International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 4460-4471

https://doi.org/10.70517/ijhsa463382

Personalized Recommendation Model of Cultural and Creative Products in Tourist Cities Based on Collaborative Filtering Algorithm

Zhengqiang He[1,*] and Yuanyuan Gu[1]
[1] College of Art, University of Sanya, Sanya, Hainan, 572000, China
Corresponding authors: (e-mail: hezhengqiang2024@126.com).

Abstract This study proposes a hybrid recommendation algorithm integrating LDA topic model and collaborative filtering, aiming to improve the accuracy and diversity of cultural and creative product recommendations in tourist cities by combining semantic analysis and user behavior modeling. The LDA topic model is utilized to extract implicit topics from user comments and product descriptions, determine the optimal number of topics through confusion and consistency indicators, and quantify the distribution of user interest preferences and product features. And combined with collaborative filtering algorithm, the user-topic association matrix is constructed, and the dynamic recommendation effect is optimized by time weight (based on Ebbinghaus forgetting curve) and distance weight (minimum diameter circle method). The experimental part validates the model performance on three datasets, Ctrip, VW Dianping and Yelp, and the RMSE of this paper's model on Ctrip dataset is 0.804, MAE is 0.752 and R-squared is 0.876 which are all better than the baseline models Caser, SLi-Rec and HGN, and on VW Dianping dataset, the RMSE's 0.791 and MAE of 0.732 also perform best, verifying its robustness. In addition, the correlation analysis of user behavior shows that the correlation coefficient of 0.946 for payment behavior and 0.913 for order placing behavior are highly correlated with interest preferences. This study effectively mitigates the data sparsity and cold-start problem through the dual-path recommendation strategy and cluster filling technique.

Index Terms collaborative filtering algorithm, tourism cultural and creative products, personalized recommendation, LDA topic model

I. Introduction

Tourism city refers to the city with more prominent urban tourism function, to meet the needs of tourists as the main function of the city, tourism city to tourism as the main industry, in which the city's cultural and creative products as an important carrier of the cultural customs and characteristics of the tourism city, plays an important role in the emotional support [1]-[4]. High-quality cultural and creative products can not only generate income for the tourism industry, but also enhance the travel experience of tourists and become part of the city's tourism culture. However, it is not easy to promote and sell these products successfully, especially in the current trend of pursuing personalized life [5]-[8].

With the rapid development and application of artificial intelligence, “personalized recommendation” has become a common term in the business field, which also lays a technical foundation for the personalized recommendation of cultural and creative products in tourism cities [9]-[11]. In the field of industrial intelligence, collaborative filtering algorithm, as one of the core technologies of personalized recommendation system, plays a crucial role in optimizing user experience and improving the accuracy of product or content recommendation, and has been widely used in the fields of e-commerce, social media, music and video recommendation, etc. [12]-[15]. In the personalized recommendation of urban cultural and creative products, the algorithm discovers the similarity between tourists and items by mining the hidden patterns in the tourists' behavioral data, so as to achieve personalized recommendation services, which plays an important role in improving the tourists' experience, enhancing the city's image and economic development [16]-[19].

A hybrid recommendation algorithm integrating LDA topic model and collaborative filtering is proposed, aiming to improve the accuracy and diversity of recommendation by combining semantic analysis and user behavior modeling. First, the LDA topic model is used to extract implicit topics from user comments and travel product descriptions, quantify user interest preferences and product features, and optimize the model performance through confusion and consistency indicators. Secondly, combined with the collaborative filtering algorithm, we extracted the theme distribution of tourism products and user interest preferences, and constructed the user-theme association matrix

through normalization and interest value calculation. We also optimize the recommendation results by user rating data and topic distribution, propose a hybrid framework combining LDA topic semantics and collaborative filtering to solve the cold-start problem, and use time and distance weights to improve the dynamic recommendation effect. Finally, the hybrid recommendation framework is designed to design a dual-path recommendation strategy based on users and items, and optimize the coverage and accuracy of the recommendation results through the clustering filling and replacement strategy. Synthesize the collaborative filtering strategy based on users and items, and introduce the time weight and distance weight to improve the algorithm to dynamically adapt to user interest changes.

II. Hybrid Recommendation Algorithm for Cultural and Creative Products in Tourist Cities Based on LDA Theme and Collaborative Filtering

II. A.LDA Subject Modeling

II. A. 1) Introduction to LDA Topic Modeling

Topic modeling is a commonly used model in natural language processing for automatically extracting topic information from a large number of documents. The core idea of topic modeling is that each document can be viewed as a mixture of multiple topics, and each topic consists of a set of words.

A commonly used topic model is the latent Dirichlet distribution (LDA).The LDA topic model assumes that each document is a mixture of multiple topics, and each topic is a probability distribution of a set of words.The LDA infers the distribution of topics for each document and the distribution of words for each topic by maximizing a likelihood function over the entire set of documents.

II. A. 2) Determination of the optimal number of topics

Through the literature combing found that a large number of empirical studies have confirmed that the effect of LDA topic extraction and the number of potential topics K value has a direct relationship, the results of the topic extraction is very sensitive to the K value, and the K value that makes the best effect of the LDA topic model is called the optimal number of topics. At present, the more common method to determine the optimal number of topics is to calculate the degree of confusion and consistency.

Perplexity is a metric used to measure the predictive performance of a topic model on new documents. It is widely used in the field of natural language processing for the evaluation of language models and topic models. For LDA topic models, the confusion degree is calculated as:

$$Perplexity = \exp\left\{-\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D}N_d}\right\} \tag{1}$$

where D is the number of documents in the test set. N_d is the number of words in document d. $P(w_d)$ is the likelihood probability of document d computed on the test set. The lower the perplexity, the better the predictive performance of the model on the test set.

Consistency is a metric used to measure the quality of topics generated by the topic model. It measures the semantic coherence of topics by evaluating the correlation between words under the same topic. In LDA, coherence is calculated as:

$$Coherence = \frac{2}{T(T-1)}\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}Sim(w_i, w_j) \tag{2}$$

where T is the number of topics, w_i and w_j are the two words in the topic, and $Sim(w_i, w_j)$ is the similarity score between the words w_i and w_j, which is usually measured using a metric such as pointwise mutual information. Higher consistency indicates that the words in the topic are more related and the topic is of better quality.

The combination of these two metrics can provide researchers with a comprehensive assessment of topic model performance and topic quality, and determine the optimal number of topics, K, for LDA topic models.

II. B.Thematic analysis based on LDA modeling

By determining the optimal number of topics for the LDA model, the model is able to effectively extract the potential topics in the reviews of tourism products. On this basis, the topic distribution characteristics of tourism products are

further analyzed and the user's interest preference in topics is quantified to provide data support for the construction of subsequent recommendation algorithms.

II. B. 1) Theme analysis of tourism products based on LDA modeling

Based on the obtained distribution of preferred topics for reviews of tourism products, the probability of each topic for all reviews of the same tourism product is summed up and divided by the total number of reviews of the same tourism product in order to obtain the normalized distribution of topics for each tourism product, which is calculated as shown in equation (3).

$$Travel\_pro_x = \frac{1}{M}\sum_{i=1}^{M}\theta_{xi} \tag{3}$$

where $Travel\_pro_x$ is the topic distribution of travel product x, θ_{xi} is the topic distribution of the i th review of travel product x, and M is the number of reviews of travel product x in the text data set of travel product reviews. After calculating by the formula, 5 tourism products were randomly selected, and Figure 1 shows the topic distribution of 5 tourism products.

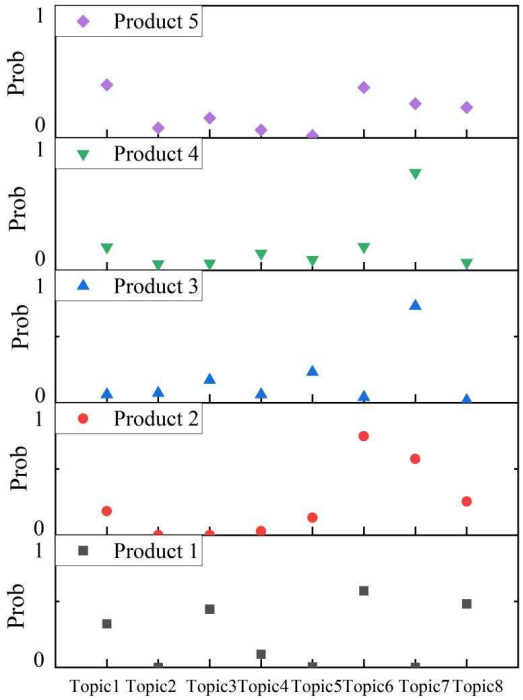

Figure 1: Tourism Products - Theme Distribution Map

Figure 1 randomly shows the distribution of topics for some of the tourism products. It can be seen at the same time through Figure 1 that the main characteristic information of tourism product 1 is related to Topic7, i.e., it is about the way of group tours. Another example is that travel product 2 is about Topic6, so this travel product belongs to cost-effective travel. As a whole, the topics generally have a large distribution only on one or two topics, while there is almost no distribution on other topics, so the topic extraction is more effective.

II. B. 2) User Interest Extraction Based on LDA Modeling

In order to obtain the degree of interest value of each user in the theme of each recommended tourism product, the user preference theme distribution is first calculated, specifically by multiplying the theme feature value of each tourism product that each user has rated in the training set with the score, and then summing up the product values of the theme feature value and the score in the same user and normalizing them to obtain the user preference theme distribution. Among them, the size of the score of the user's rating of the tourism product indicates the user's favorite degree of the characteristic theme contained in the tourism product, so the contribution of each tourism product to the construction of the user preference theme distribution is different. The calculation formula is shown in equation (4).

$$U_{ui} = \frac{\sum_{n=1}^{N} Travel\_pro_{uni} \times Score_{un}}{\sum_{i=1}^{K}\sum_{n=1}^{N} Travel\_pro_{uni} \times Score_{un}} \tag{4}$$

where the computation result U_{ui} represents the preference value of user u for the ith theme, the numerator part $Travel\_pro_{umi}$ represents the eigenvalue of the ith feature theme of the nth travel product rated by user u, $Score_{un}$ represents the rating of the nth travel product by user u, and N is the number of the travel products user u has purchased the number of tourism products, and K is the number of theme categories.

According to the formula in equation (4), the distribution of users' preferred themes is calculated, but a single distribution of users' preferred themes does not accurately portray the degree of users' interest in the themes, and it is very likely that the high value of users' preference for a certain theme is due to the fact that the average probability of this theme appearing in all tourism products is high compared to the average probability of other themes appearing in all tourism products, which leads to more tourism products being categorized as this theme, even if users are not interested in this theme. In order to arrive at a distribution that accurately reflects the user's interest level, it is necessary to test whether the distribution of user preference themes is uniform. Therefore, the average eigenvalue of each characteristic theme eigenvalue of tourism products in the training set is calculated over all the tourism products that have been rated by users, and it is observed whether the average value is close to the average value or not, and the calculation formula is shown in Equation (5).

$$Average_i = \frac{\sum_{n=1}^{N} Travel}{N} \tag{5}$$

where $Average_i$ represents the average theme value of the ith featured theme in all tourism products $Travel\_pro_{ni}$ represents the feature value of the ith theme for the nth tourism product, and N denotes the sum of the number of tourism products rated by users in the training set. The average theme value of each theme represents the distribution proportion of that theme among all tourism products. The distribution of the average theme value of each theme is calculated using Equation (4). The distribution of the average theme value of each theme is calculated using formula (5), and the average theme value is shown in Table 1.

Table 1: Average topic value

Topic		Average value of the topic
Topic1	Meal	0.0961
Topic2	Family trip	0.1519
Topic3	High cost-performance travel	0.1231
Topic4	Accommodation service	0.1126
Topic5	Tour guide service	0.1367
Topic6	Recreational activities	0.1176
Topic7	Relaxed and casual Series	0.1365
Topic8	Environment	0.1255

There is a certain gap in the average theme value of each feature theme appearing in all travel products, among which Topic2 (cost-effective travel) has the highest average feature value, indicating that this feature theme has the highest average probability of appearing in all travel products. So in order to more accurately portray the user's interest in different topics. In this paper, based on the preference value of each user for each theme calculated by Equation (4), divided by the average feature value of each theme, we obtain the ratio of the preference value of each theme to the average value of the feature theme, i.e., the user's interest value in different themes. The calculation formula is shown in equation (6).

$$I_{ui} = \frac{U_{ui}}{Average_i} \tag{6}$$

where I_{ui} denotes the interest value of user u on the ith topic, $Average_i$ denotes the average feature value of the ith featured topic, and U_{ui} denotes the preference value of user u on the ith topic. In this study, the interest value

is used to measure how much users are interested in different topics. The larger the interest value is, the greater the user's interest in the topic, and the user is also more inclined to choose tourism products with larger feature values for that topic. Setting 1 as the interest value threshold, when the interest value is greater than 1 indicates that users are interested in the theme, then the proportion of users choosing tourism products with high values of the theme's characteristics exceeds the proportion of such tourism products in all tourism products, and conversely, when the interest value is less than 1, the proportion of users choosing tourism products with high values of the theme's characteristics is lower than the proportion of such tourism products in all tourism products.

II. C.Hybrid Recommendation Algorithm for Cultural and Creative Creations in Tourist Cities Incorporating LDA Theme Model and Collaborative Filtering

After completing the association modeling between user interests and product topics, it becomes crucial to combine the topic information at the semantic level with collaborative filtering algorithms. In this section, a hybrid recommendation framework is proposed to alleviate the cold-start problem of collaborative filtering through LDA topic modeling, and at the same time introduce a dynamic weighting mechanism to adapt to the spatial and temporal changes of user interests.

The platform for cultural and creative products in tourist cities requires recommendation algorithms to balance accuracy and diversity, while alleviating the cold-start problem and data sparsity problem of collaborative over-taki algorithms. Therefore, the platform adopts the LDA topic model to calculate the topic probability distribution of rural attraction information as a way to find the association between items, so as to generate the recommendation list $TOPN_1$ instead of the item-based collaborative filtering recommendation. If the recommended user is a new user, $TOPN_1$ is the final recommendation list. If the user has historical behavior data, obtain the user's historical ratings and construct the rating matrix, use the user-based recommendation algorithm to generate the recommendation list $TOPN_2$, $TOPN_1$ and $TOPN_2$ cross to generate the final recommendation list. This method takes into account the diversity and accuracy, and at the same time, it can solve the cold-start problem caused by new users without historical behavioral data at the beginning of the system operation. The LDA model part of the recommendation model is written in Java, and the collaborative filtering recommendation part is written in Python as a script file, and Java executes the Runtime.getRuntime() command at a later stage to start the process to execute the script file and realize the call to the recommendation algorithm.

When performing user-based collaborative filtering recommendation, for the problems of sparse user rating matrix and low scalability of traditional collaborative filtering algorithms, this paper proposes to utilize Mini Batch K-Means clustering algorithm to cluster and fill the user rating matrix. The user clustering results are used as the nearest neighbor set, and the nearest neighbor range is reduced to improve scalability. When users browse the attractions in the cultural and creative products platform of tourist cities, their interest preferences will change with the change of time and distance factors, so the platform recommendation system utilizes the trend of the Ebbinghaus forgetting curve to add the time weights, and improves the distance weights with the minimum diameter circle method, and substitutes them into the collaborative filtering algorithm to generate the recommendation list.

The hybrid recommendation algorithm incorporating LDA topic modeling and collaborative filtering is divided into two major modules. The first module generates tags based on attraction information and user comments, associates tags with users and attractions, constructs user-tag matrix and attraction-tag matrix, and substitutes them into the LDA topic model to obtain the user-tag probability matrix at the semantic layer and the cultural and creative product-tag probability matrix at the semantic layer for the tourist city; the probability distributions of the user's preference for the product are obtained through the two probability matrices. Substituting Bayes' theorem formula to calculate the preference value of user u for attraction v, in order to make recommendation. The second module applies the Pearson correlation coefficient to calculate the similarity between users, and then uses the collaborative filtering algorithm with improved time and distance all-in-one to make personalized recommendations.

II. D.Hybrid collaborative filtering algorithm for tourism products

To further improve the robustness of tourism product recommendation, this section proposes a collaborative filtering optimization strategy for tourism scenarios based on the hybrid framework. The scalability of traditional algorithms under sparse data is solved by integrating the dual-path recommendation mechanism of users and items, and combining the clustering filling and substitution strategies.

II. D. 1) Hybrid collaborative filtering algorithm design

Through the analysis in the previous section, there are many problems in personalized recommendation of tourism products, and the commonly used recommendation algorithms cannot be directly applied to the recommendation of tourism products. For this reason, this paper makes improvements on the basis of traditional collaborative filtering algorithm, combines User-based CF and Item-based CF two collaborative filtering algorithms, and proposes a hybrid

collaborative filtering algorithm for tourism products. The hybrid collaborative filtering algorithm for tourism products is shown in Figure 2.

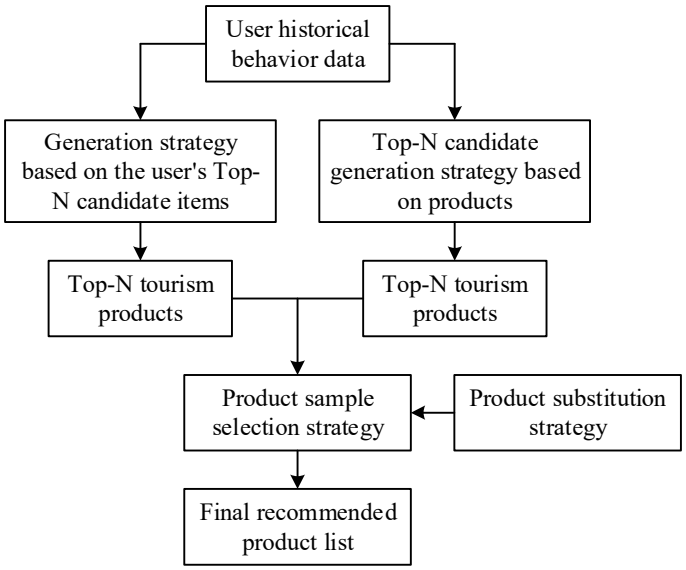

Figure 2: Hybrid Collaborative Filtering Algorithm for Tourism Products

In the above figure, the algorithm first generates Top-N tourism products by inputting users' historical behavioral data using user-based Top-N candidate generation strategy, respectively, and generates Top-M tourism products using product-based Top-M candidate generation strategy, and then integrates the recommended results based on users' clicks on the recommended products through product selection strategy, and uses product substitution strategy to replace and supplement tourism products when there are insufficient recommended products. In case of insufficiency, the product replacement strategy is used to replace and supplement the tourism products, and finally the final list of recommended products is output.

II. D. 2) User-based Top-N Candidate Generation Strategy

According to Fig. 2, the user-based Top-N candidate generation strategy in the hybrid collaborative filtering algorithm needs to utilize the following four steps: user's data representation, user's similarity computation, nearest neighbor query, and predictive scoring to generate Top-N candidates.

(1) Data representation

An m*n-order user-one-product rating matrix is built, where m denotes the user and n denotes the travel product, and the value $R_{m,n}$ in the table represents the rating of user m for travel product n.

(2) Similarity calculation of users

After obtaining the user rating matrix, the similarity between users needs to be calculated. There are three main methods to calculate the similarity between users, which are cosine similarity, modified cosine similarity and Pearson correlation similarity.

User ratings can be regarded as vectors on the n-dimensional product space, and the similarity between user u and user v is measured by the cosine angle picking between the vectors, and the larger the cosine value of the buy angle of the two asked quantities indicates that the degree of similarity between the two users is higher. Let the ratings of user u and user v on the n-dimensional product space be represented as vectors $\vec{u}$ and $\vec{v}$ respectively, then the similarity $sim(u,v)$ between user u and user v is shown in Equation (7) below:

a) Cosine similarity

$$sim(u,v) = cos(\vec{u},\vec{v}) = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\cdot\|\vec{v}\|} \tag{7}$$

The numerator on the right-hand side of the above equation is the inner product of the two user rating vectors, and the denominator is the product of the modes of the two user rating vectors.

b) Modified cosine similarity

The cosine similarity measure does not take into account the problem of rating scales of different users, so the modified cosine similarity measure improves the above problem by subtracting the average user ratings of the

products. Let the set of products that have been jointly rated by user u and user v be denoted by $I_{u,v}$ and I_u and I_v denote the set of products that have been rated by user u and user v, respectively, then the similarity of user u and user v $sim(u,v)$ is shown in Eq. (8) below:

$$sim(u,v)=\frac{\sum_{c\in I_{u,v}}(R_{u,c}-\overline{R_u})(R_{v,c}-\overline{R_v})}{\sqrt{\sum_{c\in I_u}(R_{u,c}-\overline{R_u})^2}\sqrt{\sum_{c\in I_v}(R_{v,c}-\overline{R_v})^2}} \tag{8}$$

where $R_{u,c}$ denotes user u's rating of product c, $R_{v,c}$ denotes user v's rating of product c, and $\overline{R_u}$ and $\overline{R_v}$ denote the average ratings of the product by user u and user v, respectively.

c) Pearson correlation similarity

Let the set of products that have been jointly rated by user u and user v be denoted by $I_{u,v}$, then the similarity $\sin(u,v)$ between user u and user v is shown in equation (9) below:

$$sim(u,v)=\frac{\sum_{c\in I_{u,v}}(R_{u,c}-\overline{R_u})(R_{v,c}-\overline{R_v})}{\sqrt{\sum_{c\in I_{u,v}}(R_{u,c}-\overline{R_u})^2}\sqrt{\sum_{c\in I_{u,v}}(R_{v,c}-\overline{R_v})^2}} \tag{9}$$

where $R_{u,c}$ denotes user u's rating of tourism product c, $R_{v,c}$ denotes user v's rating of tourism product c, and $\overline{R_u}$ and $\overline{R_v}$ denote the average ratings of user u and user v on tourism products, respectively.

The calculation result of Pearson correlation coefficient, r, is distributed in the interval [-1, 1], which is used to characterize the strength of the correlation between the two, and the value of r tends to 1, indicating the higher the correlation between the two, and in the practical application, only the case where the value of r is greater than zero should be considered.

(3) Nearest neighbor query

After completing the user similarity calculation, we need to generate a neighbor set $N_u=\{N_1,N_2,\cdots N_k\}$ for the recommended target user u by similarity computation, where the elements are sorted in descending order according to the size of the similarity, so that the similarity between u and N_1 is $sim(u,N_1)$ is the highest.

(4) Predictive scoring

After obtaining the set of neighbors of the recommendation target, the algorithm is needed to derive the corresponding recommendation result, i.e., to make the prediction of the score. Let the set of nearest neighbors of user u be N_u, then the predicted score P_{uj} of user u for tourism product j is calculated as shown in Equation (10) below:

$$P_{u,j}=\overline{R_u}+\frac{\sum_{n\in N_u}sim(u,n)\times(R_{n,j}-\overline{R_n})}{\sum_{n\in N_u}(|\,sim(u,n)\,|)} \tag{10}$$

The center-weighted calculation is used in the above equation, where $\overline{R_u}$ and $\overline{R_n}$ denote the average ratings of user u and user n on all tourism products, respectively, $R_{n,j}$ denotes the ratings of user n on tourism product j, and $sim(u,n)$ denotes the similarity between user u and the similarity between user n.

The above algorithm can predict the users' ratings for the unrated tourism products, and finally the Top-N tourism products with the highest ratings are taken as candidate outputs.

III. User behavior analysis and hybrid recommendation model experimental validation

The hybrid recommendation algorithm proposed in Chapter 2 needs to rely on high-quality user behavior data and experimental validation. Therefore, Chapter 3 will discuss in depth the collection method of user behavioral data, analyze the association between behavioral indicators and interest preferences, and verify the effectiveness of the model through multi-dataset experiments.

III. A. User behavior data collection and correlation analysis

III. A. 1) User behavior data collection

A complete and comprehensive study based on the perspective of user behavior analysis needs to obtain all the records of user behavior on the website during the user's life cycle, and the channels for obtaining these records include website data, Web server logs and client-side user behavior data. Website data mainly refers to the information displayed on the website page, including product name, category, user account information, user order information, user comment information, user sharing records and other data; Web server logs are the complete server logs recorded in response to user requests, including user ID, user IP address, request access time, request response time, request content link, request parameters, and user behavior data on the client side. time, request content links, request parameters (page form data), etc.; client-side user behavior data refers to the behavioral information that occurs in the user's client, the use of client-side running programs, such as can be W to develop a set of client software for the user to download and locally install, similar to the common QQ software, but also can be implanted in the user's request for the page implanted in the JavaScript code, by implanting the JavaScript agent code to collect all user behavioral data that occurs in the browser, including: page clicks, page input information and other individual user behavioral data.

III. A. 2) Correlation analysis between indicators

Based on the data structure design, it can be seen that the user behavior indicators captured in this research can be directly quantified to calculate the score, including page clicks, page browsing, page dwell time, order placement, order cancellation, payment, and user ratings, totaling seven items. In order to calculate the subsequent user behavior score based on them, it is necessary to analyze the relevance of each indicator and discuss the weight.

First of all, among the seven indicators, the user rating belongs to the accessible explicit behavioral data, which can be seen from the acquisition of website information, the indicator already has an explicit score, that is, it is divided into three levels: satisfied (5 points), general (0 points), dissatisfied (-5 points). Therefore, when measuring the relevance of other indicators, the user rating indicator can be used directly as a reference indicator, which is equivalent to an intuitive representation of the user's interest and preference.

Secondly, the ultimate interest of users in page browsing is to motivate them to consume products and services. Therefore, among the six implicit user behaviors, the most direct metrics that reflect users' interest preferences are the three behaviors of users' order placing, order cancellation and payment. However, taking into account the positive and negative aspects of placing and canceling orders for a behavior, therefore, when setting the weight of the indicators can be determined one, and reverse the other, in order to simplify the correlation analysis and avoid multi-factor interference.

Again, correlation analysis of user behavior indicators.

In this study, SPSS19.0 was used to analyze the correlation between the six indicators of page clicks, page views, page dwell time, orders, payments and user ratings, and the correlation coefficient matrix between the user behavior metrics and user interests and preferences is shown in Figure 3.

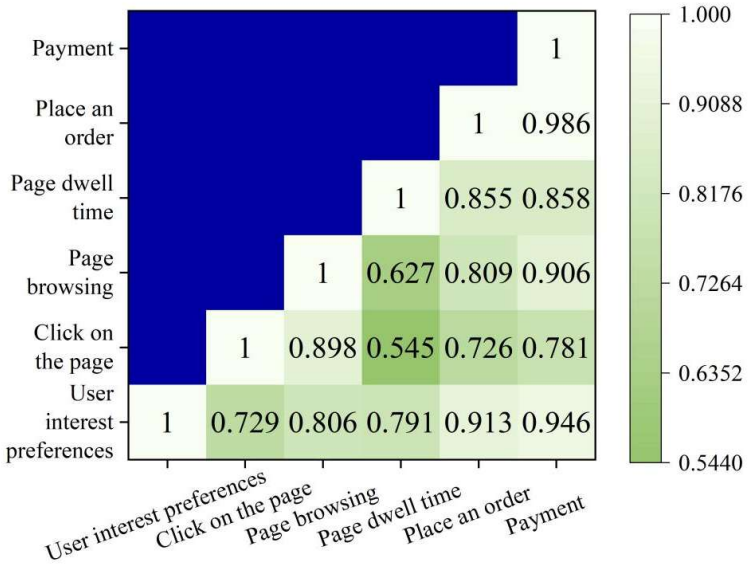

Figure 3: Correlation coefficient matrix between user behavior and interest preference

The correlation coefficients of user interest preference with page click, page view, page dwell time, order placing and payment are 0.729, 0.806, 0.791, 0.913 and 0.946 respectively, indicating that these behaviors are significantly and positively correlated with interest preference. Among them, the correlation coefficient of payment behavior is the highest, 0.946, indicating that users' actual payment behavior best reflects their interest preferences; and the strong correlation between placing an order and payment, 0.986, indicates that the two are highly correlated. In addition, the correlation coefficient between page clicking and browsing is 0.898, indicating that users' clicking and browsing behaviors on the page have high consistency.

III. B. Content-based recommendations

After completing the collection and relevance analysis of user behavioral data, the content-based recommendation model needs to be further constructed to supplement the semantic limitations of collaborative filtering algorithms.

The content-based recommendation process generates the user's interest preference model based on the user's historical data and resources, and then recommends products or services similar to the user's interest preference model. The specific realization process includes four aspects.

First, the product representation stage. Web page text mining is performed by crawler software to obtain the characterization of the product and convert it into a computer-recognizable mathematical representation;

Second, construct user model and complete user portrait. Collecting the user's historical behavior information, the user constructs the preference model and completes the user portrait;

Third, based on the results of the user model and user portrait, first, count the most commonly used labels of each user; second, for each label, count the products that have been hit with this label the most times.

Fourth, generate recommendation results, measure the user's interest in the product to be recommended, and generate a TopN list of recommended products that the user may be interested in based on the descending order of interest; here we focus on analyzing how to calculate the user's interest in the product. According to the results of user model and user portrait, the respective interest tag scores of users and cultural and creative products of tourist cities to be recommended are generated, and the keywords of a user's interest tags are "five-star service", "summer", "cost-effective", "couple" and "seafood", and the user's product interest tags are shown in Table 2.

Table 2: The product interest tag of a certain user

	Five-star service	Summer	Cost performance	Couple	Seafood	Product interest degree
User	0.67	0.82	0.75	0.33	0.45	-
Product A	0.32	0.77	0.62	0.32	0.49	1.637
Product B	0.90	0.78	0.34	0.66	0.91	2.125
Product C	0.86	0.47	0.72	0.54	0.76	2.022
Product D	0.95	0.27	0.27	0.82	0.74	1.664
Product E	0.71	0.27	0.84	0.64	0.56	1.790
Product F	0.45	0.49	0.37	0.66	0.59	1.464
Product G	0.71	0.62	0.52	0.9	0.82	2.040
Product H	0.62	0.26	0.36	0.83	0.28	1.299
Product…	…	…	…	…	…	…

From this, the user's interest level for each product is obtained:

For product A, the interest level is: 0.67*0.32+0.82*0.77+0.75*0.62+0.33%0.32+0.45*0.49=1.637

For product B, the interest level is: 0.67*0.90+0.82*0.78+0.34*0.75+0.33%0.66+0.45*0.91=2.125

For product C the interest level is: 0.67*0.86+0.82*0.47+0.34*0.72+0.33%0.54+0.45*0.76=2.022

By analogy, products D, E, F, G, and H have interest levels of 1.664, 1.790, 1.464, 2.040, and 1.299, respectively.

Finally, TopN (Product B, Product G, Product C, Product E, Product D, Product A, Product F, Product H .....) are recommended to users based on the descending order of interest.

III. C. Experimental Design and Result Evaluation of Hybrid Recommendation Models

While content-based recommendation models provide semantic-level support for hybrid frameworks, this section will comprehensively evaluate the performance of hybrid recommendation models incorporating LDA topics and collaborative filtering in real-world scenarios through experimental design and multiple dataset comparisons.

III. C. 1) Dataset and experimental environment

In order to evaluate the effectiveness of this paper's hybrid algorithm based on LDA topics and collaborative filtering, this chapter conducts a large number of experiments on three different datasets, which are the Ctrip travel dataset (Ctrip), the Dianping travel dataset (Dianping), and the Yelp travel dataset (Yelp). The three datasets are very different in terms of scale size and number of interaction records, and the corpus includes both Chinese text and English text, which can reflect the performance of the algorithm on datasets of different sizes more comprehensively.

The experimental key hyperparameters of the model proposed in the study are as follows: the LSTM unit is 64*4, the number of EANet attention heads is 4, the user tower fully connected stack layers are 256*1, 128*2, 64*1, 32*1, the item tower fully connected stack layers are 512*1, 256*2, 128*2, 64*2, 32*1, and the optimizer is Adam(lr=1e-5).

III. C. 2) Baseline model and evaluation indicators

(1) Evaluation metrics

For all the comparison experiments, this chapter uses Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R-squared) to evaluate this paper's model and the baseline model. These three metrics are used to reflect the performance of the models in predicting user ratings, thus providing a comprehensive comparison of the strengths and limitations of this paper's model and various baseline models in terms of their recommendation ranking capabilities.

Root Mean Square Error: this metric is a common metric for assessing the performance of regression models and is widely used especially in rating prediction tasks. It measures the standard deviation of the difference between the model's predicted values and the actual observed values. It is calculated by first calculating the square of the difference between each pair of predicted and true values, then averaging these squared differences, and finally calculating the square root of this average. It is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

where y_i represents the actual rating, $\hat{y}_i$ represents the predicted rating, and n represents the number of predicted targets. The smaller the value of RMSE, the higher the accuracy of the model's prediction. In recommender systems, RMSE can be used to quantify the error of predicting the user's rating of an item, and thus evaluate the predictive ability of the recommendation model.

Mean Absolute Error: this metric is another metric used to measure the prediction accuracy of a model, and is particularly applicable to the task of rating prediction. The MAE measures the mean absolute value of the difference between the model's predicted values and the actual observed values, which is determined by calculating the absolute value of the difference between each pair of predicted values and the true values, and then averaging these absolute differences. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{12}$$

where y_i represents the actual score, $\hat{y}_i$ represents the predicted score, and n represents the number of predicted targets. Compared with RMSE, MAE is not squared, so it is less sensitive to larger errors. In recommender systems, MAE can effectively reflect the average deviation between the predicted score and the actual score, so as to evaluate the accuracy of the recommendation model.

Coefficient of determination: this metric is is a statistical measure of model fit and is widely used in regression analysis. In the rating prediction task, the coefficient of determination describes the degree of correlation between the ratings predicted by the model and the ratings given by the actual users.The value of R-squared usually lies between 0 and 1, which can be interpreted as the proportion of the predicted value variance to the total variance, where the predicted value variance is the difference between the predicted values and their mean values, and the total variance is the difference between the actual values and their mean values. It is calculated as:

$$R - squared = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2} \tag{13}$$

where y_i represents the actual scores, $\hat{y}_i$ represents the predicted scores, $\bar{y}$ represents the average of the actual scores, and n represents the number of predicted targets. The range of the value of the R-squared is usually [0,1], and the higher the value of the R-squared, the better the predictive performance of the model. The higher the value of R-squared, the better the prediction performance of the model.

(2) Baseline model

In order to verify the performance of the models in this paper and to ensure the fairness of the comparison, three models in the field of sequence recommendation are used in this chapter to compare with the models proposed in the article. These models are Caser, SLi-Rec and HGN, which are the latest research in the field of sequence recommendation in recent years.

The Caser model maps sequences of items into temporal and spatial images and learns sequences of behaviors as local features of the image using filters for capturing user preferences and sequence features.

The SLi-Rec model improves on the traditional recurrent neural network using two perceptrons to better take into account situational information and control the state transitions of the “door”.The SLi-Rec model extracts the user's long and short term preferences and then uses attention to fuse the two to generate user vectors.

The HNG model integrates a Bayesian personalized ranking approach and consists of a functional gating module, an instance gating module and an interaction module. It captures the user's long term and short term interests through this hierarchical gating mechanism, which enables it to manage and utilize the information in the data in a more granular way, thus improving the relevance and accuracy of the ranking.

III. C. 3) Analysis of experimental results

Table 3 shows the prediction results of this paper's model and the baseline model on the three datasets. From the results, this paper's model achieves good performance on both datasets, achieving optimal solutions on most of the metrics.

Table 3: Compare the experimental results on different datasets

Dataset	Model	RMSE	MAE	R-squared
Ctrip Travel Dataset	Caser	0.839	0.808	0.832
	SLi-Rec	0.822	0.778	0.834
	HNG	0.849	0.765	0.858
	OURS	0.804	0.752	0.876
Dianping Travel Dataset	Caser	0.854	0.783	0.849
	SLi-Rec	0.841	0.797	0.859
	HNG	0.823	0.751	0.841
	OURS	0.791	0.732	0.864
Yelp Travel Dataset	Caser	0.859	0.821	0.799
	SLi-Rec	0.844	0.813	0.793
	HNG	0.836	0.811	0.809
	OURS	0.833	0.801	0.815

From the overall results of the experiments, it can be found that the Caser model is the most ineffective, which is because it only takes into account the user's long-term preference after the refinement of user's behavior, and does not take into account the user's recent short-term needs, resulting in its failure to fully reflect the user's current interest changes in predicting ratings. The lack of response to short-term needs makes the model underperform when dealing with recommendation tasks that are affected by temporary events, especially when facing situations such as travel recommendation where users' needs are highly influenced by situational factors.The SLi-Rec model takes into account both long- and short-term preferences, and its prediction results are significantly improved, which proves the effectiveness of taking into account both long- and short-term preferences, but it uses the same network architecture to analyze different preferences. The recurrent neural network, although suitable for shorter sequences, faces the problem of gradient disappearance when dealing with longer visitor interaction sequences, which limits the model's ability to capture long-term preference relationships, and thus SLi-Rec's prediction results on both datasets fall short of those of the proposed model.The HNG model, although integrating the user's long- and short-term preferences, does not consider the fusion of preferences, which simply combines the two. The HNG model does not consider the fusion of preferences, and it simply connects the two as the user's characteristics. This approach fails to balance the weights of the two, and the important preferences are weakened by the non-important

preferences to a certain extent, so the prediction effect of the HNG model on different datasets varies greatly, and the model lacks stability, and its prediction effect on the Ctrip dataset is much worse than that on the Yelp dataset.

The proposed model uses different network models to extract different preferences, and adopts the attention mechanism to assign weights to each preference, and it outperforms the baseline model in both RMSE and MAE metrics. On the Ctrip dataset, the RMSE of 0.804, MAE of 0.752, and R-squared of 0.876 are better than the baseline model, which indicates that its ratings prediction error is lower and the fit is better. On the Volkswagen Dianping dataset, this paper's model also performs best with an RMSE of 0.791 and an MAE of 0.732. On the Yelp dataset its RMSE of 0.833 and MAE of 0.801 as well as R-squared of 0.815 still outperform the other models, indicating that it still maintains a better explanatory ability in complex scenarios. Overall, the model in this paper significantly improves the accuracy of recommendation by integrating the advantages of LDA topic and collaborative filtering.

IV. Conclusion

In this paper, a hybrid recommendation framework integrating LDA topic model and collaborative filtering is proposed to address the problems of data sparsity, cold start and dynamic interest change in recommending cultural and creative products in tourist cities.

The analysis of user behavior shows that the correlation coefficient of payment behavior (0.946) and order placing behavior (0.913) are the core characterization of interest preference, while the page dwell time (0.791) and browsing behavior (0.806) can be used as auxiliary indicators.

The RMSE of this paper's model on Ctrip, Dianping and Yelp datasets is reduced by 4.3%, 5.9% and 0.4% respectively, MAE is reduced by 3.6%~6.5%, and R-squared is improved by 1.8%~2.7%, which verifies the advantages of fusing semantic and behavioral data.

References

[1] Postma, A., Buda, D. M., & Gugerell, K. (2017). The future of city tourism. Journal of Tourism Futures, 3(2), 95-101.

[2] Lee, P., Hunter, W. C., & Chung, N. (2020). Smart tourism city: Developments and transformations. Sustainability, 12(10), 3958.

[3] Qiu, L. (2020). Design of cultural and creative products of marine cultural tourism. Journal of Coastal Research, 112(SI), 100-102.

[4] Wu, Y. (2021). Design of tourism cultural and creative products based on regional historical and cultural elements. In E3S Web of Conferences (Vol. 251, p. 03004). EDP Sciences.

[5] Al-Ababneh, M. (2019). Creative cultural tourism as a new model for cultural tourism. Journal of Tourism Management Research, 6(2), 109-118.

[6] Liang, Y., & Qi, Z. (2021). Research on Innovative design of tourism cultural and creative products from the perspective of Huizhou intangible cultural heritage culture: Taking wood carving patterns as an example. Sci. Soc. Res, 3, 228-232.

[7] Liu, X. (2020). Design of Creative products for marine tourism culture. Journal of Coastal Research, 110(SI), 60-63.

[8] Zhang, S. N., Zhang, W. Y., Li, Y. Q., Ruan, W. Q., & Zhou, Y. (2023). Will visual peripheral cues motivate you to purchase tourism cultural and creative products? Evidence from China. Asia Pacific Journal of Tourism Research, 28(12), 1434-1451.

[9] Gorgoglione, M., Panniello, U., & Tuzhilin, A. (2019). Recommendation strategies in personalization applications. Information & Management, 56(6), 103143.

[10] Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. applied sciences, 10(21), 7748.

[11] Liao, S. H., Widowati, R., & Chan, S. C. (2025). The retail collaborative recommendations for personalized product recommendations. International Journal of Retail & Distribution Management, 53(5), 431-447.

[12] Kluver, D., Ekstrand, M. D., & Konstan, J. A. (2018). Rating-based collaborative filtering: algorithms and evaluation. Social information access: Systems and technologies, 344-390.

[13] Cunha, T., Soares, C., & de Carvalho, A. C. (2018, September). CF4CF: recommending collaborative filtering algorithms using collaborative filtering. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 357-361).

[14] Yin, C., Shi, L., Sun, R., & Wang, J. (2020). Improved collaborative filtering recommendation algorithm based on differential privacy protection. The Journal of Supercomputing, 76(7), 5161-5174.

[15] Hong, B., & Yu, M. (2019). A collaborative filtering algorithm based on correlation coefficient. Neural Computing and Applications, 31, 8317-8326.

[16] Huang, J. (2024). Personalized Recommendation Method for Cultural Creative Products in Tourism Cities Based on User Profiles. Procedia Computer Science, 243, 1133-1142.

[17] Tang, J., & Zhang, Y. (2025). Research on Personalized Design and Recommendation Systems for Cultural and Creative Products Based on User Behavior Data. International Journal of High Speed Electronics and Systems, 2540268.

[18] Zhang, S. (2025). Integrating User Profiles and Collaborative Filtering for Smart Recommendation of Tourism City Cultural and Creative Products. International Journal of High Speed Electronics and Systems, 2540296.

[19] He, Z., & Gu, Y. (2025). Research on Recommendation Framework for Personalized Push of Cultural and Creative Products in Tourist Cities Based on Multi-Level Computational User Profiling. J. COMBIN. MATH. COMBIN. COMPUT, 127, 4837-4856.