J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 5899--5926

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study on Cross-Cultural Analysis of Social Media Data and Leisure Travel Preference Prediction Supported by Cluster Analysis Algorithm

Dongxia Wu[1,✉]

[1] *School of Culture and Tourism, Huangshan Vocational and Technical College, Huangshan, Anhui, 245000, China*

ABSTRACT

In this paper, leisure tourism is taken as the entry point of the research, and the fused location key point features are added and integrated with the multidimensional features of time, location and space to construct an accurate portrait of social media tourism users. On the basis of tourism user profiles, a two-step clustering algorithm is combined to carry out cross-cultural analysis of social media data, to explore and excavate the performance of users' tourism preferences under the cross-cultural ability of social media. Meanwhile, in order to realize the prediction of leisure tourism preference, a combined model based on BP neural network and ARIMA is proposed to improve the accuracy of leisure tourism preference prediction by fully considering the linear and nonlinear laws of tourism statistics. The ARIMA-BP combination prediction model is applied to predict the leisure tourism preference in the future from 2027-2034. During the period 2027-2029, the number of leisure tourism tourists maintains a high annual growth rate of more than 15%, while the growth rate slows down after 2029, with an average annual growth rate of 4.44%. In 2033, the number of leisure tourism tourists will reach 1,691,280,000, and the leisure tourism preference of tourism users has been significantly strengthened.

Keywords: user profile; two-step clustering algorithm; BP neural network; ARIMA model; leisure tourism

1. Introduction

Multiple data show that after the epidemic, the number of global cross-border travelers in 2024 has gradually recovered to the pre-epidemic peak, and is even on the rise, with the number of cross-border travelers rising from 1.1 billion to 1.4 billion during the period of 2013-2024, and tourism revenues growing to 1.6 trillion U.S. dollars [1-3]. It is evident that the current global trend of cross-border tourism is promising. With the continuous development of tourism, cross-border tourism has become

✉ Corresponding author.
E-mail address: hszywdx@163.com (D. Wu).
Received 25 February 2024; Revised 10 May 2024; Accepted 15 December 2024; Published Online 16 April 2025.
DOI: 10.161091/jcmcc27b-324

more and more common and frequent, the study of the impact of culture on tourism behavior has received more and more attention, and scholars have begun to incorporate cultural factors into the study of tourism behavior as the key variables affecting tourists' preferences and behaviors [4-6]. As tourism, as a heterogeneous activity, will inevitably give rise to cultural exchanges between the destination and the source, tourists and hosts ultimately realize the cultural exchange and dissemination of differences across regions through mutual understanding and acceptance of each other's information [7-8]. As carriers and transmitters of culture, tourists carry the language, dress, behavior, thinking and values of their own location (i.e., the origin) to the destination, and use a series of tourism activities (e.g., residence, excursion, sightseeing, and cross-cultural contact, communication, and sociability at the destination) as a bridge to the cross-cultural transmission of tourists to the destination occurs in the collision of the two regional cultures [9-10]. A few scholars from the symbolic point of view, that the host and the guest in tourism activities use symbols to carry out cultural exchanges, symbols in tourism cross-cultural exchanges bear the basic function of expression and interpretation, due to the difference in the cultural background of the communicators, the way of expression and interpretation of the symbols and the significance of the symbols are very different, which have a significant impact on the exchanges and dissemination, and are also the source of the cultural barriers in the cross-cultural exchanges [11-13].

With the rapid development of tourism and the transformation of tourism consumption habits, tourism development is changing from the state of sightseeing tourism alone to the trend of coexistence of sightseeing, leisure, and vacation in various forms, to meet the demand for people's natural landscape, human history, leisure and entertainment integration of tourism experience [14-16]. Leisure tourism will become an important component to lead the development direction of the leisure era. By predicting the preference of leisure tourism, not only can we plan the tourism program in advance, but also give a reference to the marketing strategy and risk prevention planning of local tourism, etc [17-18]. Meanwhile, with the development of social media and the gradual expansion of the proportion of daily use, it has become an important tool for people to obtain and share information, as well as planning trips. Due to the convenience of social media and a wide range of user groups, it has had a profound impact on both cross-cultural and tourism, and its data is one of the main paths to understand tourists' travel preferences, as well as providing a path for cross-cultural communication and dissemination [19-21]. Specifically, the development of social media has changed the way people communicate and facilitated cross-cultural communication and the acquisition and dissemination of tourism information. It breaks the geographical limitations, so that people around the globe can communicate and share through various social platforms. Whether it is TikTok, Facebook, Twitter or Instagram, people can share their lives, insights and opinions with other users anytime, anywhere, and this instant and interactive communication can promote mutual understanding between different cultures, as well as news about related tourism events, so as to adjust the travel plan [22-25]. In addition, social media provide diverse and innovative ways of communication. Through multimedia forms such as text, pictures, and videos, people can more intuitively display the characteristics and uniqueness of their own culture and tourism [26]. At the same time, people can participate in multinational organizations or cross-cultural projects through social media platforms, work together with users around the world to solve problems and achieve cultural intermingling, and the analysis of the platform data for demand prediction makes the promotion and marketing of the tourism industry more targeted [27]. It can be seen that the value of social media data is extremely high, but the current social media for leisure tourism and cross-cultural analysis research is still in its infancy.

In this paper, through the fusion of offline mobile users and online tourism users' location keypoints, the TF-IDF method is used to process the textual information of online tourism users in this paper, and the LDA theme model is used to form a multi-dimensional theme of the label to add the fused location keypoint features, so that the user labeling weights have a higher degree of flexibility, and thus match the accurate labels for offline mobile users. Based on the similarity of users, the matching of user labels

is realized, and the social media tourism user profile is constructed. On the basis of tourism user profiles, the two-step clustering algorithm is introduced into the cluster analysis algorithm, the construction of the number of clusters when the data type is discrete variables is incorporated to improve the BIRCH hierarchical clustering algorithm, and the clusters generated by preclustering are clustered again using cohesive method, and the cross-cultural analysis of social media data is carried out from the two aspects of the attraction's featured theme and the user's tourism preference performance. Finally, a prediction method for the number of leisure tourism tourists based on the BP-ARIMA combination model is proposed. After describing the linear law of the data through the ARIMA model, the BP neural network is then used to predict the error of the ARIMA model and predict the nonlinear law, and then the prediction results of the two are combined to obtain the prediction value of the BP-ARIMA combination model, which realizes the prediction of the leisure tourism preference.

2. Social Media Tourism User Profile Construction

In recent years, with the rapid development of social media and the continuous increase of user location information in various types of central servers, the social media tourism user portrait has been refined by deeply integrating offline user location information and online tourism scenario information. In this chapter, a social media tourism user portrait model will be established to provide a structured data base for the subsequent cross-cultural analysis of social media data [28].

2.1. Modeling of user labels

Before generating user profiles, data modeling of various types of tourism user labels is first needed. In this paper, tourism user labels are divided into two categories: static basic labels and dynamic information labels.

Static base label: this label mainly includes the basic information of online tourism users and a series of structured information, which needs to be obtained explicitly. In this paper, the static base label is mainly obtained from the personal information in the author ID.

Dynamic information tags: most of the information in this type of tags is semi-structured information, and even part of the information is unstructured, which needs to be obtained implicitly. In fact, most of the dynamic information exists in the Internet behaviors of tourism users, such as browsing records, search records, forum interactions, and forum postings. In this paper, dynamic information labels are mainly obtained from online tourism users' forum posting behavior, i.e., labels are extracted from contextual information.

Improved information fusion-based tourism user portrait model: compared with the traditional user portrait model, this paper makes the user portrait more accurate by adding information fusion weights, and the improved formula is defined as:

$$veight(user_i, tag_j) = weight(action) \times weight(I) \times weight(location) \times \sum_{k=0}^{n} titlaS \times isFind + commentS \times \frac{count(tag_j)}{countAll} \tag{1}$$

1) $weight(action)$, behavioral weight, tourism users through the Internet generated network behavior information, such as posting, searching, browsing and interaction, etc., for different kinds of network behavior, its behavioral weight is not the same. We believe that the behavioral weight of tourism users' posting in forums is much higher than their browsing behavioral weight. Therefore, in this paper, we choose to crawl the forum posting records of online travel users.

2) $weight(location)$, the information fusion weight, after the fusion of the user's location key point features to assign weight, the weight value will be brought to the “topic-word” matrix, and ultimately get the information fusion-based “topic-word” matrix. matrix based on information fusion.

3) $titleS \times isFind + commentS \times \frac{count(tag_j)}{countAll}$, keyword position, this paper assigns weights to labels based on keyword position. In text preprocessing, since the keywords in the title of the travelogue are much more important than the keywords contained in the content of the travelogue, this paper assigns a higher weight to the title of the travelogue and a lesser weight to the content part of the travelogue.

4) $weight(I)$, the attenuation factor, this paper according to the travel user from the last behavior of the time interval to determine whether the label is accurate, we believe that the travel user's current behavior and the last behavior of the shorter the time interval, then the label accuracy is higher, the higher the weight of its behavior. It is because the user's preference habits will gradually decay over time, so this paper introduces the attenuation factor, the formula is as follows:

$$I_{weight} = \frac{1}{1+\log_c(t+1)} \tag{2}$$

Where t represents the time interval between the current behavior and the last behavior.

Combining the above elements, this paper represents the information fusion-based travel user portrait model as: label weight = behavior weight × keyword location weight × attenuation factor × information fusion weight.

2.2. Modeling the content representation of travelogues

In this paper, a vector space model is used to represent the content of the travelogue, which is not only in line with the characteristics of the keywords of the user profile in this paper, but also able to deal with the data with complex structure and large amount of data. In this paper, we use $D(t_1, w_1; t_2, w_2; ...t_n, w_n)$ represents a vector of traveler content representations, where t_n represents the keywords extracted from the traveler content, and each keyword represents a dimension. w_n represents the weight of the keywords, which is convenient to reflect the correlation between the keywords and the content of the travelogues. The keyword position factor is also taken into account, so the weights assigned to the travelogue title and the travelogue content are different, which means that the value of the keyword in the travelogue title is higher than that of the travelogue content.

2.3. TF-IDF feature extraction

By comparing a variety of feature extraction methods, it is found that the TF-IDF method is more suitable for processing the textual information of online travel users in this paper, and its advantage lies in the high universality and less difficulty, and it can intuitively reflect the distribution location and importance of keywords, and its formula is as follows [29]:

$$w_{ik} = tf_{ik} \times idf_k \tag{3}$$

Where w_{ik} represents the weight value of the keywords in the travelogues, f_{ik} represents the number of times the keywords appear in the different travelogues, and idf_k represents the inverse document frequency of the keywords in all the travelogues, which is given in the following formula:

$$idf_k = \log(\frac{N}{n_k}) \tag{4}$$

In order to eliminate the interference caused by the length of text information of online travel users to the keyword weights, we choose to normalize the method, from which we get the final keyword weights with the following formula:

$$w_{ik} = \frac{tf_{ik} \times \log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{k=1}^{n}\left[f_{ik} \times \log\left(\frac{N}{n_k}\right)\right]}} \tag{5}$$

Meanwhile, to visualize the importance of keywords in the travelogue title, we multiply the keyword weights extracted from the travelogue title by the parameter θ.

2.4. LDA Topic Modeling to Train Tour Information

In this paper, the LDA theme model is used to form a multi-dimensional theme label, which can not only comprehensively and accurately summarize the theme characteristics of the travelogue information, but also do a good job for the establishment of the user profile labeling system as a foundation [30]. The trained LDA topic model converts the "document-word" matrix formed after preprocessing word splitting into an effective "topic-word" matrix after several iterations. The training process is as follows:

1) Assign the topic z_0 randomly to each word in the travelogue, and calculate the number of word items t and documents m appearing in topic z.

2) Calculate the probability distribution $p(z_i \mid z_{-i}, d, w)$ of the assignments to the different topics

3) Assign new topics z_1 based on the probability distribution.

4) Loop the above process until the word distribution parameters in each topic and topic distribution parameters in each document are output.

2.5. Matching user labels

Matching labels for offline mobile users can not only accurately mine the preference characteristics of offline mobile users, but also accurately and efficiently retrieve tourism resources that meet the preferences and needs of such users, thus realizing the recommendation of tourism resources. The specific steps of the algorithm are as follows:

1) Represent the topic corpus and keywords with text vectors.

2) Calculate the similarity between the training text and the text to be classified.

3) Arrange the similarity in descending order to get the first K topic corpus with the highest similarity and match them to online tourism users.

4) Calculate the user similarity and finally match the labels to the offline mobile users who are similar to the online travel users:

$$sim(v_{corpus}, v_{text}) = \frac{\sum_{k=1}^{M} v_{ik} \cdot v_{jk}}{\sqrt{\sum_{k=1}^{M} v_{jk}^2} \cdot \sqrt{\sum_{k=1}^{M} v_{ik}^2}} \tag{6}$$

$$UserSim(LocTra_1, LocTra_2) = \frac{\sum_{j=1}^{m} SeqSim(seq[j])}{|LocTra_1| \times |LocTra_2|} \tag{7}$$

Due to the diversity of information and the huge amount of data on social networking platforms, in order to more accurately and comprehensively obtain the user's preferences and habits, this paper adds the fused location keypoint features to assign weights to offline mobile user labels. Traditional user profiling models generally consider the type of user behavior, keyword location, and time decay and other factors, but do not take into account the user's location key point factors. In this paper, we propose an information fusion-based travel user portrait model that simultaneously considers multi-

dimensional features such as behavior, location, and time to label offline mobile users more accurately and without error.

3. Cross-cultural analysis of social media data based on a two-step clustering algorithm

In this chapter, we will start from the perspective of leisure tourism, based on the social media tourism user portrait model constructed in this paper, and use two-step clustering algorithm to conduct cross-cultural analysis of social media data, to explore and excavate the characteristic themes of the attractions under the cross-cultural ability of social media and the performance of the users' leisure tourism preferences.

3.1. Principles of two-step cluster analysis

3.1.1. Cluster analysis. Cluster analysis algorithms have been studied for more than half a century so far, and various cluster algorithms have been proposed, developed, and evolved to enrich the cluster analysis system. The most widely used cluster analysis methods are K-Means, fuzzy clustering, BIRCH, density clustering, grid-based typical algorithm STING algorithm and so on. Data attributes in cluster analysis can be continuous, discrete or mixed, most cluster analysis algorithms are for continuous variables to calculate the distance between samples, continuous variables to calculate the distance between samples commonly used methods are: the Euclidean distance, Manhattan distance, Chebyshev distance.

If x_i and x_j are n-dimensional feature vectors:

Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{8}$$

Manhattan Distance:

$$d(x_i, x_j) = \sum_{k=1}^{n} | x_{ik} - x_{jk} | \tag{9}$$

Chebyshev distance:

$$d(x_i, x_j) = \max(| x_{i1} - x_{j1} |, \cdots, | x_{in} - x_{jn} |) \tag{10}$$

Traditional K-means clustering, hierarchical clustering and density clustering are not suitable for clustering data with mixed attributes, and the defects of K-means clustering is that it is greatly affected by noise and outliers, the selection of merging or splitting points in hierarchical clustering is more difficult, and density clustering needs to determine the parameters, and the parameters have a very significant effect on the results.

The following introduces an improved cluster analysis method based on BIRCH hierarchical clustering-two-step clustering. The two-step clustering algorithm completes the cluster analysis through two stages: preclustering and clustering, and the characteristics of two-step clustering are as follows.

1) It can be applied to the clustering of mixed-attribute data, i.e., it deals with the clustering of discrete and continuous variables at the same time;

2) Determines the number of clusters automatically;

3) Efficiently handling massive data sets.

3.1.2. Principles and steps of two-step cluster analysis. Two-step clustering is an improved algorithm for hierarchical clustering of BIRCH (Balanced Iterative Reduced Clustering), as the name

suggests it is divided into two steps, the first step is preclustering, which adopts the idea of CF tree growth in the BIRCH algorithm, and the growth of the CF tree is the process of data point insertion, which determines whether to insert the sample or not by comparing the value of the parameter one by one with the number of nodes and the radius of the sphere after the insertion of the data point to complete the dense region of data point clustering, forming many small sub-clusters [31-32]. The second step is the clustering stage, taking the results of the preclustering stage-subclusters as objects, using the log-likelihood distance measure to measure the distance of the preclustered subclusters, and selecting the cluster with the smallest distance to be merged into one cluster until the desired number of clusters.

1) Pre-clustering-improved BIRCH. The preclustering stage is the improved BIRCH clustering and also the process of clustering feature tree generation. Improved BIRCH utilizes a tree structure to achieve fast clustering, which is generally referred to as Clustering Feature Tree (CFTree).The clustering feature tree consists of root nodes, branch nodes and leaf nodes, node 1 is the root node of this CFTree, node 2 and node 3 are the branch nodes, and node 4, node 5 and node 6 are the leaf nodes. In CFTree, the nodes are composed of clustering features (CF), the clustering features of the internal nodes have pointers to the child nodes, and all the leaf nodes are linked with a bi-directional chain table.

CFTree is formed gradually by adding and updating the entries during traversing the dataset and by means of splitting the nodes, CF tree is a contour tree with three parameters: branch balancing factor B, leaf balancing factor L, and spatial threshold T; branch balancing factor B indicates that the number of clustering features of the internal nodes does not exceed B at the most, leaf balancing factor L indicates that the number of clustering features of the leaf nodes does not exceed L at the most, and the spatial threshold is the maximum sample radius threshold T for each CF in a leaf node, which indicates that all sample points in each clustered feature should be within a hypersphere with radius not exceeding T . The middle CFTree is set B = 4 and L = 3, indicating that there are at most 4 CFs in the internal nodes and at most 3 CFs in the leaf nodes.

(1) Clustering feature $\overrightarrow{CF}$ introduction. $\overrightarrow{CF}$ is the clustering feature of each node meta-item in BIRCH clustering, in BIRCH clustering define the clustering feature $\overrightarrow{CF}$ as a 3-tuple, $\overrightarrow{CF} =< N, \vec{\Sigma}_i, \vec{\Sigma}_j >$, N is the amount of data for the data points in the cluster, $\vec{\Sigma}_i$ is the linear summation of the data points, and $\vec{\Sigma}_j$ is the sum of squares of the data points. BIRCH only performs cluster analysis for numerical data, the improved BIRCH can analyze not only numerical data but also mixed-attribute data, so the definition of the clustering feature $\overrightarrow{CF}$ of the improved BIRCH is different from that of the original BIRCH definition, and the clustering feature of the improved BIRCH $\overrightarrow{CF}$ is defined as a quaternion, let a cluster C_j cluster feature $\overrightarrow{CF}_j =< N_j, \vec{\Lambda}_i, \vec{\Sigma}_i, \ddot{\vec{N}}_j >$, where:

$$\vec{\Lambda}_i = \left(\sum_{k=1}^{v} \overline{x}_{nk}, \cdots, \sum_{k=1}^{m} \overline{x}_{mk} \right)^T \tag{11}$$

$\vec{\tilde{\Lambda}}_i$ denotes the linear summation of the values of the attributes of the data points in the cluster C_j under each continuous type attribute:

$$\vec{\tilde{\Sigma}}_j = \left(\sum_{n=1}^{N_j} \vec{x}_{jn1}^2, \cdots, \sum_{n=1}^{N_j} \vec{x}_{jnD_j}^2 \right)^T \tag{12}$$

$\vec{\tilde{\Sigma}}_j$ denotes the sum of squares of the attribute values of the data points in the cluster C_j under the continuous attribute:

$$\ddot{\vec{N}}_j = ((\ddot{\vec{N}}_j)^T, \cdots, (\ddot{\vec{N}}_{jD_2})^T)^T \tag{13}$$

$\ddot{\vec{N}}_j$ denotes the measure of the cluster C_j under the categorical attribute. $\ddot{\vec{N}}_j = (\ddot{N}_{jt1}, \cdots, \ddot{N}_{j,t,x_j})^T$ is the vector formed by the number of data points of each possible value of the cluster C_j under t th

categorical attribute and $\ddot{N}_{jik}(k=1,\cdots,\varepsilon_t-1)$ denotes the number of data points for the k th value of the cluster C_j under the t th sub-type attribute.

The clustering feature $\overrightarrow{CF}$ has additivity, if the clustering features of cluster C_i and cluster C_j are $\overrightarrow{CF}_i$ and $\overrightarrow{CF}_j$, respectively, and according to the clustering feature additivity, there:

$$\overrightarrow{CF}_{(i,j)} =< N_i + N_j, \vec{\Lambda}_i + \vec{\Lambda}_j, \vec{\vec{Z}}_i + \vec{\vec{Z}}_j, \vec{\vec{N}}_i + \vec{\vec{N}}_j > \tag{14}$$

(2) Generation of clustering feature tree CFTree. Insertion process of clustering features: first take the root node as the starting point, calculate the distance between the new sample and the leaf node, search for the leaf node with the closest distance and the CF node with the closest distance in that leaf node, and if after inserting the new sample into this CF node, the radius of hypersphere corresponding to the CF node still meets the threshold T less than the set radius of the hypersphere, insert the new sample point into the CF node, and complete the new sample point insertion; if the hypersphere radius after inserting the new sample does not satisfy the set threshold T, and the number of CFs in the leaf node satisfies the threshold L which is less than the set CF number, put the new sample into the newly created CF node, update the CF and complete the insertion of the new sample; if the number of CFs in the leaf node does not satisfy the threshold L which is less than the CF number of the child node, divide the current leaf node into two, put the new sample point into the new child node and update the CF quaternion to complete the insertion.

2) Clustering stage. The clustering phase is a second clustering of the subclusters of the leaf meta-items of the preclustered generated CF tree, the subclusters are denoted as $C_1, C_2, \cdots, C_k$. The second clustering uses the coalescent method, where the core idea is to recursively merge the closest clusters until the optimal number of clusters is satisfied. The coalescent method uses log-likelihood distance to compute the distance between subclusters of the leaf element term. First, the log-likelihood distance between two of the K clusters $C_1, C_2, \cdots, C_k$ is computed; the log-likelihood distance between clusters is defined as follows:

$$d(C_i, C_j) = \zeta_i + \zeta_j - \zeta_{i,j} \tag{15}$$

Among them:

$$\zeta_i = -N_i(\frac{1}{2}\sum_{t=1}^{n}\ln(\hat{\delta}_{t,i}^2 + \hat{\lambda}_t^2) + \sum_{t=1}^{n}\hat{E}_u) \tag{16}$$

N_i denotes the data object personalities in cluster C_i:

$$\hat{\delta}_{t,i}^2 = \frac{\overline{\Sigma}_i}{N_i} - (\frac{\overline{\lambda}_t}{N_i})^2 \tag{17}$$

The $\delta_{t,i}^2$ denotes the variance under the s th continuous type attribute estimated from the data points in the cluster C_i:

$$\delta_{t,i}^2 = 1/N\sum_{x=1}^{X}(\overline{x}_{m,x} - \overline{x}_{i,x})^2 \tag{18}$$

$\hat{\delta}_s^2$ denotes the variance under the s th continuous-type attribute estimated from all data points in the data set:

$$\hat{E}_u = -\sum_{k=1}^{s_i}(\frac{\hat{N}_{ik}}{N_i}\ln\frac{\ddot{N}_{ik}}{N_i}) \tag{19}$$

$\hat{E}_n$ denotes the information entropy under the tth discrete attribute in the cluster c_i. Therefore, the log-likelihood distance between clusters can be computed using the log-likelihood distance by knowing only the clustering characteristics of the clustering resultant clusters.

Calculate and merge the two clusters with the smallest distance using the above log-likelihood distance; then continue to calculate the two-by-two log-likelihood distances for the remaining ($K-1$) clusters and merge the two clusters with the smallest distance. Until the number of clusters is the desired number. At this point, the two-step clustering is complete.

3.2. Thematic mining of attraction features based on online reviews

In this section, the mass review information of selected attractions in a travel social platform is extracted as the research data, and the text preprocessing is carried out firstly to filter the meaningless as well as repetitive data in a large number of review information, and then the keywords are extracted by using the improved TF-IDF algorithm to extract the nouns, verbs, adjectives, and gerunds as the featured keywords as the basis of the attraction featured theme analysis.

The classification of online review keywords is shown in Table 1. The filtered feature keywords evaluate the attractions from different dimensions, which helps to comprehensively understand the attraction characteristics, and finally categorize the feature keywords one by one in 11 attraction feature theme dimensions, such as human history, recreational activities, etc., and assign weights to the themes of different attractions using the improved TF-IDF algorithm.

Table 1. Classification of keywords in online reviews

Feature theme of scenic spots	Words
Human history	Heavy history, ancient, culture and so on.
Entertainment activities	Lighting performances, Hanfu, lanterns, stage plays, etc.
Tour guide service	Tour guide, explanation, business proficiency, enthusiasm, courtesy, etc.
Natural landscape	Beautiful environment, spectacular, pleasant scenery, majestic, cherry blossom and so on.
Economical benefits	Free, reasonable price, cost-effective, affordable, cheap, etc.
Restaurant cuisine	Night market, food, special snacks, pasta, etc.
Convenient transportation	Convenient transportation, subway, city center, tour bus, etc.
Architectural features	Architecture, brick and wood structure, structural design, garden style, etc.
Scenic services	Indications, landscaping, parking, public facilities, sliding rope, etc.
Scenic protection	Cultural relics protection, complete preservation, maintenance, environmental protection vehicles, etc.
Landscape impression	Experience good, beautiful, great, praise, etc.

3.3. Analysis of tourists' leisure travel preferences based on user profiles

3.3.1. User profiling results and their analysis. This section analyzes the user profile from four aspects: tourists' basic information, travel characteristics, cognitive level, and trajectory information, as shown in Table 2. Among them, the basic information of tourists includes gender, age and education, the travel characteristics include fellow travelers and travel expenses, the degree of cognition includes the cognitive level and whether they are engaged in tourism-related occupations, and the trajectory information includes the tourists' stay attractions and their stay time. In summary, the seven variables of basic information, travel characteristics and cognitive degree are used to portrait tourists through second-order clustering.

Table 2. User portrait related information description

User portrait information	Attributes	Description
Basic information	Gender	Categorical variables,1= male,2= female
	Age	Rank variable,1=under18 years old,2=18-25 years old,3=5-35 years old,4=35-60 years old,5=over 60 years old.
	Education	Rank variable,1 =junior high school and below, 2=high school and technical secondary school,3=undergraduate and junior college,4=graduate student.
Travel characteristics	Colleagues	Categorical variables,1=friends,2=family,3=parents, 4 = couples,5=a person
	Travel costs	Numerical variables, the average daily travel cost of tourists
Cognitive level	Cognitive level	Grade variables,1 = very low,2=low, 3=general,4 = high, 5 = very high
	Profession	Categorical variables, 0 = No, 1 = Yes
Trajectory information	Attractions to stay	Nominal variables, scenic spots where tourists stay
	Duration of stay	Numerical variables, the stay time of tourists at a scenic spot

The different number of cluster classes affects the accuracy of the clustering model, this section selects the number of clusters from 1 to 15, in using the BIC information criterion, the amount of change in BIC, the ratio of change in BIC, and the ratio of distance measurements to determine the optimal number of cluster clusters, as shown in Table 3. When the number of clusters is selected as 3, the comprehensive evaluation effect of its various indicators is optimal, and the distance measurement ratio is 1.565, which is the highest, and other evaluation indicators also show more advantages, therefore, 3 is selected as the optimal number of clusters, which lays the foundation for the subsequent user profiling.

Table 3. The change of cluster number

Number of clusters	BIC	BIC variation a	BIC change ratio	Distance measurement ratio
1	17389.38	-	-	-
2	15563.39	-1835.79	1	1.436
3	14327.56	-1240.01	0.668	1.565
4	13569.34	-745.878	0.397	1.307
5	13034.18	-533.785	0.292	1.323
6	12661.34	-384.577	0.208	1.178
7	12373.21	-303.621	0.164	1.043
8	12079.85	-279.429	0.146	1.115
9	11851.47	-243.909	0.127	1.064
10	11628.4	-216.936	0.109	1.208
11	11494.28	-158.121	0.074	1.019
12	11337.68	-142.686	0.073	1.132
13	11229.07	-107.68	0.061	1.05
14	11124.92	-108.455	0.052	1.161
15	11054.8	-63.033	0.039	1.029

The results of visitor clustering are shown in Fig. 1. It can be seen that under the premise that the optimal number of clusters is 3, the share of the three categories is more evenly distributed. Tourists are roughly divided into three categories.

Category 1: The main characteristics of this group of people are that they spend an average amount of money on traveling, are mostly young adults, are not engaged in occupations related to the tourism industry, have a low education, have a medium-low cognitive level, and are mostly traveling in a family group.

Category 2: The main characteristics of this group of people are that they spend more on traveling, are mostly young and middle-aged, have a higher percentage of people engaged in tourism-related occupations, have higher education and cognitive level, and mostly travel with friends or alone.

Category 3: The main characteristics of this group are lower travel expenses, mostly young people aged 18-25, mostly traveling with friends or couples, average cognitive level, and mostly with undergraduate (college) education.

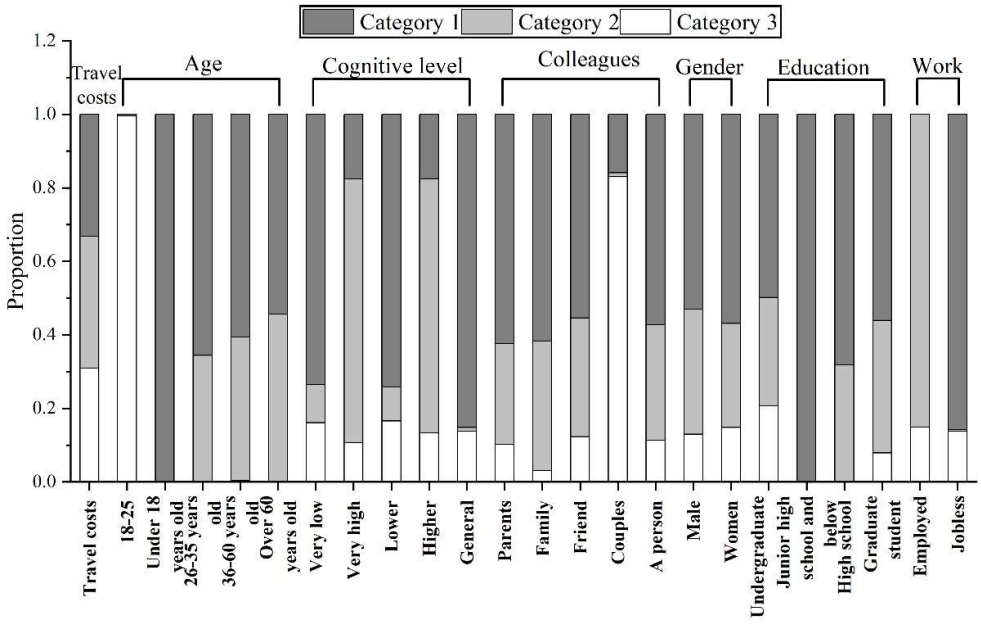


Fig. 1. Tourists clustering results analysis statistical chart

3.3.2. Results of Leisure Tourism Preference Analysis. In order to better analyze the differences in the leisure tourism preferences of different groups of people, this section adopts the multivariate Logistic regression model as the research model, selecting the category of tourists as the dependent variable, and the 11 dimensions of attraction characteristics theme humanistic history, recreational activities, tour guide service, natural landscape, affordability, catering and food, convenient transportation, architectural features, scenic area service, landscape impression, scenic area protection as the independent variable , of which the category of tourists is a categorical variable and the attraction characteristics related variables are numerical variables, as shown in Table 4.

Table 4. Data variable description

Variables	Feature theme of scenic spots	Description
Dependent variable	Tourist category	Classification variable, Type = {1, 2, 3}
Independent variable	Human history	Numerical variables
	Entertainment activities	Numerical variables
	Tour guide service	Numerical variables
	Natural landscape	Numerical variables
	Economical benefits	Numerical variables
	Restaurant cuisine	Numerical variables
	Convenient transportation	Numerical variables
	Architectural features	Numerical variables
	Scenic services	Numerical variables
	Scenic protection	Numerical variables
	Landscape impression	Numerical variables

The initial model fitting information is specifically shown in Table 5. The log-likelihood value of the model is 2135.011, with a p-value of less than 0.05, which means that the model passes the test in general, indicating that the multivariate Logistic regression model can be used to analyze and explain the relationship between visitor categories and attraction feature theme preferences.

Table 5. Model fitting information table

Model	Model fitting condition logarithmic likelihood value	chi-square	Likelihood ratio test freedom	Significance
Intercept only	2528.454	-	-	-
Finally	2135.011	392.541	22	0.000

The initial model likelihood ratio test results are shown in Table 6. The P-values of the three indicators of transportation convenience, scenic services and landscape impression are 0.498, 0.112 and 0.051, respectively, and their P-values are all greater than 0.05, and there is no significant difference for the three groups of tourists' leisure tourism preference, while the P-values of the three indicators in terms of humanities and history, recreational activities, tour guide services, architectural features, scenic area preservation, affordability, natural landscape and catering and food are all less than 0.05, and there is a significant difference in the tourists' leisure tourism preference. There is a significant difference in tourists' leisure tourism preference, therefore, in this section, the three indicators of transportation convenience, scenic area service and landscape impression are eliminated, and then the initial model is optimized.

Table 6. Likelihood ratio test of initial model

Effect	The logarithmic likelihood value of the simplified model	Chi-square	Degree of freedom	Significance
Intercept	2555.99	421.966	2	0.000
Convenient transportation	2135.395	1.376	2	0.498
Human history	2141.345	7.31	2	0.022
Entertainment activities	2155.594	21.562	2	0.000
Tour guide service	2167.079	33.033	2	0.000
Architectural features	2183.396	49.368	2	0.000
Scenic protection	2175.023	41	2	0.000
Scenic services	2138.224	4.182	2	0.112
Landscape impression	2139.825	5.791	2	0.051
Economical benefits	2164.829	30.79	2	0.000
Natural landscape	2156.955	22.929	2	0.000
Restaurant cuisine	2202.727	68.694	2	0.000

The final fitted model was formed by eliminating the three variables that were not statistically different, and the results of the likelihood ratio test of the final fitted model are shown in Table 7. The P-values of the remaining variables humanistic history, recreational activities, tour guide services, architectural features, scenic area protection, affordability, natural landscape, and food and beverage are all less than 0.05, indicating that there are significant differences in tourists' preferences for these eight attraction feature themes. Therefore, this section analyzes the leisure tourism preference of tourist groups based on the eight attraction characteristic themes of human history, recreational activities, tour guide service, architectural features, scenic area protection, affordability, natural landscape and food and beverage.

Table 7. The final model likelihood ratio test

Effect	Model fitting conditions simplify the logarithmic likelihood value of the model.	chi-square	Likelihood ratio test freedom	Significance
Intercept	2600.221	446.131	2	0.000
Human history	2167.905	13.798	2	0.001
Entertainment activities	2174.823	20.721	2	0.000
Tour guide service	2185.475	31.37	2	0.000
Architectural features	2224.146	70.051	2	0.000
Scenic protection	2187.24	33.153	2	0.000
Economical benefits	2197.812	43.706	2	0.000
Natural landscape	2178.734	24.628	2	0.000
Restaurant cuisine	2279.493	125.375	2	0.000

In order to better illustrate the differences in leisure tourism preferences among different tourists, this section constructs regression models I, II and III with category I, category II and category III as reference objects respectively, and their regression results are shown in Table 8. The regression analysis results will be somewhat different with different reference groups. When the value of coefficient B is less than 0, it means that the variable has a reverse effect on this model; when the value of coefficient B is greater than 0, it means that the variable has a positive effect on this model.

1) Analysis of the differences in leisure tourism preferences of category 1 tourist groups. Relative to the category III tourist group, there are differences between the category I tourist group and its preference for human history, recreational activities, tour guide service, architectural features, affordability, natural landscape and catering and food (Sig=0.003<0.05), of which the coefficient B values of human history and recreational activities are 0.793 and 2.896 respectively, which are greater than 0, indicating that the category I tourists prefer rich in human history, recreational activities and food to those of the category III tourists. Tourists prefer tourist attractions with rich humanistic history and diverse recreational activities. The coefficient B of affordability is greater than 0 (B=2.237>0), indicating that Category 1 tourists prefer affordable attractions compared to Category 2 tourists.

2) Analysis of differences in leisure tourism preferences of category II tourist groups. Relative to the category III tourist group, the coefficient B values of humanistic history, recreational activities, scenic spot protection and affordability for the category II tourist group are 0.95, 3.48, 6.86 and 4.013 respectively, all of which are greater than 0, indicating that the category II tourists prefer strong humanistic history, rich recreational activities, varied catering food and affordable tourist attractions compared with the category III tourists. And relative to the category one tourist group, there is a significant difference between category two tourists and their architectural features, scenic services, affordable and food and beverage cuisine (Sig=0.003<0.05). By observing the value of coefficient B, it is found that compared with category I, category II group prefers attractions with distinctive architectural features (B=1.86>0) and varied catering cuisine (B=1.065>0).

3) Differences in leisure tourism preferences of category III visitor groups. Relative to the Category I group, the coefficient B values of guide service, architectural features, natural landscape, and catering cuisine for the Category III group were 5.39, 7.567, 4.925, and 1.574 respectively, which are all greater than 0, indicating that the Category III tourist group prefers good guide service, distinctive architectural features, rich catering cuisine, and natural landscape-type attractions compared to the

Category I group. The category three group was found to prefer good tour guide service (B=4.594>0), distinctive architectural features (B=6.652>0), abundant food and beverage (B=0.498>0), and natural landscape type (B=3.032>0) attractions compared to the category two group.

Table 8. Model parameter estimates

Model	Category	Type	Coefficient B	Standard error	Wald	Degree of freedom	Significance
Model 1	2	Intercept	3.534	0.263	166.864	1	0
		Human history	0.422	0.221	3.837	1	0.05
		Entertainment activities	1.106	0.609	3.341	1	0.068
		Tour guide service	1.024	0.767	1.802	1	0.18
		Architectural features	1.86	0.637	8.749	1	0.003
		Scenic services	4.269	1.793	5.601	1	0.018
		Natural landscape	1.722	0.921	3.538	1	0.06
		Economical benefits	2.924	0.673	19.202	1	0
		Restaurant cuisine	1.065	0.166	40.487	1	0
		Intercept	-6.324	0.463	188.736	1	0
	3	Human history	-0.476	0.297	2.6	1	0.107
		Entertainment activities	-2.603	0.783	10.915	1	0.001
		Tour guide service	5.39	0.968	30.487	1	0
		Architectural features	7.567	1.01	55.752	1	0
		Scenic services	-4.856	2.122	5.27	1	0.022
		Natural landscape	4.925	1.039	22.175	1	0
		Economical benefits	-6.53	0.995	42.485	1	0
		Restaurant cuisine	1.574	0.219	53.86	1	0
Model 2	1	Intercept	4.047	0.294	185.944	1	0
		Human history	-0.149	0.188	0.576	1	0.451
		Entertainment activities	-0.566	0.618	0.851	1	0.354
		Tour guide service	-0.169	0.675	0.07	1	0.795
		Architectural features	-1.189	0.645	3.304	1	0.069
		Scenic protection	-7.817	1.431	29.44	1	0
		Economical benefits	2.237	0.658	11.111	1	0.001
		Natural landscape	-0.204	0.553	0.132	1	0.725
		Restaurant cuisine	-0.706	0.093	58.18	1	0

					6		
	3	Intercept	-2.372	0.485	24.251	1	0
		Human history	-0.956	0.273	12.84	1	0
		Entertainment activities	-3.477	0.789	19.248	1	0
		Tour guide service	4.594	0.929	24.454	1	0
		Architectural features	6.652	1.068	39.198	1	0
		Scenic protection	-6.858	1.911	12.789	1	0
		Economical benefits	-4.012	1.028	15.381	1	0
		Natural landscape	3.032	0.708	18.854	1	0
		Restaurant cuisine	0.498	0.136	13.608	1	0
Model 3	1	Intercept	6.434	0.475	188.845	1	0
		Human history	0.793	0.259	9.025	1	0.003
		Entertainment activities	2.896	0.791	13.746	1	0
		Tour guide service	-4.775	0.924	26.588	1	0
		Architectural features	-7.835	1.048	56.472	1	0
		Scenic protection	-0.962	1.776	0.284	1	0.589
		Economical benefits	6.244	1.006	37.768	1	0
		Natural landscape	-3.227	0.7	21.095	1	0
		Restaurant cuisine	-1.209	0.135	81.579	1	0
	2	Intercept	2.385	0.477	24.248	1	0
		Human history	0.95	0.264	12.85	1	0
		Entertainment activities	3.48	0.795	19.244	1	0
		Tour guide service	-4.596	0.937	24.461	1	0
		Architectural features	-6.654	1.07	39.2	1	0
		Scenic protection	6.86	1.916	12.801	1	0
		Economical benefits	4.013	1.033	15.389	1	0
		Natural landscape	-3.042	0.692	18.84	1	0

					5		
		Restaurant cuisine	-0.488	0.136	13.61	1	0

4. Prediction of Leisure Travel Preference Based on BP-ARIMA Combined Modeling

In the previous section, this paper carried out a cross-cultural analysis of social media data using a two-step clustering algorithm to mine and explore users' leisure travel preferences from the perspective of leisure travel. In this chapter, a leisure travel preference prediction method based on a combination of BP neural network and ARIMA model will be proposed to further predict users' leisure travel preferences.

4.1. Combined BP-ARIMA modeling

4.1.1. Principles of BP Neural Networks. BP neural network is a multi-layer feed-forward neural network trained based on the error back propagation algorithm, which can fit the complex non-linear relationship between input variables and output variables, and has been widely used in the field of data analysis and prediction. BP neural network consists of an input layer, a hidden layer and an output layer. The input layer receives the input data, the hidden layer maps the input to the high-dimensional feature space through a series of nonlinear transformations, and the output layer bulls into the final prediction results, and if the prediction value deviates too much from the actual value, then back-propagation is carried out, and the gradient descent method is utilized to adjust the node weights and thresholds, and the operation is terminated when the error or the number of iterations meets the requirements.

Set the BP neural network input and output as $X_i(i=1,2,\cdots,n), Y_k(k=1,2,\cdots,m)$, and n,m are the number of input neurons, and output neurons respectively.

The hidden layer neuron output component is:

$$H_j = f\left(\sum_{i=1}^{n} \omega_{i.j} X_i + \alpha_j\right) \tag{20}$$

Where: H_j is the output value of the j th neuron in the hidden layer: $\omega_{i,j}$ is the network connection weight between the input layer neuron and the hidden layer neuron: α_j is the threshold value of the j th neuron in the hidden layer: f is the transfer function of the hidden neuron.

The output layer neuron output component is:

$$Y_k = g\left(\sum_{j=1}^{l} \omega_{j,k} H_j + \beta_k\right) \tag{21}$$

Where: Y_k is the output value of the k th neuron in the output layer; $\omega_{j,k}$ is the network connection weight between the hidden layer neuron and the output layer neuron: β_k is the min value of the k th neuron in the output layer; l is the number of neurons in the hidden layer; and g is the transfer function for the output layer neuron. .

The neural network undergoes forward transfer to obtain an overall error:

$$e = \frac{1}{2}\sum_{k=1}^{m}(Y_k - Z_k)^2 \tag{22}$$

Where: e is the total error: Z_k is the desired output of the k th neuron.

Based on the gradient descent method, the correction relationship between weights and thresholds is as follows:

$$\Delta\omega_{i.j} = -\eta \frac{\partial e}{\partial \omega_{i.j}} \tag{23}$$

$$\Delta\omega_{j.k} = -\eta \frac{\partial e}{\partial \omega_{j.k}} \tag{24}$$

$$\Delta\alpha_j = -\eta \frac{\partial e}{\partial \alpha_j} \tag{25}$$

$$\Delta\beta_k = -\eta \frac{\partial e}{\partial \beta_k} \tag{26}$$

Where: η is the neural network learning rate; $\Delta\omega_{i,j}, \Delta\omega_{j,k}$ are both weight correction coefficients: $\Delta\alpha_j, \Delta\beta_k$ are both threshold correction coefficients.

4.1.2. Principles of ARIMA modeling. ARIMA model is a commonly used model for time series analysis and forecasting. ARIMA model transforms non-stationary series into stationary series through differencing, then builds autoregressive and moving semi-average models on the semi-stationary series, and predicts future values by regressing the past values against the random error term and establishing a linear relationship in the time series. ARIMA model consists of the following 3 parts.

1) AR (Autoregressive) model: describes the relationship between current and historical values and predicts future values based on past values:

$$Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \cdots + \lambda_p Y_{t-p} + \varepsilon_t \tag{27}$$

Where: Y_t is the value of the dependent variable at moment t; $\lambda_0, \lambda_1, \lambda_2, \lambda_2, \cdots, \lambda_p$ is the parameter of the AR model; ρ is the order of the AR model; and ε_t is the white noise sequence.

2) MA (Moving Average) model: the moving average process is used as a complement to the autoregressive process to predict past values based on the errors of previous predictions:

$$Y_t = \mu_0 + \mu_1 \varepsilon_{t-1} + \mu_2 \varepsilon_{t-2} + \cdots + \mu_q \varepsilon_{t-q} + \varepsilon_t \tag{28}$$

Where: $\mu_0, \mu_1, \mu_2, \cdots, \mu_q$ are the parameters of the MA model; $\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots \varepsilon_{t-q}$ is the prediction error at the corresponding moment: q is the order of the MA model.

3) I (difference method): for non-stationary series, it is necessary to transform them into stationary series by difference in order to eliminate the effect of trend and seasonality in the time series. Where d represents the number of times of differencing, that is, after d times of differencing can be transformed into a non-smooth series into a smooth series.

In summary, let $\psi_0 = \lambda_0 + \mu_0$, then the $ARIMA(p,d,q)$ model can be expressed as:

$$\begin{aligned} Y_t = {} & \psi_0 + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \cdots + \lambda_p Y_{t-p} \\ & + \mu_1 \varepsilon_{t-1} + \mu_2 \varepsilon_{t-2} + \cdots + \mu_q \varepsilon_{t-q} + \varepsilon_t \end{aligned} \tag{29}$$

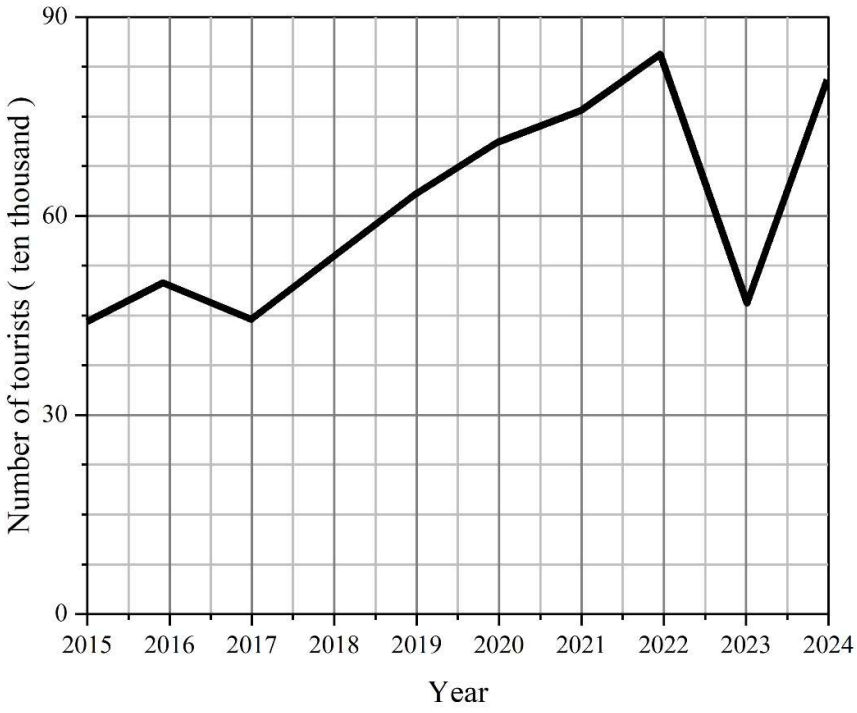
4.1.3. Combinatorial modeling. The idea of the combined prediction model is to use ARIMA model to predict first, so that the linear law of the data is described by the ARIMA model and the nonlinear law is included in the prediction error, and then the BP neural network is used to predict the error of the ARIMA model and predict the nonlinear law. Finally, the prediction result of ARIMA is added with the prediction result of BP neural network to get the prediction value of the combined prediction model.

4.2. Leisure Tourism Preference Forecast Analysis

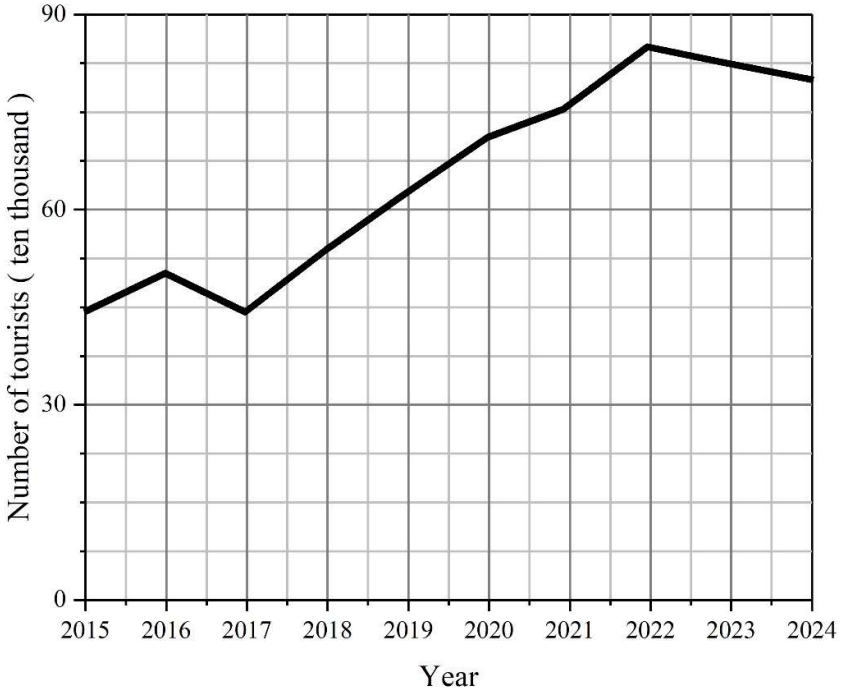
In this section, the BP-ARIMA combination model will be used to predict the leisure tourism preference of tourists, and the indicator used is the number of leisure tourism tourists, and the growth change of the number of leisure tourism tourists represents the trend of the change of leisure tourism preference of tourists.

4.2.1. ARIMA model of leisure travel preferences:

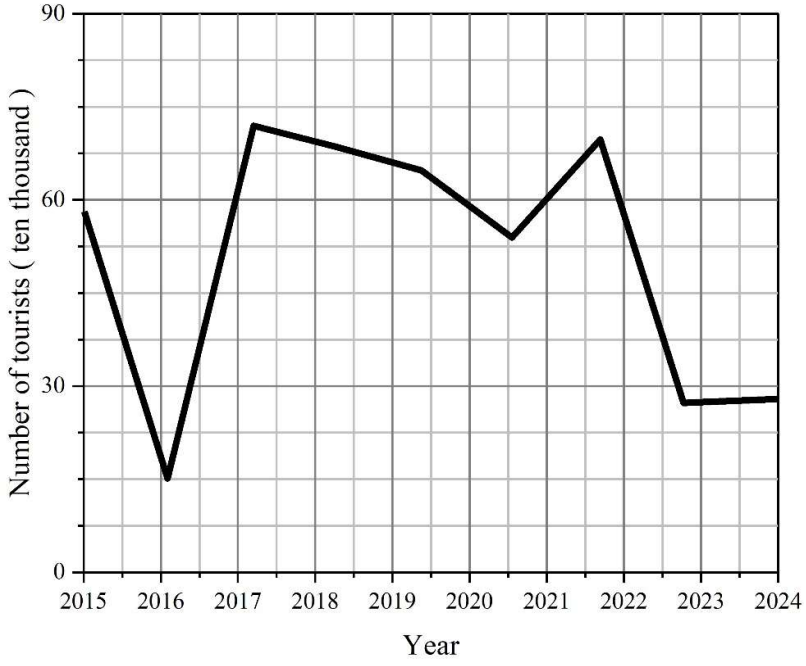
1) Smoothing of Leisure Travel Preference Historical Data. The time series trend graph of leisure tourism preference data from 2015-2024 and the transformation graph after difference processing are specifically shown in Figure 2. Figures (a) to (d) represent the time series trend at different step stages without variable transformation, after removing the data with non-normal values, after doing one difference for the variables, and after doing one difference for the variables with natural logarithm. From Figure (a), it can be seen that the general trend of the number of leisure tourism tourists can be seen to be gradually increasing, but in 2023 the data declined sharply. Therefore 2023 is an abnormal year. The average value of 2022 and 2024 is used to replace the data of 2023, and the time series of non-normal year values are eliminated. From figure (b) it can be seen that the growth of the series is different, which indicates that the time series has both an upward trend and heteroskedasticity, so it needs to be smoothened. One differencing and taking the natural logarithm after one differencing were carried out respectively, and it can be seen from Figs. (c) and (d) that the original variables have been basically smoothed out after natural logarithmic one differencing, so the parameters of the ARIMA model were set $d = 1$.



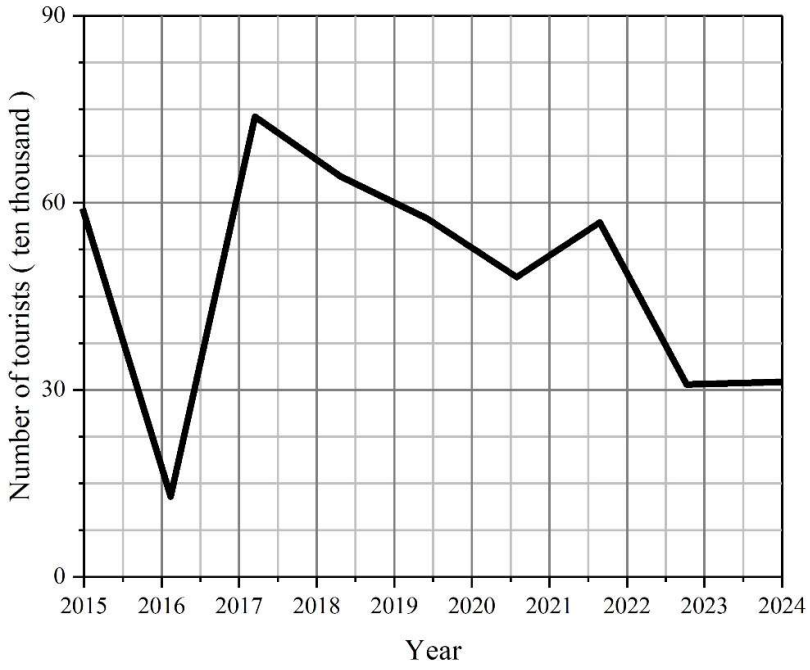
(a)Without variable conversion



(b)Data after eliminating abnormal values



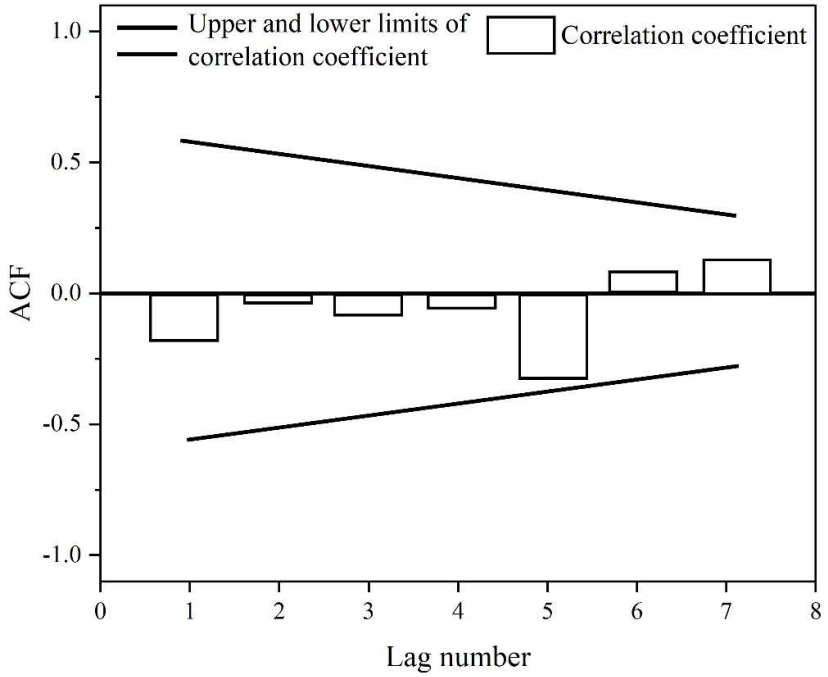
(c)Variables make a difference



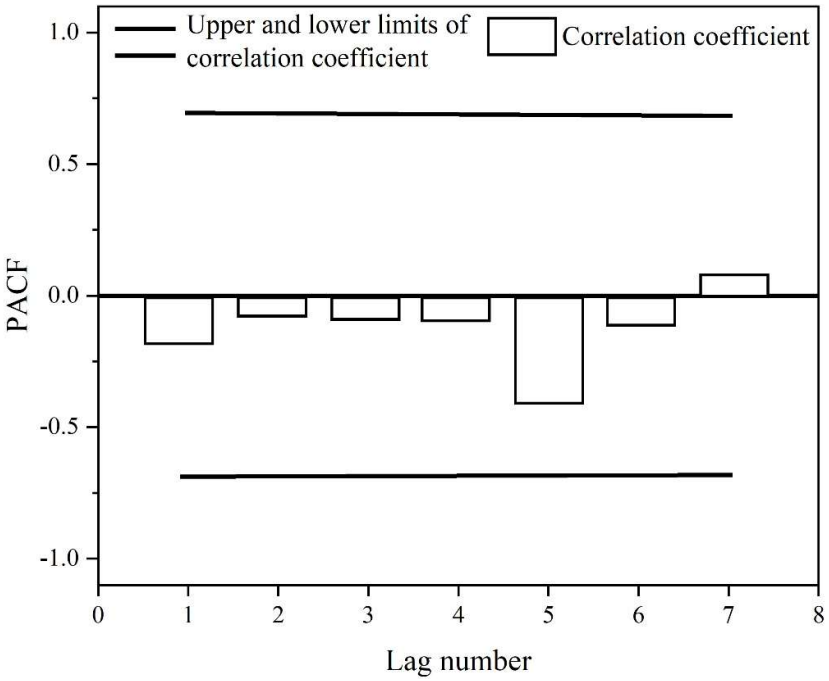
(d)The natural logarithm of the variable to do a difference

Fig. 2. Rough selection results of ARIMA model

2) Determination of ARIMA Model Parameters p and q. The parameters are pre-determined using ACF and PACF plots, as shown in Fig. 3. Figures (a) and (b) correspond to ACF and PACF, respectively. It can be seen that the data of both ACF and PACF plots are within the confidence interval. The idea of determining the parameters p and q of ARIMA model is to set different values of p and q, and by comparing the AIC value and SBC value, take the model with smaller SBC value.



(a)Charts of ACF



(b)Charts of PACF

Fig. 3. Charts of ACF and PACF

The effect indicators of the corresponding model of $ARIMA(p,1,q)$ are listed, as shown in Table 9. From the comparison of the indicators of each model, it can be seen that the model $ARIMA(2,1,2)$ is more suitable for predicting the leisure travel preference time series.

Table 9. Comparisons of AIC and SBC

Serial number	Model	AIC value	SBC value
1	$ARIMA(0,1,0)$	-13.588	-13.393
2	$ARIMA(0,1,1)$	-11.804	-11.4
3	$ARIMA(0,1,2)$	-11.239	-10.658
4	$ARIMA(1,1,0)$	-11.705	-11.316
5	$ARIMA(1,1,1)$	-10.965	-10.354
6	$ARIMA(1,1,2)$	-8.535	-7.73
7	$ARIMA(2,1,0)$	-9.482	-8.896
8	$ARIMA(2,1,0)$	-8.804	-8.024
9	$ARIMA(2,1,2)$	-6.053	-5.062

The $ARIMA(2,1,2)$ model was utilized to fit the data on the number of leisure tourism tourists, as shown in Table 10. It can be seen that the ARIMA model predicts the number of leisure tourism tourists from 2016 to 2024 are in error.

Table 10. Forecasted results with ARIMA model

Project	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Number of leisure tourists	44.2277	50.025	43.9457	54.0572	63.0189	71.0067	75.9261	85.0102	82.5067	80.0129
Predicted value	-	47.8236	53.5847	49.6567	59.8373	65.6073	72.8878	78.4089	87.2145	88.7977
Absolute error	-	2.2014	-9.639	4.4005	3.1816	5.3994	3.0383	6.6013	-4.7078	-8.7848

4.2.2. BP Neural Network Prediction. Since the error of ARIMA model prediction is only 2016-2024, the total sample size of BP neural network $N = 9$. Setting the training error requirement of BP neural network as 10-7, after 9 times of training, the prediction results are specifically shown in Table 11. From the table, it can be seen that the deviation between the predicted value and the actual value based on the combined ARIMA-BP prediction model is relatively small, while in addition the prediction deviation of the ARIMA model alone is large, and its absolute maximum deviation is 96,390,000 in 2017, and the minimum absolute deviation is 20,420,000 in 2001.

Table 11. Comparison of forecasted results with three different models

Model	Project	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
ARIMA	Number of leisure tourists	44.2277	50.025	43.9457	54.0572	63.0189	71.0067	75.9261	85.0102	82.5067	80.0129
	Predicted value	-	47.8236	53.5847	49.6567	59.8373	65.6073	72.8878	78.4089	87.2145	88.7977
	Absolute error	-	2.212	-9.6455	4.4007	3.1977	5.3994	3.0327	6.5876	-4.6939	-8.7632
BP	ARIMA error prediction	-	-	-	-	-	5.3993	3.0327	6.5875	-4.6939	-8.7632
	Relative error (%)	-	-	-	-	-	0.0019	0	0.0015	0	0
BP-ARIMA	Predicted value	-	-	-	-	-	70.0066	75.9261	85.0102	82.5067	80.0129
	Absolute error	-	-	-	-	-	0.0001	0	0	0	0

Using the ARIMA-BP combination forecasting model, the number of tourists in leisure tourism is forecasted for the future years 2027-2034, as shown in Table 12. From the table, it can be seen that the number of leisure tourism tourists will increase year by year, 2027-2029 basically maintained at a high level with an annual growth rate of more than 15%, after 2029 the growth rate slows down, with an average annual growth rate of 4.44%. The size of leisure tourism tourists will reach 1,691,280,000 in 2033, and the tendency of the tourism user's preference for leisure tourism is more significant.

Table 12. Forecasting of leisure tourism tourists in future years

Year	Number of leisure tourism tourists (ten thousand)
2027	103.356
2028	123.49
2029	142.282
2030	153.02
2031	157.047
2032	159.732
2033	169.128

5. Conclusion

In this paper, on the basis of constructing social media tourism user profiles, a two-step clustering algorithm is used to carry out cross-cultural analysis of social media data, and to mine and analyze the performance of users' leisure tourism preferences under the cross-cultural competence of social

media from the perspective of leisure tourism. A combined BP-ARIMA model is proposed to realize the prediction of leisure tourism preference.

Based on the user profiles, the leisure tourism preferences of tourists are divided into three categories of tourist groups. Comparing the category three tourist groups, the coefficient B value of category one tourist group in humanistic history and recreational activities are 0.793 and 2.896 respectively are greater than 0, which prefers tourist attractions rich in humanistic history and diverse recreational activities, while comparing the category two tourists, the coefficient B value of affordable is greater than 0 (B=2.237>0), which prefers affordable attractions. The category two tourist group compared to the category three tourist group in preferring strong humanistic history (B=0.95>0), rich recreational activities (B=3.48>0), variety of food and beverage (B=6.86>0), and affordable attractions (B=4.013>0), while comparing the category one tourist group in preferring distinctive architectural features (B=1.86>0), variety of food and beverage (B = 1.065>0) attractions. Compared to the Category 1 group, the Category 3 group preferred attractions with good guide service (B=5.39>0), distinctive architectural features (B=7.567>0), varied food and beverage offerings (B=4.925>0), and natural landscape type (B=1.574>0), while comparing to the Category 2 group, the Category 2 group preferred attractions with good guide service (B=4.594>0), distinctive architectural features (B= 6.652>0), more food and beverage (B=0.498>0), and natural landscape type (B=3.032>0) attractions.

Leisure travel preferences of travel users are predicted to always maintain a continuous growth trend in the number of leisure travelers in the coming years 2027-2034. In the years 2027-2029, basically maintain an annual growth rate of more than 15% growth level, after 2029 the growth rate began to slow down, the average annual growth rate remained at the level of 4.44%. In 2033, the number of leisure travelers will reach 1,691,280,000 tourists. Obviously, the future of tourism users will continue to enhance the preference for leisure tourism, for in the travel tourism program selection, will also increasingly favor leisure tourism program.

References

[1] Pardo, M. C., Almeida, S., & Campos, A. C. (2024). Creating new opportunities for tourism development through cross-border collaboration: Shedding light on overlooked destinations. Tourism and hospitality management, 30(3), 433-446.

[2] Akbar, I., Tazhekova, A., Myrzaliyeva, Z., Pazylkhaiyr, B., & Mominov, S. (2024). Positive outcomes of cross-border tourism development cooperation: a case of Kazakhstan, Kyrgyzstan and Uzbekistan. REGION, 11(2), 43-62.

[3] Paiva, T., Felgueira, T., Alves, C., & Costa, A. (2025). Strategies for Building Accessible and Inclusive Rural Tourism Ecosystems in Cross-Border Regions: The Case of Rural and Border Territory. Tourism and Hospitality, 6(1), 23.

[4] Kulyniak, I., Karyy, O., Bondarenko, Y., & Ohinok, S. (2024). Cross-Border Collaboration in Heritage and Cultural Tourism: The Role of Information Sharing in European Integration. Library of Progress-Library Science, Information Technology & Computer, 44(3).

[5] Yan, N., & Halpenny, E. (2019). The role of cultural difference and travel motivation in event participation: A cross-cultural perspective. International Journal of Event and Festival Management, 10(2), 155-173.

[6] McKay-Semmler, K. L. (2019). Travel analysis portfolio: Applying theories of cross-cultural communication to the task of personal travel planning. Discourse: The Journal of the SCASD, 5(1), 6.

[7] Gali, G. F., Shakhnina, I. Z., & Zagladina, E. (2021). International tourism as a substantial form of language interaction and intercultural communication. EntreLínguas, 7(7), 2.

[8] Setiawan, A. R. (2023). Tourism and Intercultural Communication: A Theoretical Study. Jurnal Komunikasi, 17(2), 186-195.

[9] AlDabbagh, M. (2019). Traditional Clothing, Souvenirs, and Food as Factors of Tourist Attraction. Journal of Home Economics, 29(1), 197-225.

[10] Zhang, J. J., & Doering, A. (2022). Introduction: the Interface of Culture and Communication Through Tourism. Tourism Culture & Communication, 22(2), 105-113.

[11] Zeng, Y. (2021). The Interactive Path of Cultural Symbols in Ethnic Tourism Consumption. International Journal of Frontiers in Sociology, 3(3).

[12] Loredana, V., Cornelia, P., Ramona, C., Diana, M., & Ioan, P. (2022). EXISTING SYMBOLS ON THE PALACES OF TIMISOARA--RESOURCES FOR THE URBAN CULTURAL TOURISM PRACTICING IN PLEVNA SQUARE. Agricultural Management/Lucrari Stiintifice Seria I, Management Agricol, 24(3).

[13] Xiong, H., & Vongphantuset, J. (2024). Cultural Internal Gaze and Cultural Symbol Alienation in Chinese Minority Cultural Tourism: A Case Study of the Yi Culture in Chuxiong. Journal of Multidisciplinary in Humanities and Social Sciences, 7(6), 3068-3082.

[14] Li, L., HOU, G. L., XIA, S. Y., & HUANG, Z. F. (2020). Spatial distribution characteristics and influencing factors of leisure tourism resources in Chengdu. Journal of Natural Resources, 35(3), 683-697.

[15] Mulet-Forteza, C., Genovart-Balaguer, J., Mauleon-Mendez, E., & Merigó, J. M. (2019). A bibliometric research in the tourism, leisure and hospitality fields. Journal of business research, 101, 819-827.

[16] Kirillova, K., Wang, D., & Lehto, X. (2018). The sociogenesis of leisure travel. Annals of Tourism Research, 69, 53-64.

[17] Wei, X., Taivan, A., & Hua, G. (2023). Love More and Buy More? Behavioral Economics Analysis of Travel Preference and Travel Consumption. Sustainability, 15(3), 1864.

[18] Beritelli, P. I. E. T. R. O. (2023). How Do Leisure Travel Decisions Come about. University of St. Gallen: St. Gallen, Switzerland.

[19] Hardt, D., & Glückstad, F. K. (2024). A social media analysis of travel preferences and attitudes, before and during Covid-19. Tourism Management, 100, 104821.

[20] Liu, X., Mehraliyev, F., Liu, C., & Schuckert, M. (2020). The roles of social media in tourists' choices of travel components. Tourist studies, 20(1), 27-48.

[21] Yuna, D., Xiaokun, L., Jianing, L., & Lu, H. (2022). Cross-cultural communication on social media: Review from the perspective of cultural psychology and neuroscience. Frontiers in Psychology, 13, 858900.

[22] Oliveira, T., Araujo, B., & Tam, C. (2020). Why do people share their travel experiences on social media?. Tourism Management, 78, 104041.

[23] Lim, K. H., Chan, J., Karunasekera, S., & Leckie, C. (2019). Tour recommendation and trip planning using location-based social media: A survey. Knowledge and Information Systems, 60, 1247-1275.

[24] Vu, H. Q., Li, G., & Law, R. (2020). Cross-country analysis of tourist activities based on venue-referenced social media data. Journal of Travel Research, 59(1), 90-106.

[25] Liao, J. C., Wang, Y. C., Tsai, C. H., & Zhao, B. (2021). Gratifications of travel photo sharing (GTPS) on social media: scale development and cross-cultural validation. Tourism Analysis, 26(4), 265-277.

[26] Amaro, S., & Duarte, P. (2017). Social media use for travel purposes: a cross cultural comparison between Portugal and the UK. Information Technology & Tourism, 17, 161-181.

[27] Li, Y., Lin, Z., & Xiao, S. (2022). Using social media big data for tourist demand forecasting: A new machine learning analytical approach. Journal of Digital Economy, 1(1), 32-43.

[28] Mirzaei Siamak & Hayati Ashkan Farrokh. (2018). An Intelligent Model for Internet Advertising Selection Based on User-Profile. International Journal of Engineering and Technology,10(4),1044-1051.

[29] Hankiz Yilahun & Askar Hamdulla. (2023). Entity extraction based on the combination of information entropy and TF-IDF. International Journal of Reasoning-based Intelligent Systems,15(1),71-78.

[30] Liang Gao,Fang Cui & Chengbo Zhang. (2024). Library Similar Literature Screening System Research Based on LDA Topic Model. Journal of Information & Knowledge Management,23(05).

[31] Bart J. J. Kremers,Jonathan Citrin,Aaron Ho & Karel L. Plassche. (2023). Two‐step clustering for data reduction combining DBSCAN and k‐means clustering. Contributions to Plasma Physics,63(5-6).

[32] Haitham Y. Adarbah,Mehdi Sookhak & Mohammed Atiquzzaman. (2024). A digital twin-based traffic light management system using BIRCH algorithm. Ad Hoc Networks,164,103613-103613.