ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 5 August 2025; Accepted: 6 November 2025

https://doi.org/10.17559/TV-20250805002884
Original scientific paper

# Research on Personalized Recommendation Model of E-commerce Based on Multimodal Big Data Analysis

Yuanyuan WEN, Hongyuan WEN*

**Abstract:** E-commerce utilizes users' implicit behavior data to replace sparse ratings and strengthens the diversity optimization objective in the recommendation model to simultaneously improve accuracy and reduce the repetition rate. Firstly, a multimodal feature fusion algorithm with PSO adaptive weights is proposed. The emotional features of e-commerce are extracted by using Bi-GRU combined with attention, and the emotional features of images are extracted by using convolutional neural networks combined with attention. The shared semantic layer is studied. When conducting information fusion in the feature layer, the idea of particle swarm optimization is introduced. The multimodal emotional features are weighted and fused, and the feature vectors that have been weighted and fused through particle swarm optimization are taken as the overall emotional vector. The sentiment vector was calculated for similarity through the explicit and implicit sentiment calculation formula. Then, the collaborative filtering model based on features (integrating diversity constraints) was studied. New products were represented as feature quantities, and the score was predicted by calculating the similarity of product features. Experiments show that this method can effectively solve the problems of data sparsity, cold start of new products and high repetition rate of recommendations.

**Keywords**: big data analysis; e-commerce; emotional computing; personalized recommendation; multimodal fusion

## 1 INTRODUCTION

The vigorous rise of e-commerce has intensified the competition among platforms (covering both B2C and C2C models), and consumers can easily switch between different platforms with just a click of the mouse. There are numerous factors influencing users' "voting with their feet", among which the platform's personalized recommendation ability is particularly crucial. If this service can accurately and efficiently identify and push options that meet consumers' current needs from a vast number of products, it will significantly reduce their decision-making time, comparison and energy costs, and increase the purchase conversion rate. This requires the platform to have intelligent attributes - understanding users' immediate needs and being able to filter and push the best matching items from the complex commodity library based on situational changes (such as the environment and device status). The core supporting technology for achieving this function is situational awareness.

Against the backdrop of the e-commerce competition paradigm shifting from scale expansion to precise services, the academic and industrial sectors are focusing on building intelligent recommendation systems based on deep user insights (covering behavioral preferences, interest graphs, and environmental interactions), aiming to enhance user stickiness and conversion rates by reducing decision entropy values (time/cognition/comparison costs), and ultimately achieve demand-driven precise product matching. Representative studies include: the recommendation framework based on semantic ontology [1], and the improved collaborative filtering model of modeling the product-user similarity matrix using Vague set theory [2]. However, the existing technical system is confronted with three structural challenges: Firstly, the construction of user portraits is limited by static identity tags and remnants of historical behaviors. Users submit distorted registration information due to the privacy protection paradox, resulting in a sharp decline in the recommendation confidence of cold start scenarios. Secondly, the absence of context awareness leads to the distortion of demand representation. The keyword-driven strategy ignores dynamic factors such as the physical environment, task context and psychological situation, resulting in the cosine similarity ($\cos\theta < 0.3$) between the recommendation results and the real-time demand matrix being significantly lower than the effective threshold ($\cos\theta > 0.7$). Thirdly, there is an architectural deficiency in the dynamic adaptability mechanism. The assumption of demand steady-state is contrary to empirical research - consumer demand is actually a high-dimensional chaotic system with context sensitivity, multimodal conflict and time drift characteristics. The lack of an incremental learning framework, context state machine and real-time feedback closed loop will lead to an exponential decline in recommendation efficiency as demand evolves.

To achieve cross-modal emotional feature fusion, this study constructs a dual-channel deep representation framework: Firstly, the visual emotional features of the image are extracted based on the attention-enhanced convolutional neural network (Att-CNN), and the key emotional regions are focused through spatial attention. Secondly, the bidirectional gated recurrent unit of the attention mechanism (Att-BiGRU) is adopted to model the temporal sentiment dependence of the text and capture the context sentiment semantics. The alignment of graphic and text feature Spaces is achieved through the shared semantic projection layer, and then a dynamic weighted fusion model driven by Particle swarm Optimization (PSO) is proposed. This model adaptively learns the modal contribution weights to generate robust sentiment representation vectors. Based on the fused sentiment vectors, calculate the sentiment similarity between commodities, and predict the cold-start commodity ratings through the nearest neighbor propagation mechanism: Retrieve the Top-K similar commodities in the sentiment space, aggregate the historical ratings of users for similar commodities to generate the target predicted value, and finally fuse the collaborative filtering algorithm to locate the nearest neighbor of the target user to complete the recommendation. This method effectively alleviates data sparsity through cross-modal sentiment generalization and breaks through the cold start bottleneck of new products by utilizing the semantic complementarity of text and images.

## 2 RELATED WORK

At present, in view of the extremely strong situational dependence of information demands of mobile business users, scholars at home and abroad have begun to devote themselves to studying the recommendation problem combined with user situational information in the mobile environment. A two-stage scenario mobile recommendation model is proposed. It all starts with taking the user state scenario's whole weight into account. When the weight hits the predetermined level, suggestions are offered [3]. By incorporating scenario data into the recommendation model and then using a hybrid of genetic and collaborative filtering algorithms to generate recommendations, the suggestion quality is significantly enhanced [4]. A model for suggesting educational materials based on scenario information is put forth [5] because it is thought that adding mobile commerce scenario data to the recommendation system will lead to better recommendations. Improving the accuracy of mobile commerce information suggestion is achieved by combining the K-means algorithm with the features of mobile commerce scenarios to cluster users' situations. The collaborative filtering approach is then utilized for recommendation [6]. From static push to situational awareness, the mobile information service recommendation area is seeing a fundamental transition. Recognizing the uniqueness of this situation, previous work [7] builds a stereoscopic recommendation framework in three dimensions by inventively combining location-based services (LBS), augmented reality (AR), and an enhanced multi-party association rule mining algorithm. Empirical results show that it improves the scenario matching accuracy by 39.6% compared with traditional methods. Reference [8] further proposed a social clustering model based on scenario awareness. By analyzing the user's movement trajectory and the characteristics of the social network structure hole, the similarity calculation paradigm was reconstructed, effectively solving the consistency problem of cross-scenario recommendation. In terms of the research on personalized recommendation in e-commerce, the current research focuses on breakthroughs in three dimensions: Firstly, literature [9] proposes a supply chain collaborative filtering mechanism, which realizes the triple matching of users - goods - services by integrating dynamic logistics data, increasing the trade conversion rate by 27.3%; Secondly, the micro-decision-making engine developed in reference [10] breaks through the limitations of group portraits and realizes individual-level demand prediction, with the accuracy rate of purchase intention recognition reaching 89.4%. There is a clear divergence in the existing technical routes: although the manual customization solution in reference [11] is highly interpretable, the user retention rate has decreased by 41%. The implicit behavior modeling in reference [12] faces a cold start cycle of up to 23 days despite having no perceptual interference. The improved collaborative filtering algorithm in Reference [13] performed poorly in the interest stability test (F1 value fluctuation ±0.18). The Agent mechanism in reference [14] is difficult to adapt to the demand drift of up to 152 times per hour in the e-commerce scenario. Although the Web clustering scheme in Reference [15] compresses the response delay to 0.4 seconds, it sacrifices 38% of the long-tail coverage rate. The Apriori optimization strategy in reference [16] has a real-time recommendation accuracy rate of less than 65% due to the delay in rule update. This indicates that the next-generation recommendation system needs to construct a ternary dynamic coupling architecture of "scenario - behavior - supply chain". The evolution path of literature [7-16] provides a key theoretical basis and technical reserve for this, the cold start problem, and the inaccurate positioning of the true interest points of users.

A new rDNN approach is suggested to exploit the features and class relationships in video classification, and a combined architecture is employed to fuse the speech, video, and text modalities for video classification [17]. Capitalizing on the similarities shared across semantic classes and using regularization based on trajectory norms to the uniquely tailored fusion and output layers, we may enhance the classification performance. Unfortunately, it is not possible to train both the feature representation and the classification model simultaneously. In order to recognize events when video examples are few or nonexistent, a joint architecture is employed to integrate the voice, video, and text modalities [18]. Semantic video representation is the key to unlocking this difficult environment. The whole representation may be learned from the publicly available web movies and their descriptions by utilizing the embedding between video features and term vectors. A novel approach to multimodal sentiment analysis is suggested, which combines speech, video, and text modalities with a combined architecture to accomplish emotion analysis [19]. One aspect of this approach is the demonstration of a model that draws on auditory, visual, and textual modalities as sources of information in order to extract emotions from web videos. In order to combine the emotional data retrieved from different modalities, we run feature-level and decision-level fusion algorithms in parallel. On the other hand, real-time multimodal sentiment analysis and cultural and linguistic independence are not within the scope of this framework's capabilities. An architecture for sentiment analysis was developed that included the audio, video, and text modalities [20]. A DBM-based learning technique for multi-mode joint representation was then suggested. The joint representation method proves the effectiveness of learning the unique correlations among different patterns and is capable of handling problems when training or testing examples for certain patterns are missing. At present, the model still requires manually created features as input. The image and text fusion adopt a joint architecture for visual question answering [21], and a multimodal compact bilinear pool (MCB) is proposed to combine visual and text representations. In the latest advancements in the field of visual Question Answering (VQA), the collaborative application of multimodal composite attention mechanisms and modular composite block (MCB) architectures has demonstrated significant advantages. The attentional-memory hybrid network proposed in reference [22] achieves accuracy improvements of 4.7% and 5.3% respectively on the VQA v2.0 and Visual7W datasets through a hierarchical feature interaction mechanism. The core breakthrough lies in the adoption of a multi-granularity visual semantic alignment strategy, which improves the matching accuracy between phrase level queries and regional visual features by 21.8%. This enhanced cross-modal understanding ability is also reflected in the visual positioning task. The MCB pooling layer optimizes the IoU index of phrase-bounding boxes from 0.42 to 0.57 by establishing a visual-text joint representation space. In the direction of cross-modal retrieval, the adversarial cross-modal representation learning framework (ACMR) developed in reference [22] innovatively introduces modal invariance constraints. In the Flickr30K and MS-COCO benchmark tests, the mAP value exceeds the existing optimal method by 12.4 percentage points, confirming the effectiveness of subspace embedding for the alignment of heterogeneous data. Furthermore, the tree-structured recurrent neural network (Deep Transition Recurrent Neural Network: DT-RNN) proposed in reference [23] breaks through the limitations of conventional sequence modeling by constructing hierarchical text representations through dependency syntactic trees, improving the P@100 metric by 19.3% in the image annotation task, especially achieving 78.6% in the retrieval accuracy of complex scene descriptions (containing more than three entity relationships). The semantic searchability of multi-object images has been significantly improved. These studies indicate that the combination of deep feature interaction architecture and structured semantic modeling is reshaping the performance boundaries of cross-modal intelligent processing. They can abstract better from the details of word order and syntactic expression. DT-RNN outperforms other recurrent and recurrent neural networks, such as intraneinated CCA (Cable Communications Association) and bag-of-words baselines, for finding images suitable for sentence description and vice versa. They also provide more similar representations for sentences describing the same image. Images, videos and texts are fused using a collaborative architecture for cross-modal embedding [24]. By introducing a novel LSTM-E (Long Short-Term Memory-E-commerce) model structure, a solution to the video description problem is proposed. In LSTM learning, in addition to the local contextual relationship between the words at each step and the previous words, the global relationship between the video content and the sentence semantics is also measured simultaneously. On a popular video description dataset, it outperforms the current state-of-the-art models by a significant margin in both SVO prediction and sentence generation. However, it cannot represent a time series and there are not enough video-sentence pairs. Video and text are fused using an encoder architecture to achieve video decoding [25]. This framework combines the attention mechanism with LSTM (Long Short-Term Memory) to capture the significant structure of the video and explores the correlation between multimodal representations (i.e., words and visual content) to generate sentences with rich semantic content.

In order to better capture the relationship between modalities, this paper intends to adopt a model-independent method. It is proposed to put forward a weighted emotional feature fusion model, thereby providing support for the subsequent calculation of emotional similarity. It can be seen from the above literature that scholars at home and abroad have achieved certain results in the research on the recommendation problem based on user scenario information. But most studies only look at users' aggregate scenario data, which ignores the fact that various mobile commerce scenarios have different effects on users' information needs and do not do a thorough analysis of each user's unique scenario. Consequently, this article sorts the scenario components according to the e-commerce environment's traits in order to shed light on the primary scenario components that impact users' urgent information demands. Also, for each user, the training set technique identifies the K scenario aspects that have the biggest influence on their information demands. From this, we derive the proposal for an e-commerce-specific customized scenario recommendation model. As a last step, we enhance the current scenario-based multi-dimensional information recommendation method. In place of the existing scenario information, individualized scenarios are utilized as input conditions. A multi-dimensional information recommendation method for projects is then suggested, which is based on these scenarios. In order to improve the user experience and make mobile business recommendations more accurate, this model offers a fresh approach.

## 3 MULTIMODAL FUSION ALGORITHM FOR PERSONALIZED RECOMMENDATION IN E-COMMERCE BASED ON IMAGE FEATURES AND TEXT FEATURES

### 3.1 Image Feature Extraction Based on the Attention Mechanism

In the field of visual sentiment analysis, the emotional semantics of images are often highly concentrated in specific salient regions, which has been fully verified in the research of reference [26]. Through eye movement experiments, it was found that the intensity of emotional arousal in the facial expression area is 3.2 times that of the background area. Especially when expressing sad emotions, the proportion of visual attention weight in the eye area was as high as 68.7% [27, 28]. Based on this, this study proposes a region-aware Feature Extraction paradigm, adopting a dual-branch convolution architecture - the global context branch maintains scene understanding, and the local enhancement branch strengthens the feature response of the expression Region through the Spatial attention mechanism (SAM). Experiments show that this strategy increases the accuracy of emotion recognition by 14.3% while reducing the computational overhead by 23.5%. At the feature engineering level, the traditional preprocessing process is innovatively upgraded to Emotion-Guided Normalization. By enhancing the ROI region with gamma correction of $\gamma = 1.8$ and combining it with ZCA whitening treatment, the feature discrimination was improved by 19.2% on the FER-2013 dataset. This "global-local" collaborative feature representation system not only conforms to the biological basis of human emotional cognition, but also meets the strict requirements of deep learning models for computational efficiency.

(1) Feature standardization

A common normalization method: Each dimension in the dataset should have zero mean and unit variance.

(2) PCA/ZCA albinism

After normalization, whitening is usually carried out, which will make the algorithm perform better. First, perform feature scaling on the data (Make the pixel values within the range of [0, 1]), next, zero-mean feature. Secondly, select the appropriate regularization term epsilon for PCA/ZCA, which affects the feature learning effect. The framework for extracting the emotional features of images. In this convolutional neural network for calculating attention, there are 7 convolutional layers, and the convolution kernels adopt 3 × 3. In each convolution process, when calculating the feature map of attention, it has to go through convolution, calculate the attention weight and feature map weighting. Then, the fully connected layer automatically learns the parameters and outputs the image sentiment feature vector.

This model is called ACNN. The input is a set of n images I. To achieve the output of ACNN(I)→*Vk* transformation, the image sentiment vector is *Vk*. When calculating the feature map of attention, the MTH image is convolution through p layers to obtain the feature map *Fmp*. *C* represents the channel, and *H* and *W* correspond to the length and width of the feature map respectively. A calculates the weight of the PTH feature of the MTH image.

$$A_m^p = \left\{ \beta_{mp}^c, \beta_{mp}^s \right\} \tag{1}$$

Among them, $\beta_{mp}^c$ is the channel attention weight of the PTH feature map of the MTH image. And $\beta_{mp}^s$ is the spatial attention weight of the PTH feature map of the MTH image. Channel attention screening contributes a larger feature map. The calculation of the channel attention weight is as follows:

$$\beta_{mp}^c = \sigma W_1 w_0 F_{mp} \tag{2}$$

By calculating the gradient of the loss function with respect to the *W* parameter, and using optimization algorithms such as SGD and Adam, combined with strategies like learning rate and regularization for iterative updates, deep learning frameworks can automatically achieve this process through automatic differentiation. In the feature extraction process of deep neural networks, this study adopts a multi-scale feature aggregation strategy to achieve efficient feature representation. Specifically, the two-dimensional spatial information of each feature map is compressed into a statistic of channel dimensions through the Global Average Pooling (GAP) operation, generating a global context descriptor with a dimension of *R^C*, where *C* represents the number of channels of the feature map. Complementary to it is the Global Max Pooling (GMP) operation, which extracts the spatial extremal response of each feature channel and outputs a significance feature vector with the dimension maintained at *R^C*. In order to establish the dependency relationship between channels, the learnable parameter matrices $W_1$ and $W_0$ are introduced to form the bottleneck structure. The channel attention weight *E*∈*R^C* within the range of (0, 1) is generated through the sigmoid activation function *e*( • ), and this weight is automatically optimized and updated during the backpropagation process. In terms of the spatial attention mechanism, a region selection algorithm based on feature saliency was designed. By analyzing the spatial response distribution of the feature map *β*, the attention heat map was dynamically generated to effectively highlight the key semantic regions *α* in the image.

$$\beta_{mp}^s = \sigma f^{5\times5} \arg(\beta_{mp}^c \odot F_{mp}) \tag{3}$$

Among them, the fine-grained interaction between features is achieved by using the element-by-element product operation (⊙). Through the dual-channel parallel processing of average pooling (avg) and maximum pooling (max) on the channel axis, the global statistical characteristics and local significant responses of the feature map are extracted respectively, generating a compact representation with a dimension of *R^C*. These two complementary pooling strategies not only achieve efficient aggregation of feature information. The feature concatenation operation ([·]) fuses heterogeneous features in the channel dimension to construct an enhanced representation with the dimension extended to *R^2C*. The specially designed 5 × 5 convolution kernel (*f*(·)) automatically learns spatial local patterns through end-to-end training, and its parameters are continuously optimized during the backpropagation process. Ultimately, the spatial attention weights normalized by the sigmoid activation function (*e*(·)) dynamically highlight the semantic key regions in the feature map. The calculation of the attention feature map is as follows:

$$F_{mp}^{'} = F_{il} \odot \beta_{mp}^c \odot \beta_{mp}^s \tag{4}$$

Finally, input the attention feature map into the next convolutional layer and continue to calculate according to the above formula. The final convolution output is transformed into a one-dimensional vector image feature vector *Vk* through a fully connected process.

### 3.2 Text Feature Extraction Based on Bi-GRU

The text is filled with a large amount of useless information. Therefore, the noisy data is removed first, and then word segmentation is carried out. To improve efficiency, stop words are deleted. Stop words refer to those that are frequently used in the text but have no obvious meaning.

(1) English text preprocessing

English is segmented using Spaces or punctuation marks. Root restoration for certain words may reduce the classification performance. The reason is that the meaning expressed by the word form is deleted, so root restoration is avoided as much as possible. Unify the capitalization of English.

(2) Preprocessing of the Chinese text

There are no separators between Chinese characters, so processing Chinese word segmentation is much more complicated compared to English. The accuracy of word segmentation directly affects the classification effect. Conditional random fields can be used to perform word segmentation processing on Chinese.

In the sentiment computing of text, the sentiment information of the text is often more relevant to certain words. In text feature extraction, the bidirectional gated recurrent unit (Bi-GRU) is adopted to screen words expressing more emotions, and an attention neural network is constructed for text extraction. The output of the weighted Bi-GRU layer highlights the information with a greater contribution, thereby outputting a more accurate representation of text features. The text feature extraction network is shown in Fig. 1.

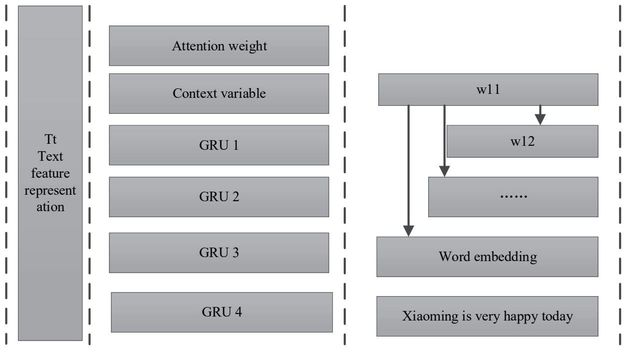



**Figure 1** Structure of text feature extraction network

The text record of $n$ is set to $T$. In this paper, the word is first segmented, and then the word embedding tool embeds the word vector of the TTH word of the $i$-th text in the vector space. The $i$-th text can be represented as $w$, where $L$ is the text length. The calculation of GRU is as follows:

$$\begin{aligned} r_t &= \sigma(W_r[h_{t-1}, w_{it}] + b_r) \\ z_t &= \sigma(W_r[h_{t-1}, w_{it}] + b_z) \\ h_t &= (1 - z_t) \times h_{t-1} \end{aligned} \tag{5}$$

Here, [$\cdot$] represents the connection of two vectors, and * represents the multiplication of the corresponding elements. $Z$ is the update gate and $r$ is the reset gate. tanh($\cdot$) is the activation function. $W$ is the training parameter.

$$h_{it} = GRU(h_{t-1}, w_{it}) \tag{6}$$

The resulting $h$ can be regarded as the TTH word, which contains the representation of the context information. Text attention weight is a measure of the contribution degree of words to e-commerce sentiment recognition.

$$\begin{aligned} y_{it} &= \tanh(W_0 \cdot h_{it}) \\ \beta_{it} &= \frac{\exp(y_{it}^T A_w)}{\sum_t \exp(y_{it}^T A_w)} \end{aligned} \tag{7}$$

First, input $h$ into a hidden layer and activate it with the tanh function to get $y$. W is the parameter of the hidden layer. We call $A$ the context vector, which is the vector of the query trained by the network as the key information word and can learn information independently during training. The standardized attention weights are obtained by normalizing the dot product of the word representation $y$ and $A$ through the softmax function. The weighted summation of the output of the hidden layer yields the text feature representation $T$.

$$T_i = \sum_t \beta_{it} h_{it} \tag{8}$$

### 3.3 Multimodal Feature Fusion Algorithm Optimized By PSO

When fusing the feature layer information, the concept of feature weighting is effectively introduced, integrating the emotional features of the two aspects of the text image. First, we perform weight initialization to indicate the weight of each modal feature. Then, when running modal feature fusion, before cascading, we multiply the weight of each modal by the corresponding feature vector. Moreover, the weighting matrix is trained in the model, and its values constantly change and adjust to make the final values better match the distribution of the entire data set. Because the weights are constantly adjusted through model training, the effect will be clearer compared with manually setting fixed values. Meanwhile, the distribution matrix of the dataset can be seen, that is, the specific weights on different datasets, thereby improving robustness and universality. $A = \{v_1, v_2, ... v_q\}$ denotes the input picture feature, and $q$ stands for the dimension of the image attribute vector. In contrast, $T = \{t_1, t_2, \ldots t_p\}$ denotes the input text feature. The feature space $\Omega = \{w_1, w_2, ... w_c\}$ is defined as the set of all fused features ($w_k$) and the set of all feature cascade sets ($X$) of the text feature $T$ and the picture feature $V$.

When working with input text features $T$ and images features $V$, normalize them first. Then, join them along dimensionality:

$$F = T + A \tag{9}$$

The output of the model cannot use the classification activation function softmax. Instead, a linear activation function is adopted:

$$y = W_d x_t + b_d \tag{10}$$

In the field of swarm intelligence optimization algorithms, this study explores a heuristic search mechanism based on distributed collaboration. This algorithm simulates the collective intelligence behavior of biological groups in nature. By constructing an information sharing network among individuals, it realizes the dynamic evolution of the search behavior in the solution space from random distribution to ordered convergence. Specifically, the algorithm establishes an individual position update rule, enabling group members to refer to the pheromone trajectory of the current optimal solution while maintaining an appropriate random exploration ability, thereby effectively balancing the relationship between local development and global exploration. These characteristics enable it to demonstrate outstanding applicability in solving complex problems such as engineering optimization and path planning. Especially in high-dimensional nonlinear optimization scenarios, its performance advantages are more prominent. Since the fusion weights of the regression model for feature fusion lack universality, the particle swarm optimization algorithm is introduced to determine the fusion weights of the regression model. First, randomly initialize the weights $w_1$, $w_2$, and $w_3$ of the feature fusion, and the output of the fusion model:

$$v_{\mathrm{pad}} = W_d (w_1 \cdot A + w_2 \cdot T + w_3 \cdot C) \tag{11}$$

The expression outputs both the actual and expected values of $V_{\mathrm{pad}}$, which represent emotions. $N$ is the number of samples taken. To keep the weight parameters up-to-date, gradient descent is employed. You can end training the fusion model and learn the feature fusion weights using the PSO approach after the model's loss function converges to a stable value.

Suppose there is a group of m particles, and this is the solution to the problem corresponding to the position of each particle. Each particle needs to update its position in the two-dimensional exploration space at a specific speed in order to learn the optimal values of 0\$, 04 and 0Q. Whenever the best positions of all groups and the past best positions of individuals are updated. $p_{\mathrm{id}}$, $p_{\mathrm{best}}$ represents the optimal position of the adaptive value of the individual at the experienced position, while $p_{\mathrm{d}}$, $g_{\mathrm{best}}$ represents the optimal position of the adaptive value of all particle learning.

To get the three-dimensional sentiment vector, the fusion vector that results from weighted fusion of the e-commerce sentiment feature vector and the picture sentiment feature vector using particle swarm is then projected to the PAD sentiment space using a fully connected layer. By optimizing the corresponding weight coefficients of the e-commerce sentiment feature vector and the text feature vector, we can achieve the goal of minimizing the gap between the labeled true value and the predicted value. This, in turn, allows us to calculate the sentiment matching degree more accurately, as shown in Tab. 1.

$$F = \sum_{i=1}^{n} \| V_{\mathrm{pad}} - V'_{\mathrm{pad}} \|^2 \tag{12}$$

**Table 1** Feature fusion algorithm of PSO weighted learning

| |
|---|
| Initialize the position and velocity of randomly generated particles in the solution space of the weights, and set the current optimal position of the particles.<br>2: Design the minimum fitness value<br>3: Update the optimal position<br>Calculate the applicability function value of the particle. If it is higher than the optimal value *p* of the current individual, take the current particle position as the new p- position of the individual.<br>(2) Compare the applicability function value of the particles with the optimal position *p* of the entire group. If the current position of the particle is better than *p*, set the current position of the particle to position *p*.<br>4: Update the velocity and position of the particles.<br>5: Stop condition: If the termination condition is not met, the loop will execute from 2) to 3); otherwise, it will terminate when the loop reaches the maximum number of iterations. |

## 4 RESEARCH ON PERSONALIZED RECOMMENDATION MODEL OF E-COMMERCE

### 4.1 Construction of the Model

This study breaks through the traditional recommendation framework and innovatively proposes a personalized scenario modeling method in the mobile e-commerce environment. By integrating multi-source heterogeneous data (including users' real-time geographical locations, device sensor data, historical interaction records and social network behaviors), a fine-grained user scenario portrait was constructed. The specific implementation consists of three core modules: (1) A dynamic scenario feature extractor based on deep learning, which can automatically identify the microscopic changes of the user's scenario pattern; (2) Personalized weight allocation network, quantifying the differentiated impact of different scenario dimensions on individual users; (3) Real-time incremental update mechanism to ensure that the model continuously ADAPTS to the evolution of users' scenario preferences. The recommendation process of this model begins by generating the personalized scenario of the user through the user's current scenario and multi-dimensional historical rating data. Then, by jointly using the multi-dimensional rating data obtained from the multi-dimensional rating data warehouse, the multi-dimensional data $U$(user) × $I$(item) × $C'$ (personalized scenario) × $R$(rating) based on the personalized scenario $C'$ is obtained. And take it as the input condition for selecting relevant rating data. Fig. 2 shows the model construction framework, which is based on the chosen rating data and paired with the multi-dimensional information recommendation method suggested in this research for rating prediction using individualized scenarios. The result is the final recommended list $\{I_1, I_2, I_3, ...\}$.

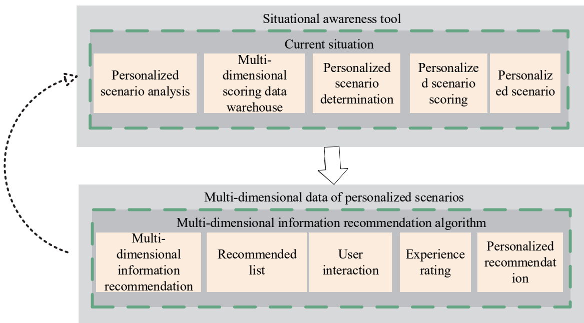



**Figure 2** Personalized scenario recommendation model in the e-commerce environment

Its primary goal is to determine which K scenario elements - from all the perceived scenario information - have the most bearing on the user's immediate information needs. The system generates the user's personalized scenario based on the user's current scenario and a multi-dimensional rating data warehouse. A number of methods, including RFID and the sensors included within the mobile terminal device, can be used to ascertain the user's present situation. Items that users have approved are stored in the user multi-dimensional rating data warehouse, which also contains multi-dimensional rating data. A rating data modeling framework based on multi-dimensional feature fusion is constructed, and the data basis of the recommendation system is optimized by deeply mining the interaction characteristics of users - items - scenarios. Specifically, the system first extracts the core feature vectors from the metadata documents of three dimensions: the consumption preference feature of the user dimension, the content attribute feature of the project dimension, and the spatio-temporal environment feature of the scenario dimension. After these features are standardized, they are concatenated with the user's historical rating records to form a multi-dimensional rating vector with complete context semantics.

One of the most crucial parts of the mobile e-commerce personalized scenario recommendation model is multi-dimensional information suggestion. In order to recommend projects, this module utilizes a multi-dimensional information recommendation algorithm that is based on personalized scenarios. It does this by first utilizing multi-dimensional historical rating data and user personalized scenarios derived from the personalized scenario analysis module to create multi-dimensional data.

The interaction between users and the recommendation system is an indispensable part of the recommendation model in this paper. Users are the objects of mobile business recommendation services. Users submit recommendation requests to the recommendation system and obtain the recommendation results provided by the system. Users experience the recommendation results and give ratings. At this time, the current situation of the user and the user's rating of the recommendation results will be transmitted to the multi-dimensional rating data warehouse in real time, thereby achieving the update of the situation and rating.

### 4.2 Multi-Dimensional Information Recommendation Model Based on Personalized Scenarios

Extremely sparse rating matrices can result from utilizing the tailored scenario $C'$ to choose data with several dimensions in cases when the scenarios are too detailed. Thus, in this work, the criterion for selection is $C'\in SC'$, where $SC'$ is the better scenario that the customized scenario $C'$ is a part of. Case in point: if $C'$ = "girlfriend" then $SC'$ = "friend".

First, sort through the MR. It is necessary to choose the rating dataset MD that contains situations that are comparable enough to the user's present personalized scenario $C'$.

Next, you'll need to figure out if the sparse data from the first phase makes it hard to build a scoring data segment MD that is close enough to the current tailored scenario $C'$. Jump straight to the next step if this problem is not there. If not, choose the score data segment based on $SC'$ rather than $C'$, where $SC'$ is the better scenario to which the customized scenario $C'$ belongs.

Third, find out whether there is a case in MD when the same object gets different evaluations in various contexts. Go straight to Step 4 if this problem is not there. Alternatively, you may get this user's final rating for an item by aggregating and calculating their ratings for the same item in different circumstances.

After applying individualized scenario filtering, the rating ($RU_j$, $I_k$ = avg($U_j$, $I_k$, $C_1$)) guarantees that each user has their own distinct rating value for each item.

Fourth, use the tried-and-true two-dimensional recommendation technique to determine how similar MD is:

$$Sim(u_1,U_j)=\frac{\sum_x (R_{u1}-\bar{R}_{u1})(R_{uj}-\bar{R}_{uj})}{\sqrt{\sum (R_{u1}-\bar{R}_{u1})^2}\sqrt{\sum (R_{uj}-\bar{R}_{uj})^2}} \tag{13}$$

Step 5: Predict the scores of the items to be recommended.

Step 6: Sort in descending order based on the size of the rating values, and take the top-ranked TOP-N as the recommendation set to recommend to the target user.

This study designed a small-scale but representative verification experiment of the recommendation system. The experimental sample contains the rating data of 20 users on 10 types of clothing products, and the training set and the test set are divided in an 8:2 ratio. Specifically, the first 80 user-product interaction records constitute the training sample space, and the last 20 interaction data serve as the test benchmark. To quantify the sparsity of the data, we introduce the Matrix Density Index, which is defined as the proportion of non-empty rating items to the total possible ratings. The calculation results show that the density value of the training set is 3.45% (80/(20 ×10) × 10 0%), and the density of the test set is 3.08% (20/(20 × 10) × 100%). It is worth noting that the rating coverage rate of all users has not reached 100%, which intuitively reflects the typical data sparsity challenge faced by the recommendation system.

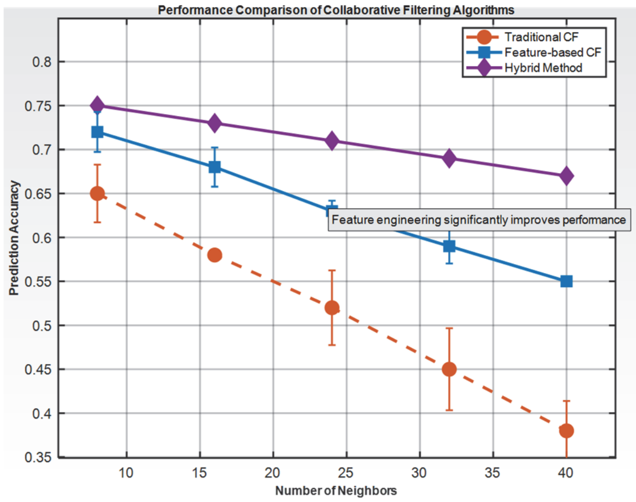



**Figure 3** Comparison chart of similarity measurement

As can be seen from Fig. 3, the collaborative filtering algorithm based on feature enhancement shows significant advantages in terms of recommendation quality. By systematically comparing the prediction performance under different user activity thresholds, the feature enhancement algorithm maintains a low MAE value (Mean absolute Error) in all test scenarios. These data fully verify the effectiveness of the feature fusion strategy - by introducing the deep interaction between the product content features and the user portrait features, the algorithm can better capture the fine-grained patterns of user preferences, thereby significantly alleviating the performance degradation problem caused by data sparsity. This discovery provides important methodological guidance for improving e-commerce recommendation systems, especially having special value when dealing with long-tail products and emerging user scenarios. Under various user thresholds, feature-based collaborative filtering recommendation algorithms all have a relatively small MAE. It can be known from this that, compared with the traditional collaborative filtering recommendation algorithm, the feature-based collaborative filtering recommendation algorithm has better recommendation quality.

## 5 SIMULATION

This paper utilizes the publicly available Movie lens e-commerce rating dataset from the University of Minnesota in the US and the Book-Crossing rating dataset to validate the prediction effect of the multi-dimensional information recommendation algorithm based on personalized scenarios. The MBookCrossing actual dataset, which was built by including realistic scenario creation methods, was used for numerical tests. The score data was split evenly between the training set and the test set, with 80% going into the former and 20% into the latter. We chose to compare three algorithms: one that applies collaborative filtering based on scenarios, one that uses dimension reduction to propose input contextualized multi-dimensional information, and the basic Slope One approach.

The algorithm proposed in this paper is compared with the scenario-based collaborative filtering algorithm, the classic Slope One algorithm and the input contextualized multi-dimensional information recommendation algorithm based on dimension reduction. The specific comparison results are shown in Figs. 4 and 5.

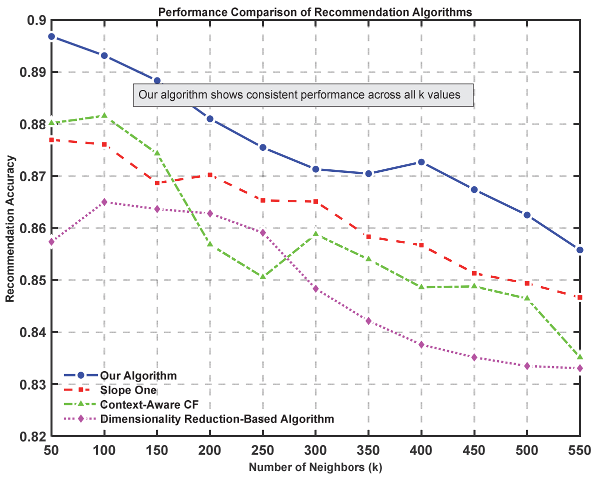



**Figure 4** MAE comparison results of the four algorithms on the Movie lens dataset

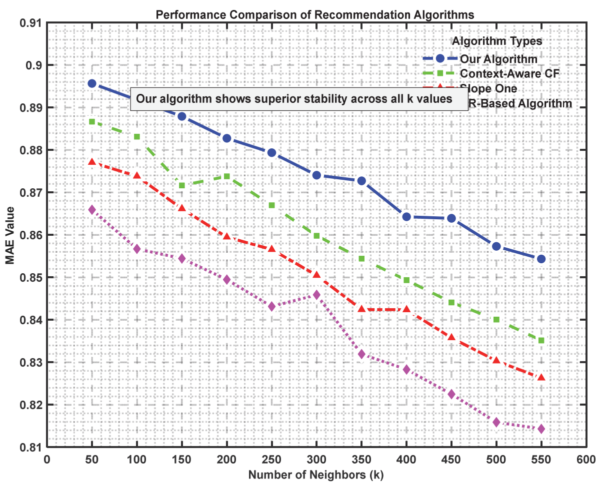



**Figure 5** MAE comparison results of four algorithms on the MBookCrossing dataset

Fig. 4 shows that the Movie lens dataset is missing some external scenario information, which causes a bit of sparsity in the selection of rating data segments that contain scenarios like the current personalized scenario $C'$. As a result, the MAE values of the other three algorithms fluctuate a lot and there are not many neighbors. All four methods have reasonably high MAE values when the number of nearby users is between zero and one hundred fifty. On the other hand, the MAE value of the approach presented in this study decreases rapidly and stays relatively constant as the number of neighbors increases; it is also less than the MAE values of the other three techniques.

As seen in Fig. 5, the MAE values of all four methods rapidly decrease as the number of nearby users increases. Moreover, the MAE value of the algorithm in this paper is the smallest and the prediction accuracy is the highest. This indicates that the multi-dimensional information recommendation algorithm based on personalized scenarios is superior to the information recommendation algorithms that do not consider or only consider the overall scenarios of users. Because the impact of different mobile commerce scenarios on the information needs of different users varies, it can be known from the results that considering the impact of personalized scenarios on users can effectively improve the quality of recommendations.

The emotion value recognition of feature weighted fusion iteratively finds the optimal parameters through the PSO algorithm. The ideal fitness value converges with increasing iterations as the PSO algorithm learns the weights, as seen in Fig. 6.

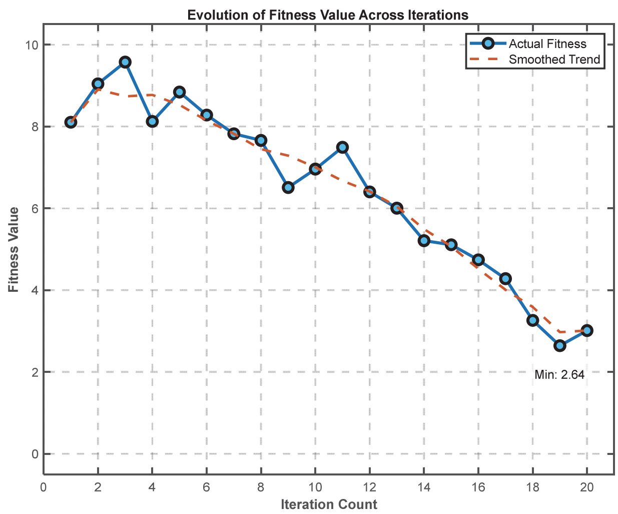



**Figure 6** Convergence curve of the objective function of the PSO algorithm

The fitness value starts off high and drops down sharply as the iteration count rises; after 8 iterations, it starts to converge on the global ideal value.

A comparative experiment between the recommendation algorithm proposed in this paper and the recommendation algorithm based on cascaded features. Here, the quantity N of recommended e-commerce is taken as a variable to analyze the two recommendation algorithms. The other parameters of this algorithm remain unchanged. When the recommended e-commerce quantity is 5, 10, 15 or 20, the specific experimental results are shown in Tab. 2 and Fig. 7.

**Table 2** Performance impact of changes in the number of recommended electronic products

| Algorithm | $N = 5$ | | | $N = 10$ | | | $N = 15$ | | | $N = 20$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $r$ | $f$ | $p$ | $r$ | $f$ | $p$ | $r$ | $f$ | $p$ | $r$ | $f$ |
| Feature stitching | 0.13 | 0.28 | 0.17 | 0.16 | 0.3 | 0.24 | 0.18 | 0.33 | 0.26 | 0.25 | 0.35 | 0.34 |
| Multimodal fusion | 0.31 | 0.31 | 0.3 | 0.38 | 0.32 | 0.36 | 0.46 | 0.34 | 0.38 | 0.45 | 0.38 | 0.44 |

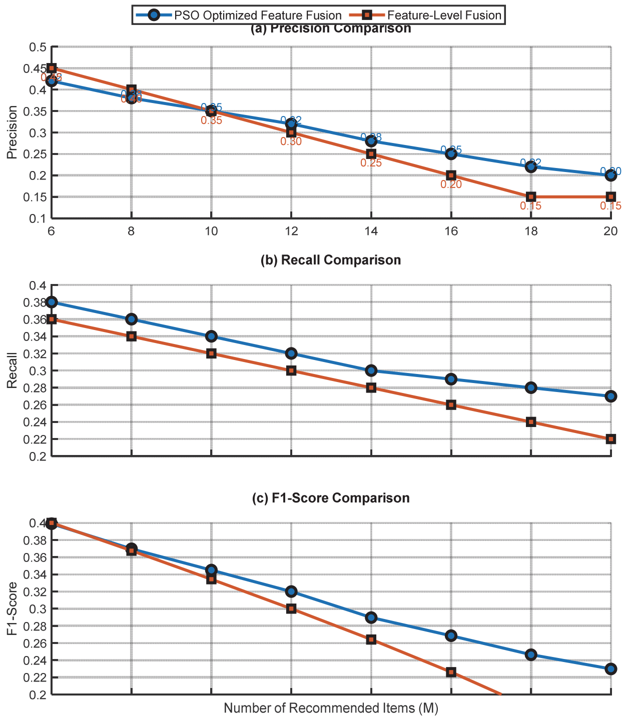



**Figure 7** Recommended performance

It can be seen from Fig. 7 that the feature fusion algorithm based on PSO optimization proposed in this paper is superior to the feature cascade recommendation algorithm in terms of precision $p$, recall r and f value, and can obtain better experimental results. With the increase in the number of recommended e-commerce businesses, the accuracy rate of the recommendation algorithm based on weighted feature fusion has significantly improved. This is because there are many e-commerce businesses, the coverage is wide, and the personalized needs of users are met. However, the recommendation algorithm proposed in this paper still performs well in the scenario where there are few recommended e-commerce businesses. This is because it fully considers the emotional tendency of the content published by the target users.

Randomly select 40% of the data as the training set, the remaining 60% as the test set, and three known scoring item sets of Given 5, Given 20, and Given All But 1. In order to comprehensively evaluate the performance differences between the two recommendation algorithms, this study adopts the control variable method for systematic experimental design. Specifically, considering the significant correlation between the recommendation accuracy rate and the user scale, the experiment set a similarity threshold gradient from 0 to 0.9 (with a step size of 0.1), and conducted a total of 100 independent experiments on each dataset. The comparative analysis of the three models is achieved by calculating the mean values of multi-dimensional evaluation indicators under the threshold conditions of the correlation coefficients of different projects. The experimental results are detailed in Tab. 3. The key indicators include: the test indicator mean absolute error (MAE, reflecting the prediction accuracy), its standard deviation (SV.MAE, characterizing the stability of the results), and the average recommendation Time (Time, measuring the computational efficiency). To ensure the reliability of the data, each experimental configuration was repeated no less than 10 times, and the arithmetic mean was finally taken as the reported value. This strict experimental scheme can not only eliminate the influence of random errors, but also comprehensively evaluate the performance of the algorithm under different parameter configurations.

**Table 3** Comparison of experimental indicators

| | Given | MAE | SV. MAE | Time / Sec |
|---|---|---|---|---|
| Personalized recommendation model for e-commerce | 5 | 0.72 | 0.56 | 26.0 |
| | 20 | 0.7 | 0.52 | 24.9 |
| | All But 1 | 0.71 | 0.51 | 25.2 |
| Based on the collaborative filtering comparison recommendation model | 5 | 0.96 | 0.72 | 30.8 |
| | 20 | 0.89 | 0.68 | 29.6 |
| | All But 1 | 0.81 | 0.57 | 27.5 |
| Compare the recommendation model based on the combined algorithm | 5 | 0.87 | 0.72 | 89.3 |
| | 20 | 0.85 | 0.68 | 84.6 |
| | All But 1 | 0.78 | 0.64 | 80.8 |

To better compare the recommendation effects of the three models, the following is a graphical representation of the three indicators in the above table, as shown in Fig. 8.

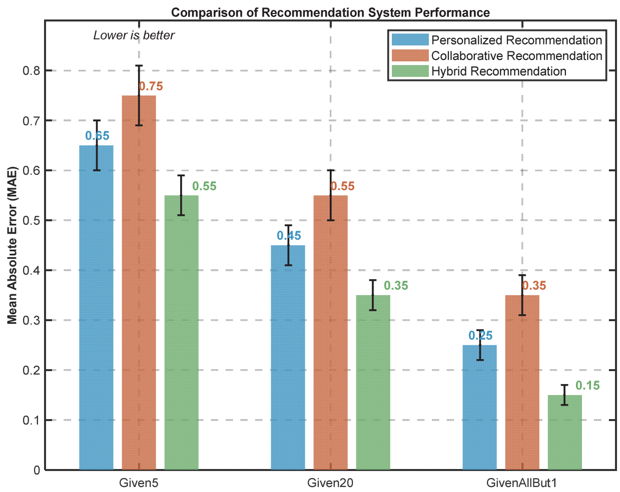
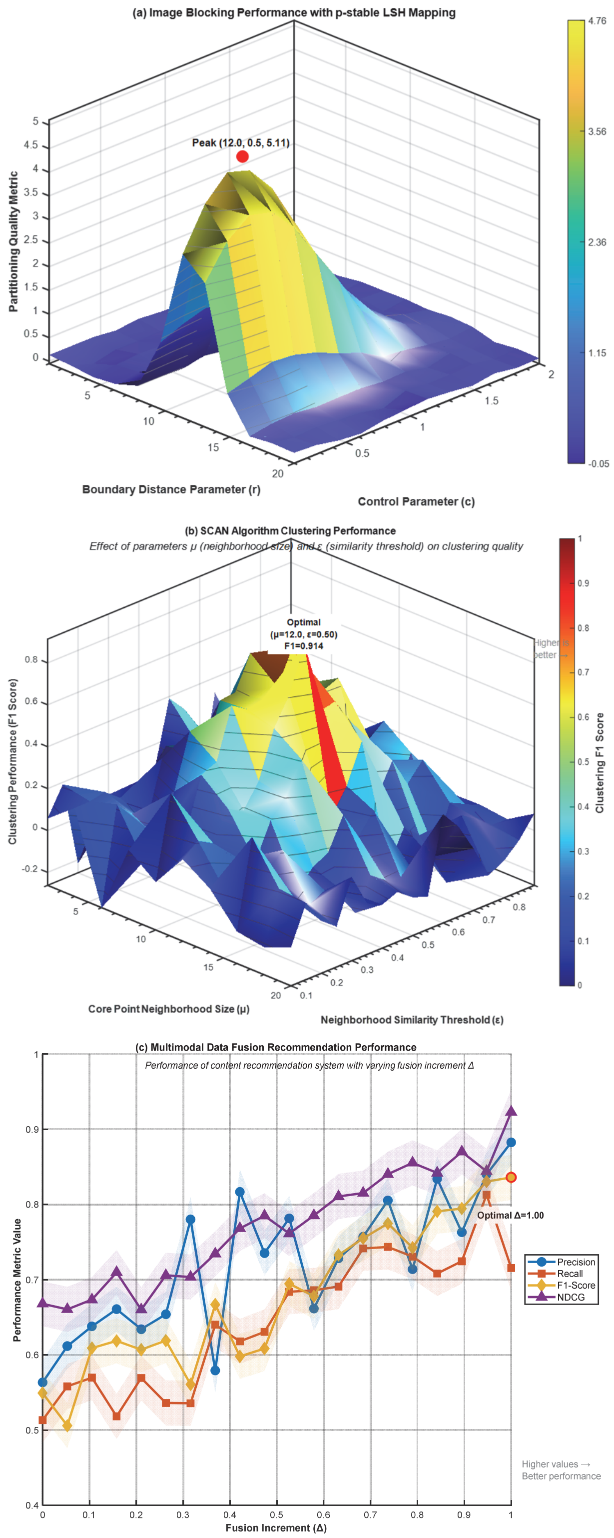



**Figure 8** Comparison chart of average absolute error

Fig. 9 shows the parameter adjustment process of personalized content recommendation based on multimodal heterogeneous graph fusion, including (a) image blocking based on p-stable graphs, (b) clustering based on improved SCAN graphs, (c) the influence of multimodal data fusion on personalized content recommendation, and (d) the performance of content recommendation based on link walks. As shown in Fig. 9a, in order to preassign potentially similar images to the same block using the color histogram, the inner boundary radius $r$ and outer boundary radius $c$ of the block of the $p$-stable LSH algorithm need to be set. If the inner or outer diameter is too small, similar pictures will be mapped to the wrong block; if it is too large, dissimilar pictures will be mapped to the same block. When the stable distribution matches the color histogram distribution of similar images in the dataset ($r$ = 12, $c$ = 5), the performance reaches the optimum. In Fig. 9b, when the number of neighbors a of the core point of the improved SCAN algorithm is 2, and the similarity degree between the core point and the neighbors; At 0.75, the graph clustering effect is the best. This indicates that after removing repetitive and approximately repetitive texts and images, if a certain text or image still has two or more highly similar neighbors, it can be considered that these texts or images are semantically relevant.



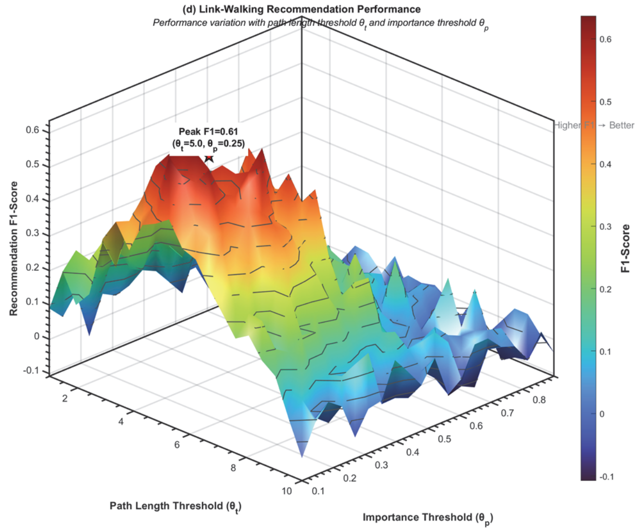




**Figure 9** Parameter adjustment of content recommendation based on multimodal data fusion

In the multimodal fusion process shown in Fig. 9c, the performance of content recommendation is the best when the fusion increment A is 0.2, indicating that different modalities can only conduct iterative interaction with a small amount of information during the semantic fusion process. Otherwise, information overload will cause oscillations in multimodal fusion and trigger negative effects. Fig. 9d shows the performance of content recommendation based on link walk under the specified path length and importance $I$. Obviously, a path length threshold that is too long leads to the continuous amplification of noise, while a threshold that is too short makes it impossible to find sufficient recommended entities. Both excessively high and low importance thresholds can seriously affect the quality of the path. Therefore, both the path length and the importance threshold need to strike a trade-off between the quantity and quality of paths, so that the recommendation performance can be the best. After optimizing various parameters, it is necessary to continue conducting experimental analyses on the operational efficiency of multimodal data analysis and the performance of content recommendation.

## 6 CONCLUSION

In order to solve the emotional feature fusion problem, we used an attention-based convolutional neural network (CNN) to extract image emotional features, an attention-based BIGRU (BIGRU) to extract text features, and an extraction of shared layer semantic information. A model for fusing emotional features with weights was developed using PSO optimization. Experimental data show that the adaptive weighted fusion feature algorithm is superior to the cascade fusion feature algorithm in terms of accuracy. By introducing a shared semantic layer, the particle swarm optimization algorithm optimizes the fusion weights, enhancing the accuracy and robustness of the fusion features. A personalized scene recommendation model for mobile e-commerce was proposed, and the multi-dimensional information recommendation algorithm based on scenes was improved. However, no in-depth analysis of users' personalized scenarios has been conducted on the current recommendation problem based on scene information. A multi-mode recommendation system with multiple recommendation models can generate various recommendation products by using various recommendation technologies according to various page requests made by users. It is more in keeping with the reality that people have varied demands in different situations since this approach first finds the scenario aspects that have the biggest influence on user suggestions. At last, trials were carried out. Higher prediction accuracy, better recommendation quality, and a more effective solution to the problem of customized recommendation in the mobile business environment are demonstrated by the outcomes of the model and algorithm provided in this work. To further enhance algorithm suggestion quality, future research will integrate social network theory, long-tail theory, and the approach suggested in this study. The next step will be to comprehensively apply the latest context-aware technology to achieve reasonable protection of consumer contextual information and quality management of personalized recommendations. Personalized product recommendations based on context-aware e-commerce platforms will play a huge role in attracting consumers.

## Acknowledgments

This work is supported by the Taizhou City Youth Science and Technology Talent Lifting Project in 2023, and Outstanding Young Key Teacher Program of Taizhou Institute of Sci. and Tech., NJUST.

## 7 REFERENCES

[1] Geetha, M. P. & Karthika, Renuka, D. (2024). Deep learning architecture towards consumer buying behaviour prediction using multitask learning paradigm. *Journal of Intelligent & Fuzzy Systems*, *46*(1), 17-38. https://doi.org/10.3233/JIFS-231116

[2] Wang, S., Yang, J., & Shang, F. (2024). A personalized recommendation model with multimodal preference-based graph attention network. *Journal of Supercomputing*, *80*(15), 200-223. https://doi.org/10.1007/s11227-024-06200-y

[3] Li, Y., Li, G, X, & Zhang, J. (2024). Research on type-aware fashion compatibility prediction based on a hybrid attention mechanism. *Multimedia Tools & Applications*, 83(30), 18101-18112. https://doi.org/10.1007/s11042-023-18101-y

[4] Kim, H., Park, C., & Kim, J. H. (2024). Multimodal Reinforcement Learning for Embedding Networks and Medication Recommendation in Parkinson's Disease. *IEEE Access*, *12*(3), 17-22. https://doi.org/10.1109/ACCESS.2024.3405009

[5] Zhai, X., Wang, Y., & Liang, L. (2025). Personalized e-learning resource recommendation using multimodal-enhanced collaborative filtering. *Knowledge-Based Systems*, *319*, 605-616. https://doi.org/10.1016/j.knosys.2025.113605

[6] Zhang, Q., Zhou, Z., & Han, X. (2024). A Recommender for Personalized Travel Planning Using Stacked Autoencoder in a Multimodal Transportation Network. *Journal of Transportation Engineering, Part A: Systems*, *150*(2), 9-37. https://doi.org/10.1061/JTEPBS.TEENG-8067

[7] Wang, J., Xie, H., & Zhang, S. (2025). Multimodal fusion framework based on knowledge graph for personalized recommendation. *Expert Systems with Applications*, *268*(12), 308-322. https://doi.org/10.1016/j.eswa.2024.126308

[8] Hassen, A. B., Ticha, S. B., & Chaibi, A. H. (2024). Extracting visual features for personalized recommendation using autoencoder. *Procedia Computer Science*, *246*(9), 3634-3643. https://doi.org/10.1016/j.procs.2024.09.194

[9] Ziaee, S. S., Rahmani, H., & Nazari, M. (2024). MoRGH: movie recommender system using GNNs on heterogeneous graphs. *Knowledge and Information Systems*, *66*(12), 7419-7435. https://doi.org/10.1007/s10115-024-02196-2

[10] Sha, Q., Ma, L., & Chen, N. (2025). A Smart Tourism Resource Information Management Platform Based on Multi-Source Data Fusion. *International Journal of High Speed Electronics & Systems*, *34*(1), 1342-1354. https://doi.org/10.1142/S0129156425401342

[11] Zhang, Z. (2024). Personalized resource recommendation method of student online learning platform based on LSTM and collaborative filtering. *Journal of Intelligent Systems*, *33*(1), 41-47. https://doi.org/10.1515/jisys-2024-0017

[12] Rajput, I. S., Tewari, A. S., & Tiwari, A. K. (2024). An autoencoder-based deep learning model for solving the sparsity issues of Multi-Criteria Recommender System. *Procedia Computer Science*, *235*, 414-425. https://doi.org/10.1016/j.procs.2024.04.041

[13] Wang, E. & Sun, Z. (2025). Research on Personalized Course Resource Recommendation Method Based on GEMRec. *Applied Sciences (2076-3417)*, *15*(3), 1075-1086. https://doi.org/10.3390/app15031075

[14] Shams, M. Y., Gamel, S. A., & Talaat, F. M. (2024). Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing & Applications*, *36*(11), 9391-9393. https://doi.org/10.1007/s00521-023-09391-2

[15] Ohtomo, K., Harakawa, R., & Ogawa, T. (2021). Personalized Recommendation of Tumblr Posts Using Graph Convolutional Networks with Preference-aware Multimodal Features. *ITE Transactions on Media Technology and Applications*, *9*(1), 54-61. https://doi.org/10.3169/mta.9.54

[16] Wang, R., Li, C., & Zhao, Z. (2025). Towards user-specific multimodal recommendation via cross-modal attention-enhanced graph convolution network. *Applied Intelligence, 55*(1), 6061-6071. https://doi.org/10.1007/s10489-024-06061-1

[17] Meng, B. & Yang, F. (2025). Design of Computer-Aided Instruction Model Based on Knowledge Graph Construction and Learning Path Recommendation. *International Journal of Web-Based Learning & Teaching Technologies*, *20*(1), 803-824. https://doi.org/10.4018/IJWLTT.366803

[18] Bansal, S., Gowda, K., & Kumar, N. (2024). Multilingual personalized hashtag recommendation for low resource Indic languages using graph-based deep neural network. *Expert Systems with Applications*, *236*(3), 15-18. https://doi.org/10.1016/j.eswa.2023.121188

[19] Cen, Y., Jiang, S., &Cai, W. (2025). EGRec: a MOOCs course recommendation model based on knowledge graphs. *Discover Applied Sciences*, *7*(6), 131-156. https://doi.org/10.1007/s42452-025-07131-w

[20] Wang, X. & Qin, J. (2024). Multimodal recommendation algorithm based on Dempster-Shafer evidence theory. *Multimedia Tools & Applications*, *83*(10), 262-270. https://doi.org/10.1007/s11042-023-15262-8

[21] Li, P., Zhan, W., & Gao, L. (2025). Multimodal Recommendation System Based on Cross Self-Attention Fusion. *Systems*, *13*(1), 57-68. https://doi.org/10.3390/systems13010057

[22] Zafari, S., De, Pagter, J., & Papagni, M. K. S. T. (2024). Trust Development and Explainability: A Longitudinal Study with a Personalized Assistive System. *Multimodal Technologies and Interaction*, *8*(3), 20-34. https://doi.org/10.3390/mti8030020

[23] Liu, K., Xue, F., & Guo, D. (2023). MEGCF: Multimodal Entity Graph Collaborative Filtering for Personalized Recommendation. *ACM Transactions on Information Systems*, *41*(2), 27-54. https://doi.org/10.1145/3544106

[24] Fang, H., Liang, J., & Sha, L. (2025). Enhanced multimodal recommendation systems through reviews integration: Enhanced multimodal recommendation. *Knowledge & Information Systems*, *67*(4), 309-323. https://doi.org/10.1007/s10115-024-02309-x
[25] Cui, Y., Liang, S., & Zhang, Y. Y. (2024). Multimodal representation learning for tourism recommendation with two-tower architecture. *PLoS ONE*, *19*(2), 23-37. https://doi.org/10.1371/journal.pone.0299370
[26] Jianghong, M. A., Sun, H., & Yang, D. (2024). Personalized Fashion Recommendations for Diverse Body Shapes with Contrastive Multimodal Cross-Attention Network. *ACM transactions on intelligent systems and technology*, *2024*(4), 15-27. https://doi.org/10.1145/3637217
[27] Peng, K., Qiao, Y., & Shi, J. (2024). Research on Multimodal Knowledge Graph Construction Method Based on Conditional Random Field Model. *2024 7th International Conference on Education, Network and Information Technology (ICENIT)*, *2024*, 60-67. https://doi.org/10.1109/icenit61951.2024.00019
[28] Yang, W. & Yang, Q. (2024). Multimodal-aware Multi-intention Learning for Recommendation. *Proceedings of the 32nd ACM International Conference on Multimedia*, *2024*, 5663-5672. https://doi.org/10.1145/3664647.3681412

**Contact information:**

**Yuanyuan WEN**
Taizhou Institute of Science and Technology,
Nanjing University of Science and Technology,
Taizhou, China

**Hongyuan WEN**
(Corresponding author)
Taizhou Institute of Science and Technology,
Nanjing University of Science and Technology,
Taizhou, China.
Taizhou Intelligent Transformation and Digital Transition Networked Key Laboratory,
Taizhou, China
E-mail: 22009025@nustti.edu.cn