algorithm are more accurate and run faster [16]. Z. Feng and J. Zhang, inspired by the mixed model KMA and the expectation maximization algorithm, proposed a KMA for non-parametric mixed regression and mixed Gaussian process estimation, and conducted numerical simulation, comparison, and analysis on real datasets. The simulation indicates that the algorithm has effectiveness and competitiveness [17]. F. Moodi research team proposes an improved K-means CA that iterates in view of changes in the distance from points to the centroid and eliminates points with distances greater than the threshold. The experiment indicates that the clustering accuracy of this method can be improved by 41.85%, which is helpful for big data analysis [18].

In total, many researchers have carried research and design on user RA, relationship strength, and K-means CA, but the applicability of EC user recommendations still needs to be improved. Therefore, this study proposes an EC user RA based on social relationship characteristics and improved KMA, aiming for enhancing the accuracy of user recommendations and promote the further development of EC recommendations.

3 Design of E-Commerce User Recommendation Algorithm Based on Social Relationship Characteristics and Improved K-Means Algorithm

The accuracy of traditional topic model is low. In the first section, a dynamic topic model in view of adaptive time slice division is proposed. This model combines unsupervised Bayesian models for iteration, achieving document set training and time slice partitioning. The time slices obtained by partitioning are used as training parameters for the dynamic model, to obtain the obtain the text topic probability distribution (TTPD) and topic word probability distribution (TWPD). The feature extraction (FE) method based on substitution cost is used to extract features such as user behavior. The second section of this chapter proposes an EC user RA in view of the improved KMA. This algorithm finds the initial clustering center (CC) through the calculation of inter-class distance and sample mean within the cluster, and then uses the KMA for clustering. Finally, it utilizes a friend RA based on community discovery to complete EC user recommendations.

3.1 Design of Strength Calculation Model Based on Social Relationship Characteristics

The structure of social networks affects information dissemination and user communication. In complex EC networks, social relationships contribute to the extraction of effective feature attributes [19]. This study utilizes FE and model transformation methods to transform the network structure of social EC platforms and represent data information using the strength of social relationships. The traditional topic model does not consider the time factor, and its accuracy of topic extraction is low. The basic dynamic topic model takes into consideration the temporal type of the text and divides the time of the text, but fails to consider the difference of topics in adjacent time slices [20]. To improve the accuracy of topic extraction, this research proposes the Automatic Time Division Dynamic Topic Model (ATD-DTM) based on adaptive time slice division. The vector representation of the theme is illustrated in Eq. (1):

$$\overline{v}_i^j = \frac{1}{N}\sum_{n=1}^{N} v_{in}^j \tag{1}$$

In Eq. (1), the time slice division is represented by j; the topic vector is represented by $\overline{v}_i^j$; the number of feature words with the highest probability is N; the feature word vector is represented by v_{in}^j. The calculation of topic difference is illustrated in Eq. (2):

$$D_{j,j+1} = \frac{1}{K}\sum_{q=1}^{K} KL_{\min q}^{j,j+1} \tag{2}$$

In Eq. (2), the topic difference is represented by $D_{j,j+1}$; the number of theme difference vector elements is represented by K; the minimum distance between two adjacent time slots is represented by $KL_{\min q}^{j,j+1}$. The comprehensive similarity calculation of time slices is illustrated in Eq. (3):

$$S_{j,j+1} = \frac{1}{D_{j,j+1}} \tag{3}$$

In Eq. (3), the comprehensive similarity of time slices is represented by $S_{j,j+1}$. The comprehensive similarity and difference of time slices are inversely proportional. The process of adaptive time slice partitioning method is illustrated in Fig. 1.

Research will use unsupervised Bayesian models as iterative models to achieve document set training and time slice partitioning. The adaptive time slice partitioning method is as follows: first, two initial adjacent time slices are set, and the minimum time slice and upper time limit are set. A certain maximum value is used as the initial value of time slice similarity, and a text set is constructed based on the initial adjacent time slices. Then, an unsupervised Bayesian model is used to train the text set, and the calculated time slice similarity is used to determine whether the iteration will continue, divide time slices through continuous iteration and round-trip operations. The divided time slices are used as training parameters

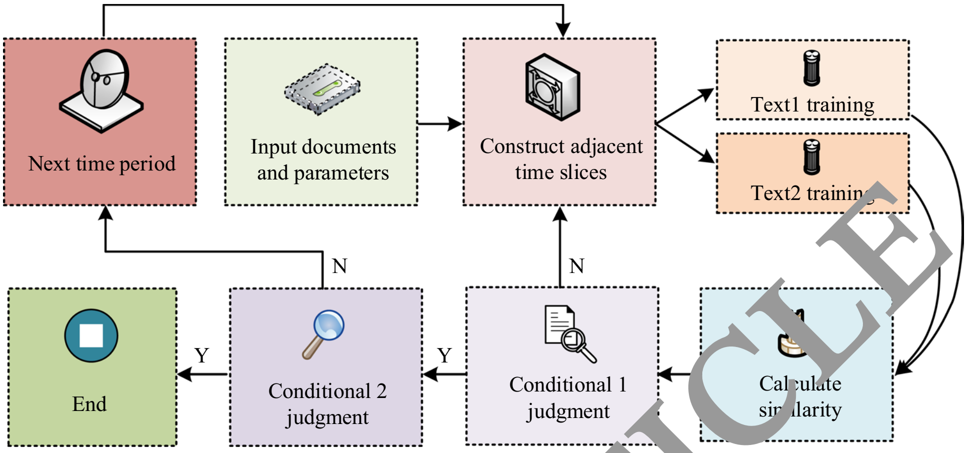


Fig. 1 Process of adaptive time slice partitioning method

for the dynamic model to obtain the TTPD and TWPD. Under various reasons, user interests will change over time. The equation for calculating time weight is illustrated in Eq. (4):

$$f(t) = \exp(\frac{t}{T_0}) \tag{4}$$

In Eq. (4), the occurrence time of the project is represented by t; its time weight is represented by $f(t)$; the attenuation index is represented by T_0. The time weight reconstruction formula combined with adaptive time slice partitioning is illustrated in Eq. (5):

$$f(t) = \exp(-\frac{n-t}{T_0}) \tag{5}$$

In Eq. (5), the current time is represented by n. The earlier the time, the smaller the project time weight and its impact on user interest. The formula for calculating the user comment vector is indicated in Eq. (6):

$$q_u = \sum_{j=1}^{n} \exp(-\frac{n-t}{T_0}) q_u^j \tag{6}$$

In Eq. (6), the user comment vector is represented by q_u, and the comment vector of user u in time slot j is represented by q_u^j. There are three commonly used similarity calculations: cosine similarity (CS), Pearson correlation coefficient, and Jaccard similarity. CS refers to the cosine angle value between two vectors in a spatial model. The study uses CS to measure user similarity, and the formula for calculating user comment similarity is indicated in Eq. (7):

$$S_{comment}(u,v) = \frac{\sum_{k=1}^{K} q_{u,k} \times q_{v,k}}{\sqrt{\sum_{k=1}^{K} q_{u,k}^2} \times \sqrt{\sum_{k=1}^{K} q_{v,k}^2}} \tag{7}$$

In Eq. (7), the similarity of user comments should be represented by $S_{comment}(u,v)$; the probability distributions of users u and v on the k-th topic are represented by $q_{u,k}$ and $q_{v,k}$, respectively. The calculating the strength of social relationships based on comment features is indicated in Fig. 2.

The document set contains a collection of comments and product information; its preprocessing is divided into word segmentation and word removal, and an unsupervised Bayesian model is used for iteration to obtain adaptive time division results. The dynamic topic model trains the partition results to obtain the TTPD and TWPD, and then obtains the user comment vector through attenuation calculation, and uses the user comment similarity calculation results to measure user interest preferences.

The types of components in social networks are diverse, and using only textual information to measure user relationships is not comprehensive. Extracting effective features can help enhance the recommendation models [21]. The traditional method of extracting user behavior features is simple and lacks in-depth analysis of behavior features. Therefore, the study presented a FE method in view of alternative cost. The calculation of product effectiveness is indicated in Eq. (8):

$$e_{s,u,i} = p_u^T \cdot q_i \tag{8}$$

In Eq. (8), the effectiveness of the product is represented by $e_{s,u,i}$; user features are represented by p_u; the product features are represented by q_i. The calculation of substitute efficacy value is indicated in Eq. (9):

$$e_{s,u,i_{sub}} = \max_{i \in I(s) \backslash i_b} e_{s,u,i} \tag{9}$$

In Eq. (9), the effectiveness value of the substitute is represented by $e_{s,u,i_{sub}}$; the set of unit sequence products is represented by $I(s)$; the products purchased by users are represented by i_b. Based on product frequency and sorting, as well as different types of user behavior, a behavior impact function $g(s,u,i)$ is proposed, and its calculation formula is indicated in Eq. (10):

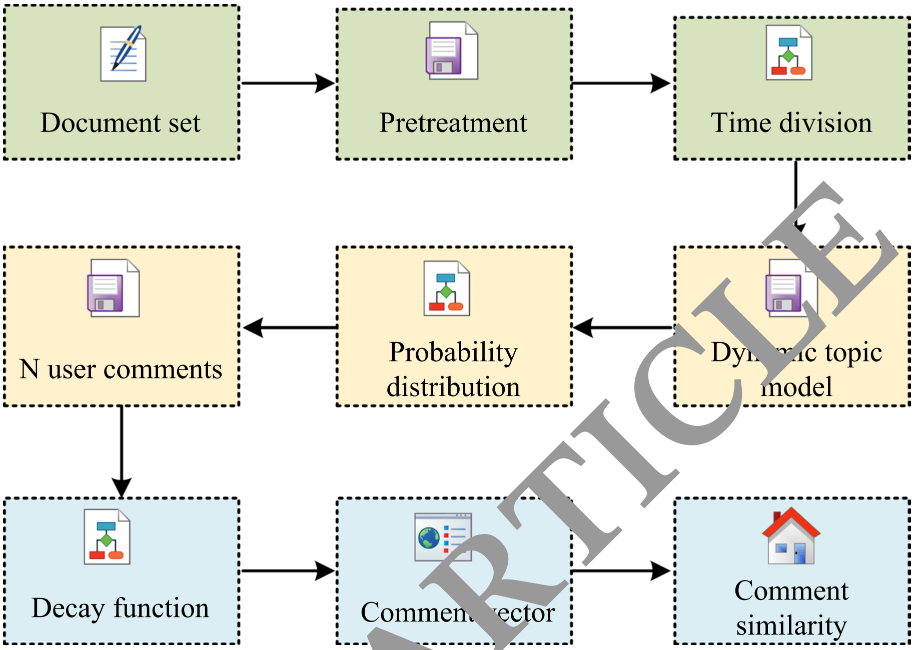


Fig. 2 A model for calculating the strength of social relationships based on comment features

$$g(s,u,i) = \sum\nolimits_{r \in R(s,u,i)} \exp\left[-\left(\mu \cdot \frac{N - r_{collect}}{N} + (1-\mu) \cdot \frac{N - r_{click}}{N}\right)\right] \tag{10}$$

In Eq. (10), the sequence length is served as by N; the set of product sequence rankings is served as by $R(s, u, i)$, the browsing behavior, bookmarking behavior, and the weights are represented by r_{click}, $r_{collect}$, and $r_{collect}$. The CS calculation formula based on behavioral characteristics is demonstrated in Eq. (11):

$$S_{behavior}(u,v) = \frac{\sum_{m=1}^{N} p_{u,m} \times p_{v,m}}{\sqrt{\sum_{m=1}^{N} p_{u,m}^2} \times \sqrt{\sum_{m=1}^{N} p_{v,m}^2}} \tag{11}$$

In Eq. (11), the similarity of user behavior is represented by $S_{behavior}(u, v)$; the m th feature in the user u and v behavior feature vectors is represented by $p_{u,m}$ and $p_{v,m}$, respectively. The first step of the social relationship strength calculation method based on behavioral characteristics is to preprocess behavioral data, then complete the division of row sequences in the sample set, calculate each substitute product, and combine gradient descent method for iterative optimization until convergence. The similarity of user personal information is a supplement to the similarity of user comments and behavior. Users of similar age, distance, and gender have similar consumption preferences. The relevant calculating formula is demonstrated in Eq. (12):

$$S_{info}(u,v) = w_1 S_{age}(u,v) + w_2 S_{dis}(u,v) + w_3 S_{sex}(u,v) \tag{12}$$

In Eq. (12), the similarity of personal information is represented by $S_{info}(u, v)$; age similarity is represented by $S_{age}(u, v)$; the distance similarity is represented by $S_{dis}(u, v)$; gender similarity is represented by $S_{sex}(u, v)$; the sum of the weight parameters w_1, w_2, and w_3 is 1. The formula for calculating the total similarity of users is demonstrated in Eq. (13):

$$S_{total} = \alpha \cdot S_{comment} + \beta \cdot S_{behavior} + \gamma \cdot S_{info} + b \tag{13}$$

In Eq. (13), the total similarity of users is represented by S_{total}; the weights of user comments, behavior, and personal information similarity are represented by α, β, and γ, respectively; the sum of the three is 1, and the constant term is represented by b. The calculation model for social relationship strength is demonstrated in Fig. 3.

A social EC platform includes a set of user nodes, a set of product nodes, and a set of edges in users and products. The collection information between users and products is relatively complex. The model is analyzed through user comments and titles, combined with the dynamic topic model of adaptive time division to achieve probability solution, and the total similarity is calculated through the similarity of user comments, behaviors, and personal information.

3.2 Design of E-commerce User Recommendation Algorithm Based on Improved K-Means Algorithm

The increase in internet capacity makes it difficult to effectively improve computing speed and recommendation accuracy [22]. Therefore, the study proposes a friend RA based on community discovery, combined with an improved KMA for achieving user community partitioning. The KMA is a classic CA. The idea is for extracting K points as the CCs,

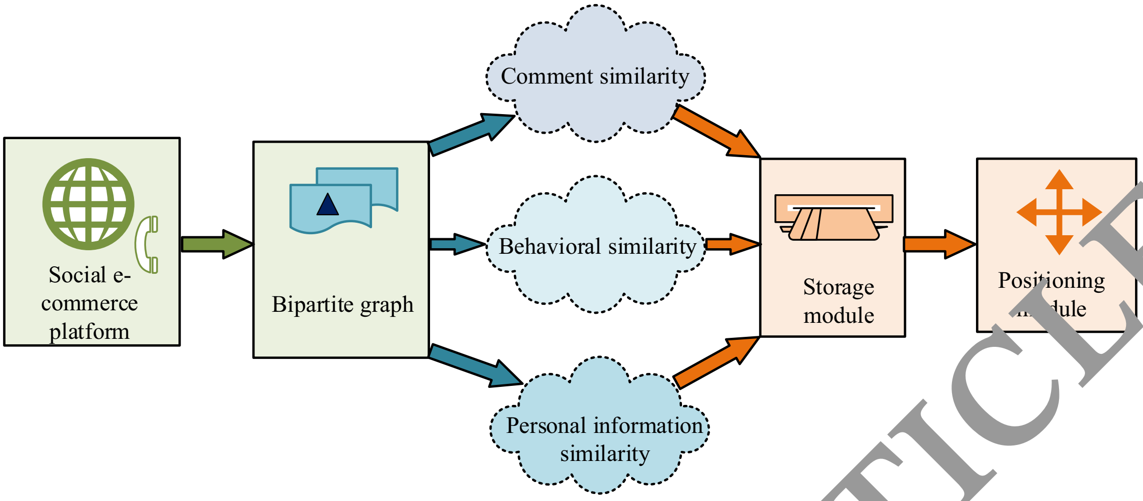


Fig. 3 A model for calculating the strength of social relationships

and use the distance measurement formula to calculate the distance between the sample and the CC. By redividing the clustering and calculating the mean, the calculation is repeated until the CC remains unchanged or the target converges [23]. The linear time complexity of the KMA makes it efficient and easy to apply, but it requires inputting the number of clusters in advance for experienced personnel to successfully complete this step. The randomness of initial clustering leads to unstable clustering results. The clustering effect under different CCs is demonstrated in Fig. 4.

Figure 4a shows the original distribution of the dataset; Fig. 4b shows that a more suitable CC was selected, resulting in ideal clustering results; Fig. 4c shows the selection of unsuitable CC, resulting in less-than-ideal clustering results. From this, it illustrates that the traditional KMA is influenced by the initial CC. When the initial CC is relatively discrete, the clustering effect is better. K-means++ is one of the improved algorithms of KMA, with the idea that the initial CC should be as far away as possible [24]. The first step is to randomly select an initial CC in the sample dataset, and calculate the minimum distance and selection probability from the data point to the CC. Then, combined with roulette wheel, complete the selection of the next initial CC, repeat the operation until all the initial CC are chosen, and finally use the traditional KMA for clustering. The improved KMA proposed in this paper takes the density distribution and distance as the criteria for selecting the initial CC. The formula for calculating the average distance between data points is demonstrated in Eq. (14):

$$Rad = \frac{1}{C_{|D|}^{2}} \times \sum d(x_i, x_j) \tag{14}$$

In Eq. (14), the dataset is represented by D; the average distance or neighborhood radius of all data is represented by Rad; the quantity of samples in the dataset is represented by $|D|$; the number of sets of any two data points is represented by $C_{|D|}^{2}$; the distance between two data points is represented by $d(x_i, x_j)$. The distribution density of data points is calculated as demonstrated in Eq. (15):

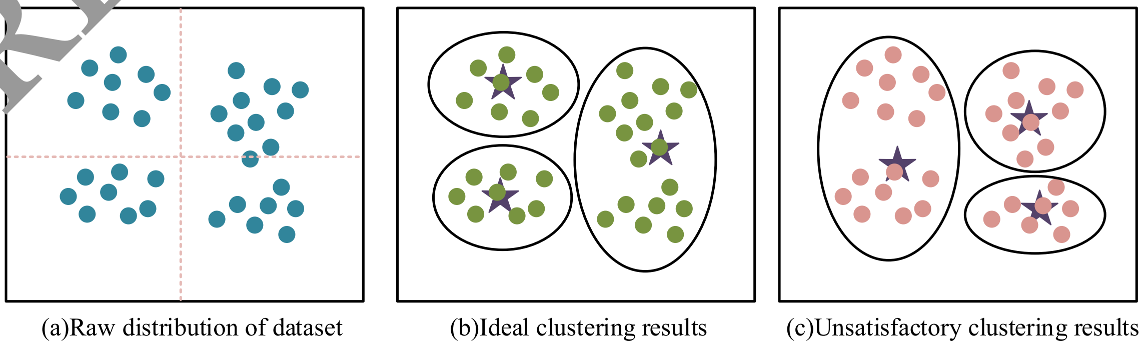

(a)Raw distribution of dataset (b)Ideal clustering results (c)Unsatisfactory clustering results

Fig. 4 A model for calculating the strength of social relationships

$$Dens(x) = \sum_{i=1}^{|D|} f(Rad - |x_i - x|) \tag{15}$$

In Eq. (15), the distribution density of data points is represented by $Dens(x)$; the choice function is represented by $f(x)$, and its values are 0 and 1. The minimum distance calculation formula for data points between two types is demonstrated in Eq. (16):

$$dis(k) = \min_{x_i \in n_i, x_j \in n_j} d(x_i, x_j) \tag{16}$$

In Eq. (16), the minimum distance between data points between two classes, i.e., the distance between classes, is represented by $dis(k)$. The input of the improved KMA is the dataset and the quantity of clusters, and the maximum distribution density cluster is obtained by calculating the neighborhood radius and distribution density. By calculating the distance between classes and the mean of samples in the cluster, the initial CC is found, and then the KMA is used for clustering. The process of friend RA based on community discovery is demonstrated in Fig. 5.

Friends in socialized EC platforms refer to the users who initiate the group recommended by the system. If the target user accepts the original user's group, the system successfully implements the user recommendation and the two users establish a friend relationship [25, 26]. The input of the friend RA in view of community discovery is the number of clusters and the length of the recommendation list. It constructs a user feature vector matrix through user comment vectors, user behavior vectors, and user personal information vectors, which is used as input for the improved KMA, and then divides user communities. It calculates the range in the target user and the CC and selects the nearest community as the nearest neighbor search space. It calculates the similarity in users in the nearest neighbor search space and forms a user recommendation list as the output of the algorithm. The study evaluates the effectiveness of algorithm recommendations through accuracy, which is defined as the proportion of successfully recommended friends to the number of user views among the friends recommended by the target user. The recommended accuracy calculation formula is demonstrated in Eq. (17) [27]:

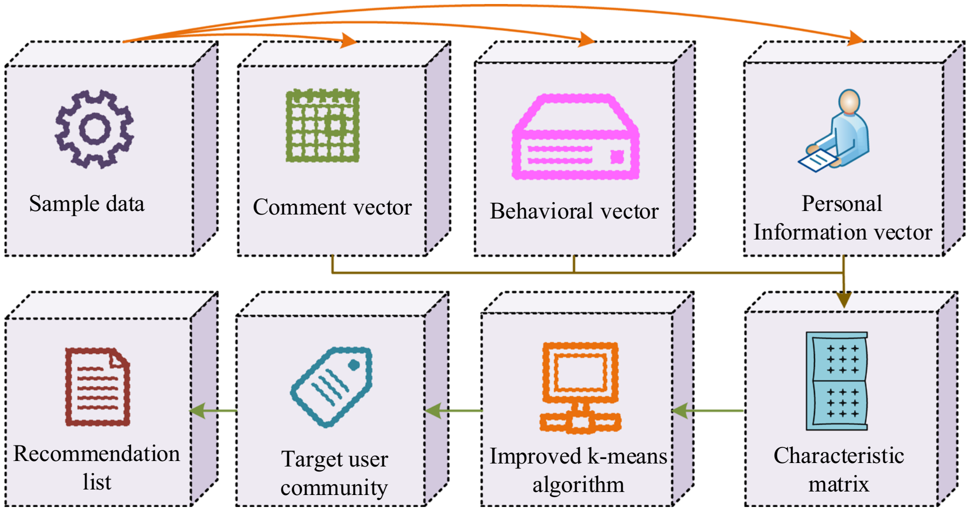
$$precision = \frac{T_{sf}}{U_l} \tag{17}$$

In Eq. (17), the recommended accuracy is represented by *precision*; the number of successfully recommended friends is represented by T_{sf}; the number of user views is represented by U_l.

4 Experiment and Analysis of E-commerce User Recommendation Algorithm Based on Social Relationship Characteristics and Improved K-Means Algorithm

The first section of this chapter first proposes to extract a dataset from a certain social EC merchant, which includes the consumption situation of 1500 users. Then the effect of the ATD-DTM model proposed in the study is verified, and the potential Dirichlet distribution model and the dynamic topic model are compared. The experiment analyzes user behavior data for testing the rationality of the construction of the behavior influence function, and then compares traditional FE methods based on splicing for testing the presented FE method in view of alternative costs. The second section of this chapter first proposes a test dataset selected from the Glass, Iris, and Wine datasets, and compares it with the traditional KMA and K-means++ algorithm for testing the effectiveness of the EC user RA based on the improved KMA. Finally, it compares the traditional concatenated

Fig. 5 Process of friend recommendation algorithm based on community discovery



feature recommendation methods for testing the presented FE method in view of alternative costs.

4.1 Verification of Model Effectiveness Based on Social Relationship Feature Intensity Calculation

For testing the effectiveness of the social relationship feature intensity calculation model, a dataset was extracted from a certain social EC merchant, which includes the consumption situation of 1500 users. The division of the dataset is demonstrated in Table 1.

The experimental dataset is separated into three parts. The first part is the user personal information dataset that includes user ID, gender, age, and location; The second part is a user behavior dataset that includes users, products, their IDs, and times; The third part is a user evaluation dataset that includes information such as users, products, comments, and time; Due to commercial confidentiality considerations, user, product, and their category IDs need to be desensitized and finally represented in numbers, with time information accurate to the hour. For verifying the ATD-DTM model presented by the research, the experiment compares the potential Dirichlet distribution (LDA) model with the dynamic topic model (DTM). The determination of the number of topics of each model is demonstrated in Fig. 6.

In Fig. 6, Perplexity is the evaluation standard for determining the effect of the number of model themes. The more its value is, the better the model effect will be. When the number of topics is 5, the Perplexity of LDA model is 700; the Perplexity of DTM model is 595; the minimum Perplexity of ATD-DTM model is 545; the curve is relatively stable when the quantity of topics is small, illustrating that the model has better explanatory power, the clustering effect, and the accuracy of topic extraction are the highest. When the quantity of topics is 17, the Perplexity of the model is basically stable, and the quantity of topics is set to 17. The distribution of product sorting and user behavior is revealed in Fig. 7.

Figure 7a shows that the frequency of purchasing goods is mostly distributed in the top 40%, indicating that the more frequently products appear, the easier they are to be purchased; The ranking of purchased products is mostly

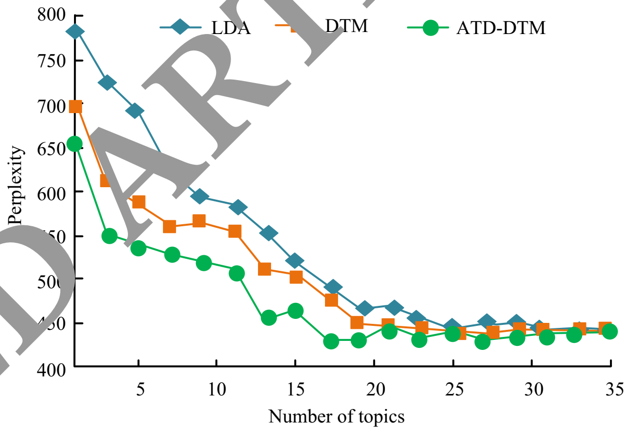


Fig. 6 Determination of the number of topics in each model

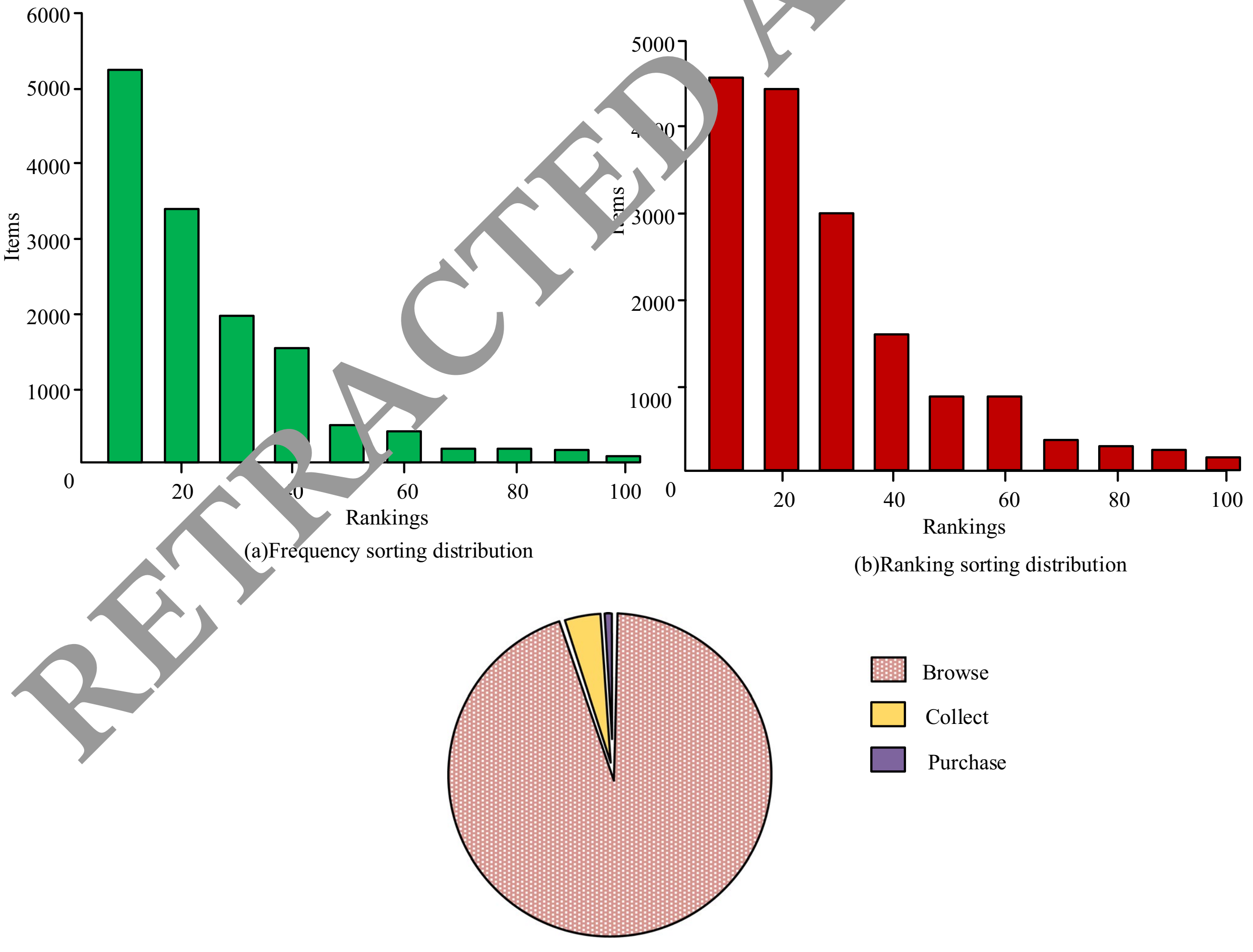
Table 1 Dataset partitioning

Dataset name	Field	Description	Extraction instructions
User personal information	Sex	Gender identification	Male and female are represented by 0 and 1, respectively
	Age	Age identification	/
	Location	User location	/
User behavior	User ID	User identification	Field desensitization
	Item ID	Item identification	Field desensitization
	Item category ID	Item category identification	Field desensitization
	Behavior type	Types of user behavior towards products	Browse, bookmark, and purchase values 1, 2, and 3
	Time 1	Behavioral time	Accurate to the hour
User feedback	Item information	Product title information	/
	Comment	User comments	/
	Time 2	User comment time	Accurate to the hour
	Record ID	Users in domain name identification	Field desensitization
	User ID	User identification	Field desensitization
	Item ID	Item identification	Field desensitization
	Item category ID	Product type identification	Field desensitization

distributed in the top 60%, indicating that the lower the ranking of product operations, the higher the probability of purchase. In Fig. 7b, browsing behavior accounts for 95%, collecting behavior accounts for 4%, and purchasing behavior accounts for 1%, indicating that the purpose of collecting is stronger than browsing. The data indicate that the construction of the behavior influence function is reasonable. The collection weight analysis results of behavior influence function under the training of Logistic Regression (LR), Gradient Boosting Decision Tree (GBDT) and random forest (RF) are revealed in Fig. 8.

In Fig. 8, under the three classification models, the increase in the weight of the collection behavior leads to an initial increase and then a decrease in F1 scores. When the weight of collection behavior is 0.8, the gradient lifting decision tree and random forest model achieve the best training effect, with F1 values of 0.82 and 0.83 respectively; when the weight of the collection behavior is 0.9, the logistic regression model achieves the best effect, with an F1 value of 0.77. According to the voting law, set the weight of the collection behavior to 0.8. For validating the presented alternative cost-based FE method (labeled A), the experiment compared the traditional concatenation extraction method (labeled B). The predicted comparison results under LR, GBDT, and RF model training are revealed in Fig. 9.

Figure 9 shows that under the training of LR model, GBDT model, and RF model, the highest F1 scores of the FE method based on alternative cost are 0.7[illegible]5, 0.845, and 0.821, respectively; The highest F1 scores of traditional stitching FE methods are 0.7[illegible]2, 0.[illegible]10, and 0.782, respectively. According to the analysis of the experimental dataset, the F1 score of the FE method based on alternative cost is the highest among the three models, indicating that the prediction effect of this method is good and can provide high FE accuracy and prediction accuracy.



(a)Frequency sorting distribution

(b)Ranking sorting distribution

(c)Proportion of user product interaction behavior

Fig. 7 Distribution of product ranking and proportion of user behavior

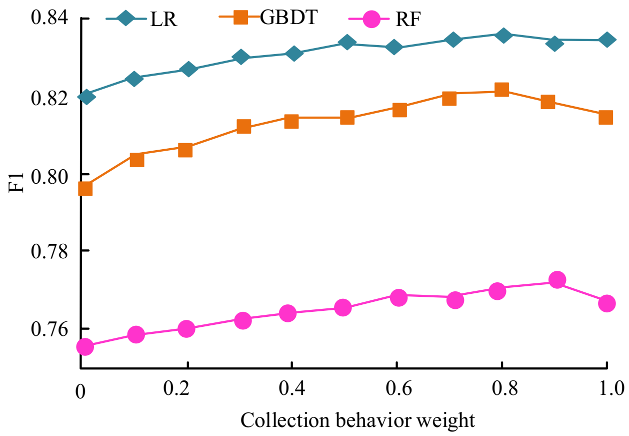


Fig. 8 The collection weight analysis results of the behavior influence function

4.2 Verification of the Effectiveness of E-commerce User Recommendation Algorithm Based on Improved K-Means Algorithm

For testing the effectiveness of the EC user RA based on the improved KMA, the experiment selected data from the Glass, Iris, and Wine datasets as the test dataset, and compared it with the traditional KMA and K-means++ algorithm. The clustering test dataset division is shown in Table 2.

As shown in Table 2, the sample sizes for the Glass, Iris, and Wine datasets are 625, 150, and 178, respectively, with feature attribute dimensions of 4, 4, and 13, and category numbers of 3. The clustering effects of each algorithm under each dataset are revealed in Table 3.

Table 3 shows that the traditional KMA possesses the most significant fluctuation in clustering performance, with the lowest average accuracy (AA) in the Glass dataset and an AA of 70.4%. The highest accuracy is 78.5%, and the lowest is 69.5%. The K-means++ algorithm has less significant fluctuations in clustering, and its highest AA is in the Wine dataset, with an AA of 95.4%. The highest accuracy is 95.7%, and the lowest is 95.1%. The clustering performance of the improved KMA does not fluctuate, and the highest, lowest, and AA rates remain consistent across all datasets, with AA rates of 78.9%, 84.5%, and 95.9%. The data show that the improved KMA has stable clustering performance and higher clustering accuracy than traditional KMA and K-means++ algorithms. The experimental data on the number of neighbors and clusters are revealed in Fig. 10.

Figure 10a indicates the number of nearest neighbors under the optimal weight on the accuracy of model recommendation. The experiment was tested by combining the logistic regression model and gradient descent method. It can be concluded that when the number of neighbors is less than 80, the recommendation accuracy increases rapidly, while when the quantity of neighbors is over 80, the recommendation accuracy fluctuates and increases. Figure 10b shows the clustering effect of the improved KMA when the quantity of nearest neighbors is 30. It reveals that when the quantity of clusters is below 8, the clustering accuracy shows an upward trend; When the quantity of clusters is greater than 8, the clustering accuracy shows a downward trend, indicating that the accuracy is highest when the quantity of clusters is 8. Therefore, the model sets

Table 2 Cluster testing dataset partitioning

Data set	Number of samples	Feature attribute dimension	Number of categories
Glass	625	4	3
Iris	150	4	3
Wine	178	13	3

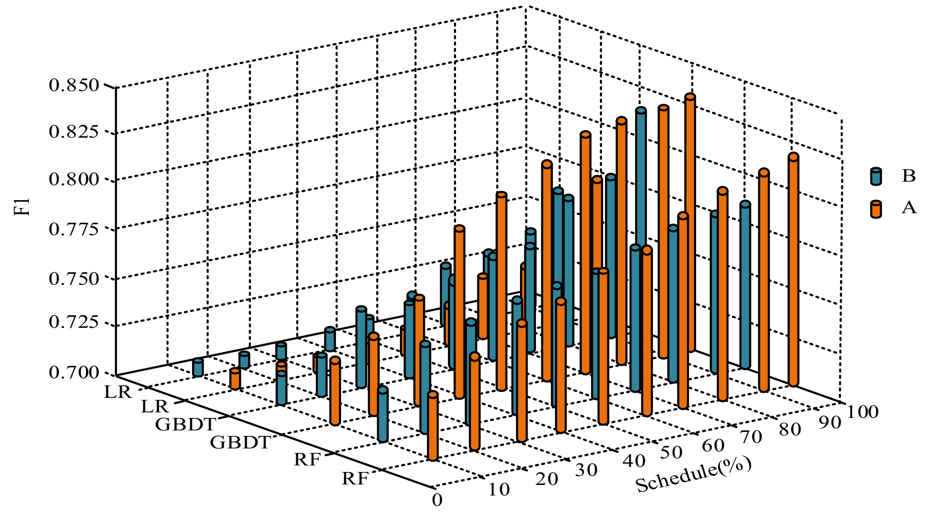


Fig. 9 Comparison of predictions under three models

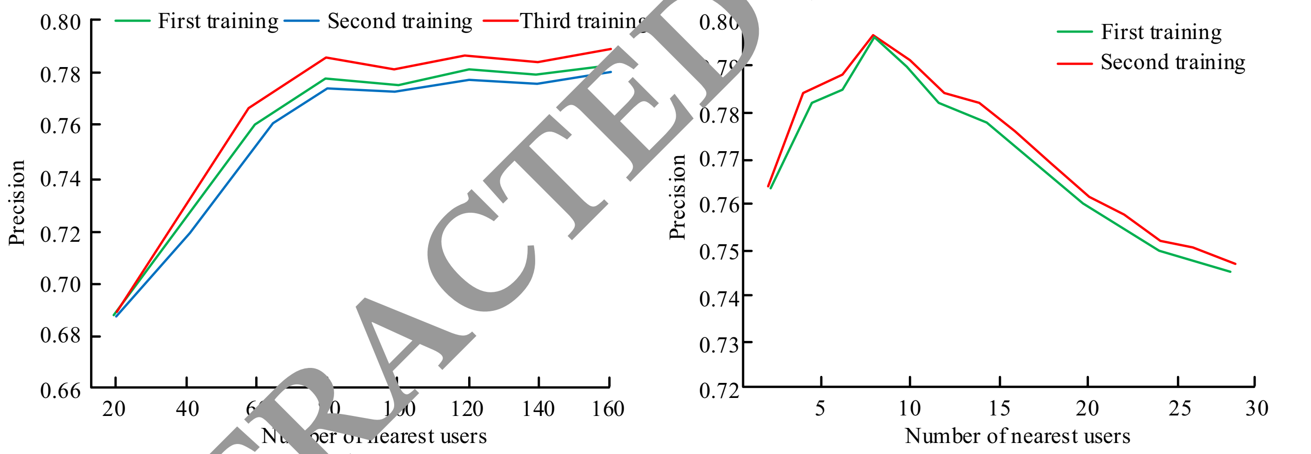
Table 3 Clustering performance of various algorithms in the dataset

Data set	Algorithm	Maximum accuracy (%)	Minimum accuracy (%)	Average accuracy (%)
Glass	K-means	78.5	69.5	70.4
	K-means++	78.7	74.5	76.1
	Improve K-means	78.9	78.9	78.9
Iris	K-means	83.6	79.1	80.8
	K-means++	83.9	80.9	82.2
	Improve K-means	84.5	84.5	84.5
Wine	K-means	95.4	94.9	95.2
	K-means++	95.7	95.1	95.4
	Improve K-means	95.9	95.9	95.9

the quantity of clusters to 8. For testing the effectiveness of the community discovery-based friend RA (labeled as CD), the K-means++ algorithm was used as a comparison in the experiment. The recommended effects of each algorithm are revealed in Fig. 11.

In Fig. 11, the recommendation accuracy of K-means++ algorithm combined with multi-feature fusion is higher than that of K-means++ algorithm combined with three types of independent features. This indicates that the analysis of social relationship strength through multi-feature fusion is more comprehensive and can enhance the accuracy of UP description. The community discovery-based friend RA presented in this study has the highest accuracy, illustrating that improving the KMA can further improve recommendation accuracy. The comparative experimental results of user comment feature algorithms are revealed in Fig. 12.

In Fig. 12, the user comment feature recommendation performance of traditional DTM and LDA models is poor. The effectiveness of the ATD-DTM model considering time division has been improved. However, the ATD-DTM model with the addition of a time decay function has been further improved, indicating that the time decay function can adapt to changes in UP. This makes the calculation results of user comment vectors more in line with the actual situation of users, further improving the accuracy of recommendations. The experiment compared the

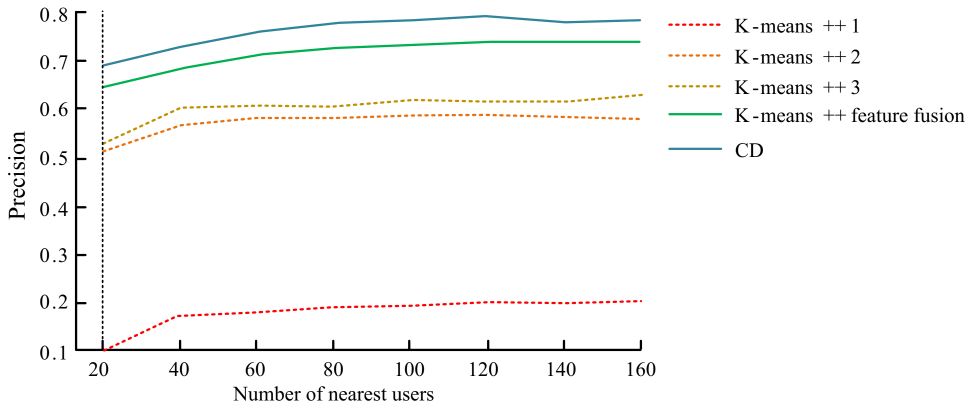


(a)Accuracy of nearest neighbor number under optimal weight

(b) Improving the clustering performance of K-means algorithm

Fig. 10 Experimental data on the number of neighbors and clusters

Fig. 11 Comparison of friend recommendation algorithms



traditional concatenated feature recommendation method (marked as A) to verify the presented FE method in view of alternative cost (marked as B). The comparative experiment of the user behavior feature algorithm is revealed in Fig. 13.

Figure 13 reveals that as the quantity of neighboring users grows, the accuracy of each RA gradually increases. Experimental data show that the proposed FE method based on substitution cost has better performance in recommending behavioral features. When the number of user behaviors is 180, the accuracy of the traditional concatenated user behavior feature recommendation is 0.58, and the accuracy of the FE method based on alternative cost is 0.63, which improves the accuracy by about 9%.

5 Conclusion

The network information is complex and rich, and the effective information extracted from it helps to enhance the services. For improving the user recommendations on social EC platforms, a EC user RA based on social relationship features and improved KMA is presented. This algorithm uses the dynamic topic model based on adaptive time slice division for improving the topic extraction, and combines the EC user RA in view of improved KMA to complete friend recommendation. The experimental data show that the ATD-DTM model has less Perplexity than the LDA and DTM models, and its curve is more stable when the number of topics is small. This indicates that the model has good explanatory power, with the highest clustering effect and topic extraction accuracy. When the

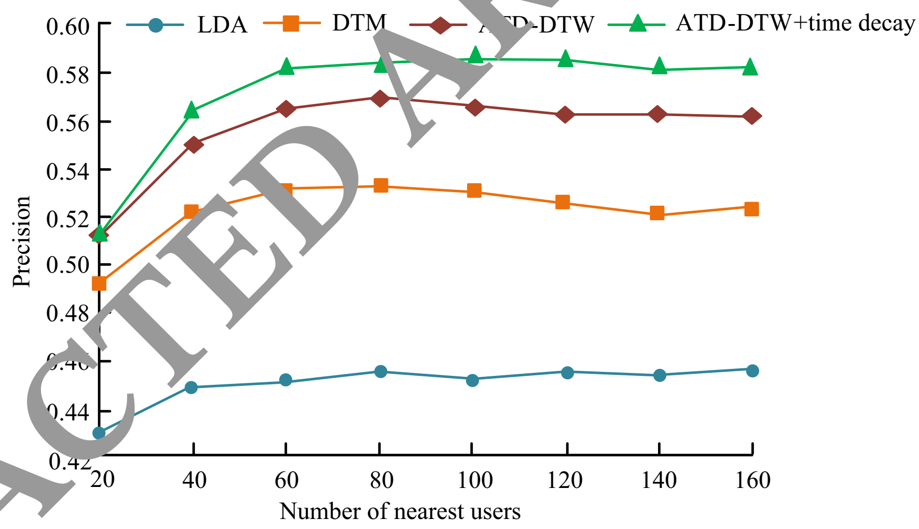


Fig. 12 Comparative experimental results of user comment feature algorithms

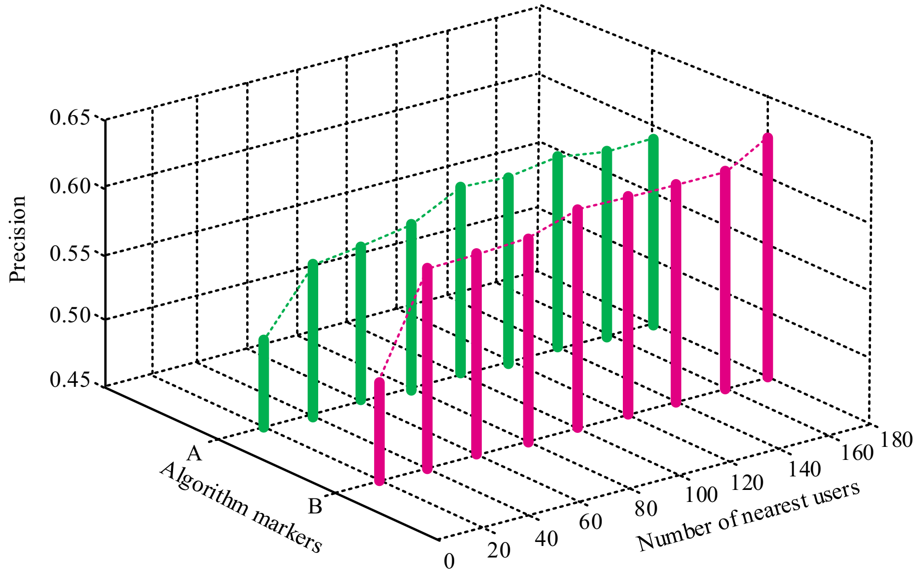


Fig. 13 Comparative experimental results of user behavior feature algorithms

number of topics is 17, the Perplexity of the model is basically stable, and the quantity of topics is set to 17. The F1 score of the FE method based on alternative cost under the three models exceeds that of the traditional concatenation FE method, illustrating that the prediction effect of this method is good and can provide high FE accuracy and prediction accuracy. The clustering performance of the improved KMA does not fluctuate, and the highest accuracy, lowest accuracy, and AA remain consistent under the three datasets, with AA of 78.9%, 84.5%, and 95.9%, respectively. The data show that the improved KMA has stable clustering performance and higher clustering accuracy than traditional KMA and K-means++ algorithms. The recommendation accuracy of K-means++ algorithm and multi-feature fusion exceeds that of K-means++ algorithm and three independent features, illustrating that the social relationship strength analysis of multi-feature fusion is more comprehensive and can improve the accuracy of UP description. The community discovery-based friend RA in the study has the highest accuracy, indicating that improving the KMA can further improve recommendation accuracy. The proposed FE method based on substitution cost has better performance in recommending behavioral features. When the number of user behaviors is 180, the accuracy of the traditional concatenated user behavior feature recommendation is 0.58, and the accuracy of the FE method based on alternative cost is 0.63, which improves the accuracy by about 9%. In future research, user sentiment analysis or part of speech analysis can be added to user comment features to further explore the information in user comments and improve the effectiveness of the algorithm; time slice partitioning can increase the aggregation degree of the main graph for measurement, further improving the accuracy of time partitioning algorithms; the validation dataset can be enriched by updating or self-correcting to improve the applicability of the algorithm.

Author Contributions XS wrote original draft.

Funding The research is supported by Philosophy and Social Science research project of Jiangsu Universities in 2022, "Research on the Development Path of Rural Vocational Education in Jiangsu Province under the strategy of Rural Revitalization" (No. 2022SJYB1985).

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest None to declare.

References

1. Tian, X., Xu, D., Du, D., Gai, L.: The spherical k-means++ algorithm via local search scheme. J. Comb. Optim. **44**(4), 2375–2394 (2022). https://doi.org/10.1007/s10878-021-00737-x
2. Gupta, M.K., Chandra, P.: Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: a review. Multimedia Tools Appl. **81**(26), 37007–37032 (2022). https://doi.org/10.1007/s11042-021-11255-7
3. Cao, E.: A personalised recommendation algorithm for e-commerce network information based on two-dimensional correlation. Int. J. Auton. Adaptive Commun. Syst. **15**(4), 345–360 (2022). https://doi.org/10.1504/IJAACS.2022.127411
4. Nguyen, T., Hsu, P.F.: More personalized, more useful? reinvestigating recommendation mechanisms in E-commerce. Int. J. Electron. Commer. **26**(1), 90–122 (2022). https://doi.org/10.1080/10864415.2021.2010006
5. Song, S., Peng, W., Zeng, Y.: Optimal add-on items recommendation service strength strategy for e-commerce platform with full-reduction-promotion. RAIRO-Oper. Res. **56**(2), 1031–1049 (2022). https://doi.org/10.1051/ro/2022037
6. Gao, C., Huang, C., Yu, D., Fu, H., Lin, T., Jin, D., Li, Y.: Item recommendation for word-of-mouth scenario in social E-commerce. IEEE Trans. Knowl. Data Eng. **34**(6), 2798–2809 (2020). https://doi.org/10.1109/TKDE.2020.3017509
7. Xie, N., Chen, D., Fan, Y., Zhu, M.: The acquisition method of the user's Kansei needs based on double matrix recommendation algorithm. J. Intell. Fuzzy Syst. **41**(2), 3809–3820 (2021). https://doi.org/10.3233/JIFS-191241
8. Ding, X., Liu, X.: User privacy protection algorithm of perceptual recommendation system based on group recommendation. Int. J. Auton. Adaptive Commun. Syst. **13**(2), 135–150 (2020). https://doi.org/10.1504/IJAACS.2020.109809
9. Chen, Y.: Research on personalized recommendation algorithm based on user preference in mobile e-commerce. Information Syst. E-business Manag. **18**(4), 837–850 (2020). https://doi.org/10.1007/s10257-022-00597-w
10. Shin, D.: How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. Comput. Hum. Behav. **109**(Aug), 106344–106353 (2020). https://doi.org/10.1016/j.chb.2020.106344
11. Xiang, D., Zhang, Z.: Cross-border e-commerce personalized recommendation based on fuzzy association specifications combined with complex preference model. Math. Probl. Eng. **2020**(Pt.38), 8871126–8871134 (2020). https://doi.org/10.1155/2020/8871126
12. Cong, H.: Personalized recommendation of film and television culture based on an intelligent classification algorithm. Pers. Ubiquit. Comput. **24**(2), 165–176 (2020). https://doi.org/10.1007/s00779-019-01271-8
13. Zhang, B., Zhang, Y., Bai, Y., Lian, J., Li, M.: Multi-dimensional recommendation scheme for social networks considering a user relationship strength perspective. Comput. Informatics. **39**(1–2), 105–140 (2020). https://doi.org/10.31577/CAI_2020_1-2_105
14. Chader, A., Haddadou, H., Hamda, L., Hidouci, W.: The strength of considering tie strength in social interest profiling. J. Web Eng. **19**(3/4), 457–502 (2020). https://doi.org/10.13052/jwe1540-9589.19345

15. Pan, Y., He, F., Yu, H., Li, H.: Learning adaptive trust strength with user roles of truster and trustee for trust-aware recommender systems. Appl. Intell. **50**(2), 314–327 (2020). https://doi.org/10.1007/s10489-019-01542-0
16. Shen, G., Jiang, Z.: Optimisation of K-means algorithm based on sample density canopy. Int. J. Ad Hoc Ubiquit. Comput. **38**(1–3), 62–69 (2021). https://doi.org/10.1504/IJAHUC.2021.119087
17. Feng, Z., Zhang, J.: Nonparametric K-means algorithm with applications in economic and functional data. Commun. Stat. Theory Methods **51**(2), 537–551 (2022). https://doi.org/10.1080/03610926.2020.1752383
18. Moodi, F., Saadatfar, H.: An improved K-means algorithm for big data. IET Softw. **16**(1), 48–59 (2022). https://doi.org/10.1049/sfw2.12032
19. Li, B., Li, J., Ou, X.: Hybrid recommendation algorithm of cross-border e-commerce items based on artificial intelligence and multiview collaborative fusion. Neural Comput. Appl. **34**(9), 6753–6762 (2022). https://doi.org/10.1007/s00521-021-06249-3
20. Bohra, S., Bartere, M.: Implementing a hybrid recommendation system to personalize customer experience in E-commerce domain. ECS Trans. **107**(1), 9211–9220 (2022). https://doi.org/10.1149/10701.9211ecst
21. Tewari, A.S., Parhi, I., Turjman, F.A., Abhishek, K., Ghalib, M.R., Shankar, A.: User-centric hybrid semi-autoencoder recommendation system. Multimedia Tools Appl. **81**(16), 23091–23104 (2022). https://doi.org/10.1007/s11042-021-11039-z
22. ShiB, P.U., Muthu, A., BSivaparthipan, C.: RETRACTED ARTICLE: deep learning-assisted heuristic data management in the E-commerce recommendation system. Arab. J. Sci. Eng. **48**(3), 4145–4145 (2021). https://doi.org/10.1007/s13369-021-06081-w
23. Tan, C., Zhao, H., Ding, H.: Statistical initialization of intrinsic K-means clustering on homogeneous manifolds. Appl. Intell. **53**(5), 4959–4978 (2022). https://doi.org/10.1007/s10489-022-03698-8
24. Zhang, R., Lu, S., Wang, X., Yu, H., Liu, Z.: A multi-model fusion soft measurement method for cement clinker f-CaO content based on K-means++ and EMD-MKRVM. Trans. Inst. Meas. Control. **45**(2), 287–301 (2022). https://doi.org/10.1177/0142331222111001
25. Jasinska-Piadlo, A., Bond, R., Biglarbeigi, P., Brisk, R., Campbell, P., Browne, F., McEneaneny, D.: Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset. Int. J. Data Sci. Anal. **15**(1), 49–66 (2022). https://doi.org/10.1007/s41060-022-00346-9
26. Chen, Z.: Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. J. Comput. Cogn. Eng. 1(3), 103–108 (2022). https://doi.org/10.47852/bonviewJCCE149145205514
27. Wang, X., Cheng, M., Eaton, J., Hsieh, C., Wu, S.: Fake node attacks on graph convolutional networks. J. Comput. Cogn. Eng. **1**(4), 165–175 (2022). https://doi.org/10.47852/bonviewJCCE2202321

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.