International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5974-5986

https://doi.org/10.70517/ijhsa463514

Cross-cultural analysis and travel preference prediction of social media data: a study based on the K-nearest neighbor algorithm

Dongxia Wu[1,*]
[1] School of Culture and Tourism, Huangshan Vocational and Technical College, Huangshan, Anhui, 245000, China
Corresponding authors: (e-mail: hszywdx@163.com).

Abstract Travel-related content on social media platforms is exploding, and there are significant differences in travel behaviors and preference expressions among tourists from different cultural backgrounds. This study integrates text mining techniques and K-nearest neighbor algorithm to cross-culturally analyze travel data on social media platforms and predict tourists' preferences. The study crawled 3500 travel tips from Poor Traveler and GoWhere.com, and obtained 3000 valid data after cleaning. The TF-IDF algorithm is used to extract 50 high-frequency feature words, and the correlation matrix is constructed through the Ochiai coefficient, and the hierarchical clustering method is used to classify the tourism behaviors into three major categories and seven subclasses, namely, scenic area unique resource perception, entertainment experience behavior, and facility and service appeals. Meanwhile, an improved KNN algorithm based on vector orthogonalization and updated out-of-sample prediction method is proposed to predict the passenger flow at subway stations in A city. The results show that the average time-sharing prediction error of the whole network under 5-minute time granularity is 11.64%, and the cumulative all-day prediction error is 2.37%, and the prediction accuracy of the model is significantly better than that of the traditional method. The study found that more than 90% of the successfully matched samples were within one year before the prediction date, and the prediction accuracy was higher at the sites with higher passenger flow. This study provides an effective cross-cultural analysis framework and tourist preference prediction tool for the tourism industry, which can help companies develop accurate marketing strategies and personalized service plans.

Index Terms Social media data, Cross-cultural analysis, Tourist preference prediction, K-nearest neighbor algorithm, Text mining, Hierarchical clustering

I. Introduction

The emergence and rapid development of social media has changed the way we live and interact with each other. Social media has become an important tool for communication and exchange, both in personal life and business environment. In the field of tourism, social platforms such as TikTok, Xiaohongshu, and Instagram produce hundreds of millions of data about tourism-related data on a daily basis, especially after the epidemic, the rise of cross-border tourism has led to an explosive growth in the generation of tourism-related content on social media, which increases tourists' willingness to travel [1]-[3]. These contents reveal the differences in cultural expression of different cultural groups, such as the implicit emotional expression of Eastern tourists, while Western tourists' emotional expression is direct and open. East Asian tourists often generate content in the form of “text + pictures”, while European and American tourists often generate content in the form of video. Individual cultural tourists often rely on social media platforms for travel information, while collectivist tourists often rely on mutual recommendations [4]-[6]. Literature [7] reveals cultural differences in the selection of accommodations by tourists from different cultural backgrounds during tourism, comparing English-language tourists, Chinese-language tourists show positive comments about landlords, while more objective and strict comments about accommodation conditions, location and price. These different cultural expressions provide a valuable database for cross-cultural analysis. In addition, by analyzing and mining the characteristics and influencing factors of tourism-related user-generated content on social media platforms, studying users' attitudes and preferences towards tourism destinations, and proposing relevant prediction and recommendation algorithms, it can help enterprises better understand tourists' needs and preferences, so as to provide more personalized services and products, and provide the tourism industry with a basis for market research and competitive analysis [8]-[10].

Literature [11] realized a multidimensional analysis of the cross-cultural content of cultural tourism promotion content on the Instagram platform through LIWC2015 software, revealing the cultural differences in the form of tourism promotion. Literature [12] used RQDA (R Qualitative Data Analysis) and mutual information to analyze the differences in the visual representation of linear tourism destinations by tourists from different cultural backgrounds, and in the destination social networking site pictures show that their differences are mainly manifested in the spirit, culture, and ideology of the tourists, which are manifested in the visual symbol level of the tourists' photographs in the cultural differences. Literature [13] utilized user profiling and two-step clustering algorithms in social media for cross-cultural analysis and constructed a hybrid model of back-propagation neural network and autoregressive moving average model for predicting travel preferences. Literature [14] found that among a host of machine learning algorithms such as support vector machines, decision trees, and K-nearest neighbor algorithms, linear support vector machines possessed a better prediction of tourists' preferences, with an accuracy rate of more than 95%. Literature [15] proposed a tourist preference prediction system which is designed by multiple neural networks, ReLU activation function, Sigmoid output layer, binary cross-entropy function, application programming interface, etc., which is a prediction system with user questionnaire data as input. Literature [16] for tourism destination preference prediction, used edge computing, random forest algorithm, multivariate logistic selection model for preference feature construction, feature selection and preliminary ranking, feature sequence ranking, and finally determine feature weights, combined with link prediction method for preference prediction.

At present, cross-cultural analysis research is still in its infancy, with most limitations, such as user ecological differences, unspecified relevant cultural indicators of social media platforms, and poor algorithm interpretability. On travel preference prediction, some algorithms have poor performance for capturing the nature of cultural interactions and are prone to stereotyping. The k-nearest neighbor algorithm, on the other hand, is a classification method that does not learn the instances and their relationships, but stores the data directly, and when a classification prediction is needed, it searches for the K nearest points, and then makes a prediction based on the categories of these points. The k-nearest neighbor algorithm is widely used for data analysis and event prediction because of its computational simplicity, good predictability, and fast computation for large-scale data sets [17], [18].

Social media platforms have become the main channels for tourists to share their travel experiences and obtain travel information, and the huge amount of travel-related content generated every day contains rich differences in tourists' behavioral patterns and cultural expressions. Tourists from different cultural backgrounds show distinct cultural characteristics in terms of content generation, emotional expression, and travel decision-making: Eastern tourists tend to express their emotions implicitly, and content generation is often based on “text + pictures”; Western tourists are direct and open in their emotional expression, and prefer video; individualistic tourists rely on platforms to obtain travel information, while collectivistic tourists are more trustful of mutual recommendations. These cultural differences provide a unique perspective for understanding the travel behavior of global tourists. However, there are many challenges in analyzing cross-cultural tourism data: significant differences in user ecology, the lack of clear cultural indicators on social media platforms, and the poor performance of existing algorithms in capturing cultural interactions that are prone to stereotyping. Meanwhile, accurate prediction of tourists' preferences is important for destination management and enterprise marketing decisions, which can help relevant organizations provide more personalized services and products. K-nearest neighbor algorithms have been widely used in the field of data classification and prediction due to their advantages of simple computation, strong interpretability, and fast processing speed of large-scale data. This study proposes a comprehensive analysis framework that combines text mining techniques and improved K-nearest neighbor algorithm. Firstly, we obtain the text data of travel guide through web crawler technology, extract high-frequency feature words using TF-IDF algorithm, construct co-occurrence matrix and correlation matrix, and adopt hierarchical clustering method to identify the travel behavior patterns under different cultural backgrounds. Secondly, to address the problems of strong indicator correlation and unreasonable weight allocation of traditional KNN algorithm in travel preference prediction, an improvement scheme based on vector orthogonalization and updating out-of-sample prediction method is proposed to improve the prediction accuracy by dealing with the correlation problem and dynamically updating the training samples. The framework effectively integrates qualitative analysis and quantitative prediction, providing a new methodological support for cross-cultural research and precision marketing in tourism.

II. Sources of research data

In this paper, we use web crawler technology [19] to crawl the text data of travel tips about city A from Poor Travel and Go.com, and obtain a total of 3500 travel tips from 2014-2024. Due to the misinformation and logical problems in the part of the acquired travel guides, the data need to be cleaned. Firstly, attraction poi records that are not in the study area are removed, then duplicate travel guides are removed, and secondly, guides with less than 2 records of attraction visits in the study area are removed. After batch cleaning the data based on the above rules, the cleaned data is manually verified to remove advertisements, travel itineraries with logical errors, and tips that are not real travel experiences. Finally, 3000 useful travel tips were obtained. The time series data of Baidu index can reflect the changes of network heat and attention of different attractions in different time periods, and this paper crawled the overall daily average of the search index of the main attractions in City A from Baidu index platform from 2021 to 2024. In addition, subway line and station data of City A were obtained from Gaode Map for predicting travel preferences.

III. Cross-cultural analysis of social media data

This chapter analyzes the behavioral differences of tourists based on the collected textual data of travel tips in City A, as a reaction to the characteristics of tourism behavior in different cultural contexts.

III. A. Methods of analysis

III. A. 1) Text mining analysis

In order to extract and utilize the useful information hidden in online text data, text mining and content analysis are widely used in tourism research. Text mining mainly consists of three typical phases: data collection, data mining and result output, where the data mining phase includes two sub-steps: data preprocessing and pattern discovery.

The first step data collection. As mentioned before, web crawler technology is used for data collection.

Second step data mining. The collected online text data are analyzed to extract useful information through two sub-stages: data preprocessing and pattern discovery. In data preprocessing, different techniques are used for different research purposes and the common operations used in the existing tourism literature are data cleaning, lexical segmentation and lexical labeling.

(1) Data cleaning: is used to detect and remove inaccurate or useless records such as spelling mistakes, deactivated words, non-target languages, etc. from online text data in order to leave valuable tourism information.

(2) Segmentation: aims to decompose travel-related text data into words, phrases or other meaningful elements. Segmentation allows filtering valuable keywords from a large number of sentences about tourist attractions, travel sentiments, and so on.

(3) Lexical annotation: marking the lexical nature of each word in a sentence, such as noun, adjective or adverb, in order to remove unimportant other labeled words.

Pattern discovery is another key stage of text mining, aiming at exploring interesting information in documents. Typical techniques used in existing tourism research are LDA analysis, sentiment analysis, statistical analysis, clustering and classification, text summarization and dependency modeling.

(1) LDA analysis: a topic model used to identify abstract “themes” in text data, e.g., LDA is used to quickly discover mixed themes from a large number of reviews about factors that influence customer satisfaction in hotels.

(2) Sentiment analysis: identifying tourists' attitudes towards tourism products or attractions by categorizing text data into sentiment categories: positive, negative or neutral.

(3) Statistical analysis: In tourism research, descriptive statistics, t-tests, correlation matrices, Kruskal-Wallis tests, and correspondence analyses are widely used for a variety of information contained in online textual data, such as review data, reviewer characteristics, and so on.

(4) Clustering and Classification: Clustering is to discover similar objects to form a collection, and common methods include hierarchical clustering, K-clustering, hierarchical clustering, and so on. Classification is based on the existing classification system table for classification.

(5) Text summarization: used to automatically generate concise summaries of single or multiple documents for refining key information from the original text.

(6) Dependency modeling: used to capture the relationship between text data and tourism factors, traveler behavior, etc. In the current research, Bayesian ordered logit model, linear regression model, Tobit regression model and other models have been applied.

The third step result output. The interesting information extracted through data mining is converted into useful knowledge to further serve tourism research. According to related studies, the valuable knowledge covers tourist satisfaction, consumption preference, destination image, travel routes, review features, etc.

(1) Chinese Segmentation

Chinese participle has been in constant exploration, according to the realization principle and characteristics of Chinese participle, Chinese participle algorithm can be roughly divided into string matching based participle algorithm, statistical participle algorithm and semantic participle algorithm [20]. At present, there is no conclusive statement that which kind of word separation algorithm has high accuracy, for mature word separation system, can not rely on only one kind of word separation algorithm, often a combination of multiple word separation algorithms for Chinese word separation. The main participle tools are Jieba participle, HIT LTP, Ansj participle, etc. Jieba participle accuracy is up to 97% or more, and supports customized dictionaries, it is a participle system with higher efficiency and accuracy, so this paper chooses Jieba to perform participle. the principle of Jieba participle is: based on the statistical lexicon, a prefix lexicon is constructed. Then the prefix dictionary is used to slice the input sentence to get all possible slice results. According to the cut position, construct a directed acyclic graph (DAG). The maximum probability path is computed by dynamic programming to get the final form of the cut.Jieba Segmentation has three modes: full mode is the default mode, which will slice all the possible segmentation cases in the sentence. Precise mode tries to slice the sentence most accurately and is suitable for text analysis. The search engine mode is based on the precise mode and continues to cut long words.

1) Segmentation algorithm based on string matching

The idea of string matching based participle algorithm is: first, set up a matching rule, and then match the Chinese string sequence to be participle and the words in the “rich enough” participle dictionary according to the set rule, and if the matching is successful under the set rule, then the participle is realized. This algorithm can be divided into the following categories according to the different search directions: forward matching, reverse matching and two-way matching. Simply using the string-based matching algorithm has a lot of defects, so it is often necessary to combine with other algorithms to improve the effectiveness of the word.

2) Statistics-based word separation algorithm

Statistically based word-splitting algorithms are based on the idea of probabilistic combination is: it is considered that the frequency of simultaneous occurrences of neighboring Chinese characters somewhat represents the probability that they form a word. If the algorithm is expressed in mathematical language: that is, for a string Y, there may exist m kinds of segmentation results:

$$\left\{\begin{matrix} X_{11} & X_{12} & \cdots & X_{1n_1} \\ X_{21} & X_{22} & \cdots & X_{2n_2} \\ \vdots & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn_m} \end{matrix}\right\} \tag{1}$$

where $n_i(i=1,2,....m)$ denotes the number of words in the i th participle result. If we want to obtain the expected optimal participle result j we have to make the distribution probability corresponding to this participle result the largest among all participle results, i.e.:

$$j = \arg\max_{i} P(X_{i1}, X_{i1}, ..., X_{im_i}) \tag{2}$$

Since $P(X_{i1}, X_{i1}, ..., X_{im_i})$ involves the joint distribution of n_i words, it is more difficult to solve. Therefore, Markov's assumption is often introduced in the actual solution, i.e., it is considered that the occurrence of the next word is only related to one or several words before it.

If the n th word is related to the preceding 1 word, the distribution probability of the partition result is shown in Equation (3):

$$P(X_{i1}, X_{i1}, ..., X_{im_i}) = P(X_{i1})P(X_{i2} \mid X_{i1})...P(X_{in_i} \mid X_{i(n_i-1)}) \tag{3}$$

If the n th word is related to the previous 2 words, the distribution probability of the partition result is shown in Equation (4):

$$P(X_{i1}, X_{i1}, ..., X_{im_i}) = P(X_{i1})P(X_{i2} \mid X_{i1})P(X_{i3} \mid X_{i1}, X_{i2})...P(X_{in_i} \mid X_{i(n_i-2)}, X_{i(n_i-1)}) \tag{4}$$

Eq. (3) is called the binary model and Eq. (4) is called the ternary model, which can be generalized to the N model accordingly.

3) Semantic-based word segmentation algorithm

The main idea of the semantic-based participle algorithm is that the text is participle based on the Chinese grammar, combined with the context. Due to the complexity of Chinese language rules and semantics, this humanassisted semi-supervised learning requires a lot of linguistic knowledge and information, so the semantic-based participle system is still in the experimental stage.

(2) TF-IDF Algorithm

TF-IDF, i.e., Word Frequency-Inverse Document Frequency, is an unsupervised statistical algorithm commonly used in the field of information retrieval and data mining to assess the importance of words in a particular document in a corpus. The importance of a word is directly proportional to the number of times it appears in that document and inversely proportional to the frequency of the word in other documents in the corpus [21].

TF, or word frequency, represents the number of times a word appears in a particular document, and this number is usually normalized to avoid TF bias towards long documents. Defining any word as T_i, the TF of the word T_i is calculated as:

$$TF_{i,j}=\frac{N_{i,j}}{\sum_k N_{k,j}} \tag{5}$$

In Eq. (5), $N_{i,j}$ is the number of times the word T_i occurs in the document D_j, and $\sum_k N_{k,j}$ is the sum of the occurrences of all words in the document D_j.

IDF, i.e., Inverse Document Frequency, represents the distribution of documents containing a given word in the corpus. The IDF of the word T_i is calculated as:

$$IDF_i=\log\frac{|D|}{1+|\{j:T_i\in D_j\}|} \tag{6}$$

In Eq. (6), $|D|$ is the total number of documents in the corpus, and $1+|\{j:T_i\in D_j\}|$ is the number of documents containing the word T_i, and the addition of 1 is to avoid the denominator to be 0 (i.e., the case that all documents do not contain the word).

TF-IDF is calculated as in Equation (7), it can be seen that the larger the TF (the more times a word appears in a document) and the smaller the IDF (the fewer times the word appears in the corpus), the larger the TF-IDF value is, which means that the word is more important and representative of the article.

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i=\frac{N_{i,j}}{\sum_k N_{k,j}}\times\log\frac{|D|}{1+|\{j:T_i\in D_j\}|} \tag{7}$$

III. A. 2) Co-occurrence analysis

Co-occurrence refers to the phenomenon of co-occurrence of specific keywords in a text collection. Co-occurrence analysis is a quantitative study of the co-occurrence phenomenon to reveal the content associations of the information and the knowledge implied by the feature items, and its basic principle is to reflect the strength of the association between these words by counting the co-occurrence of high-frequency words in the text collection [22]. There are four main stages in the specific process of co-occurrence analysis: first, determine the analysis data set. The second is to determine the object of analysis: extract the high-frequency words related to the object of study from the text collection. The third is to construct a binary co-occurrence matrix to derive quantitative information such as co-occurrence frequency. Fourth, to conduct co-occurrence analysis, which is generally combined with social network analysis for visualization and presentation.

In this paper, co-occurrence analysis is mainly realized by NLPIR software, whose calculated quantitative information includes co-occurrence frequency, binary probability, and binary word pair information entropy. Co-occurrence frequency refers to the number of times two words co-occur in the previous and subsequent order, which is used to reflect the strength of the association between them. It is generally believed that the higher the co-occurrence frequency of a word pair, the closer the relationship between the words. Binary probability refers to the probability of occurrence of the co-occurring word pairs, and the binary word pair information entropy indicates the breadth of information contained in the phrase, and its information entropy is calculated by the formula:

$$H(p,q)=-\sum_{x\in X}p(x)\log q(x) \tag{8}$$

III. B. Analysis of results

III. B. 1) Text Segmentation

Segmentation is the basic work of text mining, and the text needs to be preprocessed during segmentation.

Configure the user dictionary. As the sample data belongs to the special whisker domain of tourism, some tourism-specific words in it. Similarly, place names, attraction names, activity names, etc. are not included in the dictionary that comes with the lexicon software, which may not be able to automatically identify and correctly cut, thus affecting the subsequent analysis results. In view of this, this paper summarizes the relevant special vocabularies and compiles a user word list as a supplement by collating relevant information about City A, the sample of this paper, from Ctrip.com, Baidu Encyclopedia, and official websites. In addition, the official names of some words may differ from those expressed by tourists in their travelogues, which is generally the difference between the whole process and the abbreviation. However, similar is the hotel, inn such words, although from the functional point of view of the meaning of similar, but in fact reflects the different characteristics of tourism behavior.

The main purpose of configuring the deactivation dictionary is to filter out the words that are not related to the research content and affect the research efficiency, including Chinese and English punctuation, special symbols, prepositions, conjunctions, intonation words, and strings that have certain rules but no meaning.

After completing the document preprocessing, the sample data were subjected to word splitting. A total of 21,000 words were obtained.

III. B. 2) High Frequency Feature Word Extraction

The words obtained through word separation are the basic units constituting the text, if all of these huge numbers of words are retained, it will make the text dimension too high, which seriously affects the analysis effect, we need to streamline it and extract the valuable information that meets the demand, i.e., to extract the feature words that can reflect the tourism behavior as the knowledge unit.

In this paper, we use the TF-IDF algorithm for feature word extraction, which utilizes word word frequency and inverse document frequency to assess the importance of a word in the corpus. The final constructed high-frequency feature words are shown in Table 1. Among them, “inn, starry sky, guest room, ancient town, history” is a high-frequency word in the text.

Table 1 High frequency special term

N	Special	Frequency	N	Special	Frequency
1	Tavern	1000	26	Walk	588
2	Starry sky	940	27	Music fountain	531
3	Guest room	913	28	Night scene	506
4	Ancient town	897	29	Queueing	504
5	History	875	30	Villa	484
6	Lantern	838	31	Feature	472
7	Hot spring	820	32	Snack	416
8	Leisure	814	33	Perimeter	412
9	Facilities	812	34	Dining room	390
10	Admission ticket	806	35	Boutique hotel	381
11	Fountains	789	36	Peripheral swimming	360
12	Swimming pool	767	37	Performing	353
13	Lodge	757	38	Hot pot	342
14	Cable car	753	39	Academy	337
15	Jiangnan water township	752	40	Scenic spot	336
16	Room	735	41	Photography	336
17	Activity	721	42	Water dance show	308
18	Vacation	715	43	Service	269
19	Plaza	712	44	Free line	267
20	Rest	700	45	Drive	260
21	Consignment	694	46	Parking lot	240
22	Civil house	692	47	Wine shop	224
23	environment	685	48	Tent	206
24	perform	682	49	Travel	168
25	Dyeing garden	658	50	Ferry	141

III. B. 3) Co-occurrence and correlation matrix construction

The 50 high-frequency feature words extracted from the text of the tourism strategy of city A are the knowledge units composing the knowledge of tourism behavior, but the current state of these feature words is scattered and isolated, and only according to the frequency of occurrence from high to low can not show the connection between them. Therefore, it is necessary to rely on the total text set to construct a feature word co-occurrence matrix to solve the representation of word vector proximity, and to seek the intrinsic connection between the knowledge contents. The co-occurrence matrix counts the frequency of every two words appearing in the same text, which can reflect the degree of closeness between words. There are the same feature words in different travel guide texts, among the 50 extracted high-frequency feature words, if any two words appear in the same travel guide, the number of co-occurrence of these two feature words will be accumulated once, the number of the same feature words in all the travel guide texts will be counted, and the co-occurrence relationship will be deposited into a two-dimensional array to get a symmetric matrix of 50×50. The final high-frequency feature word co-occurrence matrix (part) is shown in Figure 1.

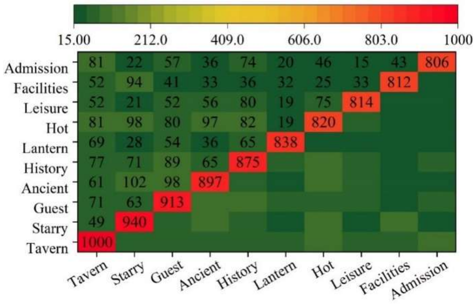


Figure 1: The common eigenvalues of the high frequency special term (part)

The values in the matrix, i.e., co-occurrence frequency, reflect the strength of co-occurrence relationship between knowledge units, and in co-word analysis, what is needed is the strength of correlation between knowledge units, so it is necessary to use the correlation coefficient to calculate the similarity between concepts based on co-occurrence relationship, and to convert the eigen-word co-occurrence matrix into a similarity matrix. In this paper, the similarity is calculated using the Ochiia coefficient, which is often used to transform the correlation matrix of binarization matrix. The Ochiia coefficient between words is calculated by the following formula:

$$\textit{Ochiia}\ \text{coefficient}=\frac{\text{Frequency of co-occurrence of two keywords AB}}{\sqrt{\text{The total frequency of A}}\cdot\sqrt{\text{The total frequency of B}}} \tag{9}$$

The feature word similarity matrix (partial) is obtained by calculation in EXCEL as shown in Figure 2. In the similarity matrix, the value range is [0, 1], and the larger the value between two words, the closer the relationship between the words and the higher the similarity. Conversely, the smaller the data between two words, the more distant the two keywords are from each other and the worse the similarity. The number “1” on the diagonal in the correlation matrix indicates the degree of association between the keyword and itself.

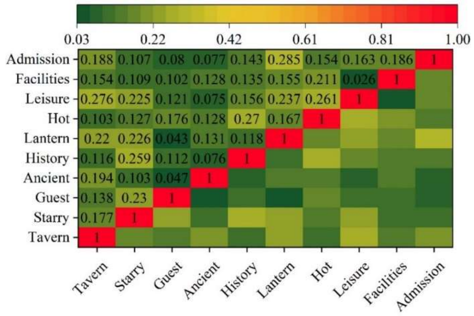


Figure 2: A similar matrix of a special term (part)

The similarity matrix is converted to a dissimilarity matrix, where dissimilarity value = (1 - similarity value), and the eigenword dissimilarity matrix (partial) is shown in Figure 3. The larger the value in the dissimilarity matrix, the lower the similarity, and the smaller the value, the higher the similarity.

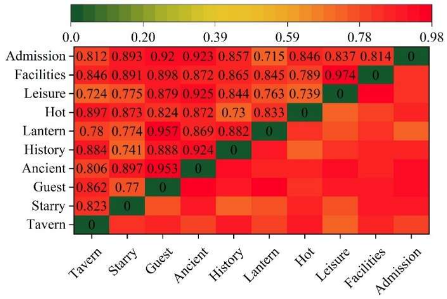


Figure 3: The different matrix of the special term (part)

III. B. 4) Cluster analysis

In this section, the clustering analysis will be carried out by using the feature word dissimilarity matrix obtained from the conversion, adopting the method of hierarchical clustering, calculating the affinity relationship between nodes in accordance with the numerical characteristics of the things, and simplifying the complex covariate mesh relationship between many analyzed objects into the relationship between the class groups of relatively less numerical values, i.e., the information that is closely associated with the information is aggregated to form the class clusters. Through the cluster analysis, the tourism behavior information is organized and integrated to achieve the discovery of systematic tourism behavior knowledge. The dissimilarity matrix is imported into SPSS software for hierarchical clustering analysis, and the inter-group connection distance averaging method is selected to obtain the hierarchical clustering analysis results shown in Figure 4. The clustering results are divided into three major categories: scenic unique resource perception, entertainment experience behavior, and facility and service appeals, as well as seven subcategories based on the knowledge content associations within the class groups: creative cultural display attractions, natural and humanistic landscapes, performance activities, experience behaviors, lodging behaviors, catering behaviors, and transportation behaviors. Based on the results of this classification, a cross-cultural analysis can be conducted to explore the differences in tourist behavior in different cultural contexts.

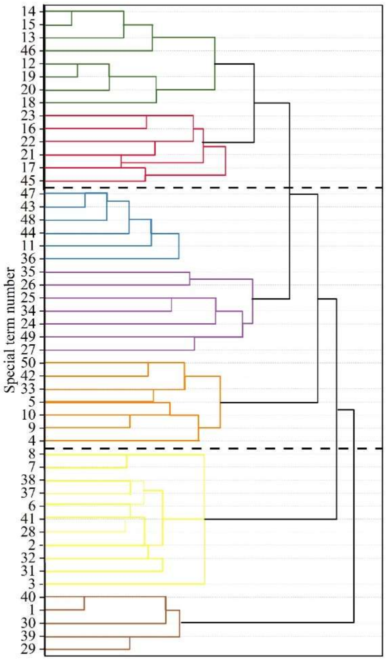


Figure 4: Hierarchical cluster analysis results

IV. Travel preference prediction based on K-nearest neighbor algorithm

This chapter predicts tourists' travel preferences based on the collected data of subway lines and stations in A city. And the prediction method and prediction results are analyzed relatedly.

IV. A. Forecasting methodology

IV. A. 1) Traditional KNN algorithm

The KNN algorithm classifies the samples to be classified by comparing the categories belonging to the K nearest neighbors to the samples to be classified [23], the basic idea is to calculate the distance between the samples to be classified and the training samples and select the K training samples closest to the samples to be classified, and determine the category with the greatest number of these K samples as the category of the samples to be classified, and the specific steps are:

(1) Determine the value of K.

(2) Calculate the distance from the training sample point to the point of the sample to be categorized using the Euclidean distance, the specific formula is:

$$d(a,b)=\sqrt{\sum_{i=1}^{k}(X_i^a-X_i^b)^2} \tag{10}$$

In the above equation, X_i^a denotes the i th indicator value of a object, and X^b denotes the b th indicator value of b object, where a and b are the sample points to be categorized and the training sample points, respectively.

(3) Select the K training sample points that are closest to the sample point to be classified.

(4) Use the category with the most of these K training sample points as the category of the sample to be classified.

From the above steps, it can be seen that the KNN classification algorithm is simple to operate, so it is often used as an attempted method for data classification. However, there are the following two problems in its classification based on travel preferences.

(1) Due to the strong correlation of different indicators of tourism preference and the very large difference in values between different indicators, it may lead to the problem of unreasonable weight allocation for distance calculation.

(2) In the traditional KNN algorithm, the time is simply divided into training set and testing set, while in the actual traveling process, travelers will update their data according to the daily information and produce new judgments.

Based on the above two problems, this paper proposes to improve the KNN algorithm by using vector orthogonalization and updating out-of-sample prediction method in the hope of obtaining a more suitable model for making predictions on travel preferences.

IV. A. 2) Improving the KNN algorithm

The process of the improved KNN algorithm obtained in this paper based on vector orthogonalization and updated out-of-sample prediction method is as follows:

(1) The matrix containing travel preferences is orthogonalized to obtain mutually independent factors.

(2) Determine the value of K.

(3) Select only one sample to be classified at a time, and the training sample follows up with the prediction point (after the prediction of the first day is completed, when predicting the second day, the training sample is also overall to the previous day, the first day's data is updated into the training sample, and the first day's data in the original training sample is moved out).

(4) Repeat steps (2) to (4) in the traditional KNN algorithm.

Where step (1) orthogonalization is done as follows:

The existing group of variables is a matrix of $m\times n$, where n (column vector) denotes different indicators and m (row vector) denotes different samples (e.g., different times of day, travel preferences), i.e., the same columns denote the same indicators for different samples and the same rows denote different indicators for a single sample:

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \tag{11}$$

Standardization of X:

$$\overline{x}_J = \frac{\sum_{i=1}^{m} x_{ij}}{m}, \sigma_J = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \overline{x}_J)^2}{m}}, y_{ij} = \frac{x_{ij} - \overline{x}_J}{\sigma_J} \tag{12}$$

The new matrix obtained is the normalized matrix of X and from this, the matrix of correlation coefficients R is calculated, reflecting the correlation between the indicators:

$$R = Y^T Y = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix} \tag{13}$$

Calculating the n eigenvalues (λ) and corresponding eigenvectors (μ) of the R matrix yields the factor loading matrix A that responds to the relationship of each independent factor to the original indicator:

$$A = \begin{bmatrix} \mu_{11}\sqrt{\lambda_1} & \cdots & \mu_{1n}\sqrt{\lambda_n} \\ \vdots & \ddots & \vdots \\ \mu_{m1}\sqrt{\lambda_1} & \cdots & \mu_{mn}\sqrt{\lambda_n} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{14}$$

The matrix is then obtained by multiplying the original variable matrix X by the factor loading matrix:

$$Factor = XA \tag{15}$$

Each column of this matrix, i.e., factors independent of each other, can be substituted into the subsequent steps of the improved KNN algorithm.

IV. B. Analysis of forecast results

Relying on the data of subway lines and stations in City A to analyze the accuracy of the prediction model, the all-day time-sharing average absolute percentage error E_T and the cumulative all-day average absolute percentage error E_D at the stations are used as the evaluation indexes, and the formula is:

$$E_T = \frac{1}{N} \sum_{t=1}^{N} \frac{| P_{T,t}^i - V_{T,t}^i |}{V_{T,t}^i} \cdot 100\% \tag{16}$$

$$E_D = \frac{| \sum_{i=1}^{N} P_{T,t}^i - \sum_{i=1}^{N} V_{T,t}^i |}{\sum_{i=1}^{N} V_{T,t}^i} \cdot 100\% \tag{17}$$

First, the prediction effect of the model at different time granularity is tested, and the results are shown in Table 2. Compared with the more similar methods of the prediction model in this paper, the average E_T of the whole network under 15 min granularity is 12.4%, which proves that the improved KNN algorithm has stronger timeliness and accuracy. In the subsequent analysis, 5 min time granularity is used by default.

Table 2: Forecast error under different granularities (%)

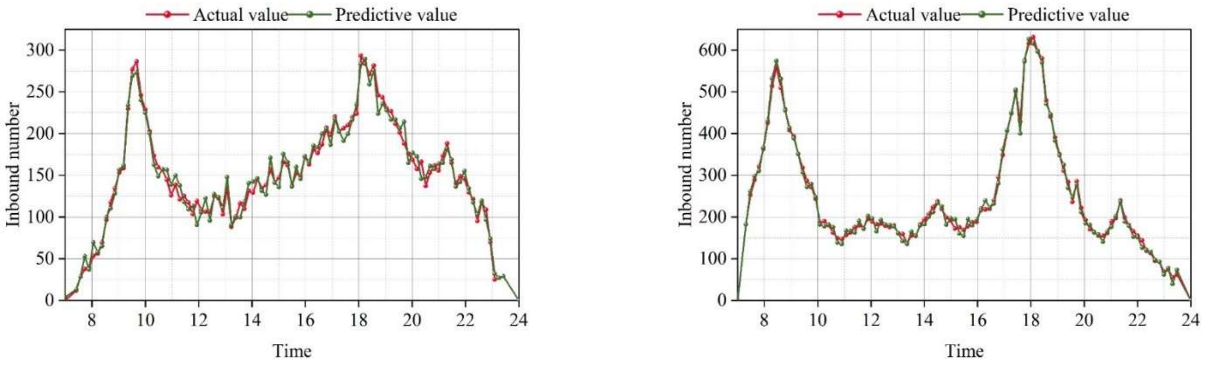
Forecast date	E_D			E_T		
	5min	15min	1h	5min	15min	1h
2024/12/23	2.22	1.09	0.81	11.17	9.84	6.9
2024/12/24	2.13	1.24	0.81	11.91	9.24	7.04
2024/12/25	2.05	1.21	0.8	10.68	8.47	6.58
2024/12/26	2.28	1.12	0.98	11.73	9.51	6.78
2024/12/30	2.75	1.34	1.29	12.23	10.46	8.3
2024/12/31	2.79	1.29	0.93	11.64	9.66	7.6

Then, the analysis was performed at the line level and the results are shown in Table 3. Stations with higher passenger flow tend to have better prediction accuracy. In the case of Line 7, for example, there are about 90 periods per day when the inbound volume is less than 5 passengers, resulting in a higher error for this station, but such effects can be ignored in the application.

Table 3 Forecast error of each line

Line	Station number	Site 5 minAverage inbound quantity/number	Average E_T /%	Min E_T /%
Line 1	17	264.48	7.07	5.33
Line 2	26	204.49	8.14	6.2
Line 3	17	269.88	7.61	4.41
Line 4	14	247.6	8.94	5.04
Line 5	15	73.14	14.85	7.66
Line 6	25	184.8	10.31	9.42
Line 7	31	127.11	13.98	7.92
Line 8	10	23.93	23.65	6.84
Line 9	14	142.62	13.53	6.47
Line 10	22	65.51	16.24	9.35

Finally, taking 2 typical stations with different passenger flow patterns as an example, the complete prediction results of a certain day are randomly selected to be shown, and the prediction results are shown in Fig. 5, with (a) and (b) denoting the prediction results of Station 1 and Station 2, respectively. It can be seen that the method shows good results in the prediction of different types of stations, and basically there is no local deviation phenomenon.



(a) Station 1

(b) Station 2

Figure 5: Number of arrivals forecast

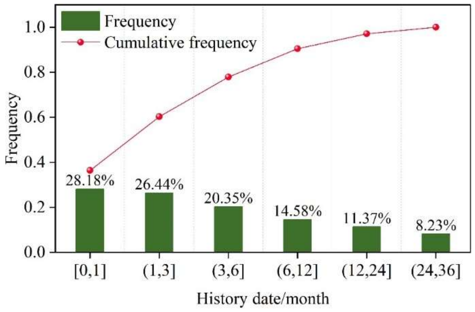


Figure 6: Frequency distribution histogram of sample selection

Finally, the applicability of the model under different data conditions is explored. Based on the long-term real-time prediction record of 2024, the actual utility played by historical data in the prediction in different periods is analyzed, and the frequency of its being successfully matched as a nearest neighbor is counted, and the results are shown in Fig. 6. More than 90% of the samples that were actually matched were within one year before the forecast date. Therefore, to ensure the prediction accuracy, the historical database should try to cover the passenger flow data of the recent year. For special scenarios such as holidays, historical samples of similar scenarios with longer time frames should be provided in conjunction with the actual situation.

V. Conclusion

Through the in-depth mining of 3,000 travel tips, a tourism behavior knowledge system covering 50 high-frequency feature words was successfully constructed, and the core tourism elements represented by “inn” (1,000 times), “starry sky” (940 times), and “guest room” (913 times) were revealed. The cross-cultural analysis framework effectively categorizes tourism behaviors into three major categories and seven sub-dimensions, which provides a scientific basis for understanding the behavioral differences of tourists from different cultural backgrounds. The improved K-Nearest Neighbor algorithm performs excellently in predicting tourists' preferences, with the time-averaged absolute percentage error of the time-sharing under the 5-minute time granularity controlled within 12.23%, which is significantly better than that of the traditional method under the 15-minute granularity (12.4%). The timeliness analysis of the prediction model shows that historical data within the last year accounted for more than 90% of the successfully matched samples, verifying the important impact of data timeliness on prediction accuracy. The analysis of the difference in the prediction accuracy of different subway lines shows that the passenger flow is positively correlated with the prediction accuracy, e.g., the prediction error of Line 1, which has the largest passenger flow, is only 7.07%, while the error of Line 8, which has smaller passenger flow, reaches 23.65%. This study not only enriches the theory of cross-cultural analysis of social media data, but also provides practical tools for tourism enterprises to formulate differentiated marketing strategies and optimize the allocation of service resources, which is of great value in promoting the digital transformation of tourism.

References

[1] Chen, J., Becken, S., & Stantic, B. (2022). Harnessing social media to understand tourist mobility: The role of information technology and big data. Tourism Review, 77(4), 1219-1233.

[2] Latif, K., Malik, M. Y., Pitafi, A. H., Kanwal, S., & Latif, Z. (2020). If you travel, I travel: Testing a model of when and how travel-related content exposure on Facebook triggers the intention to visit a tourist destination. Sage Open, 10(2), 2158244020925511.

[3] Da Mota, V. T., & Pickering, C. (2020). Using social media to assess nature-based tourism: Current research and future trends. Journal of Outdoor Recreation and Tourism, 30, 100295.

[4] Veiga, D. A., Frizzo, G. B., & Silva, T. H. (2019, October). Cross-cultural study of tourists mobility using social media. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 313-316).

[5] Sharmin, F., Sultan, M. T., Wang, D., Badulescu, A., & Li, B. (2021). Cultural dimensions and social media empowerment in digital era: travel-related continuance usage intention. Sustainability, 13(19), 10820.

[6] Wu, M., Jiang, C., Zhang, Y., Cao, J., Cheng, Y., & Liu, Y. (2021). Culture vs. distance: comparing the effects of geographic segmentation variables on tourists’ destination images based on social media data. Computational Urban Science, 1, 1-14.

[7] Xi, Y., Ma, C., Yang, Q., & Jiang, Y. (2022). A cross-cultural analysis of tourists’ perceptions of Airbnb attributes. International Journal of Hospitality & Tourism Administration, 23(4), 754-787.

[8] Ana, M. I., & Istudor, L. G. (2019). The role of Social Media and user-generated-content in Millennials travel behavior. Management dynamics in the knowledge economy, 7(1/23), 87-104.

[9] Ji, G. M., Cheah, J. H., Sigala, M., Ng, S. I., & Choo, W. C. (2023). Tell me about your culture, to predict your tourism activity preferences and evaluations: cross-country evidence based on user-generated content. Asia Pacific Journal of Tourism Research, 28(10), 1052-1070.

[10] Sun, Y., Ma, H., & Chan, E. H. (2017). A model to measure tourist preference toward scenic spots based on social media data: A case of Dapeng in China. Sustainability, 10(1), 43.

[11] Mele, E., Kerkhof, P., & Cantoni, L. (2021). Analyzing cultural tourism promotion on Instagram: a cross-cultural perspective. Journal of Travel & Tourism Marketing, 38(3), 326-340.

[12] Wei, Y., & Wu, T. (2021). Visual representation of a linear tourist destination based on social network photos: a comparative analysis of cross-cultural perspectives. Journal of Tourism and Cultural Change, 19(6), 781-804.

[13] Wu, D. (2025). A Study on Cross-Cultural Analysis of Social Media Data and Leisure Travel Preference Prediction Supported by Cluster Analysis Algorithm. J. COMBIN. MATH. COMBIN. COMPUT, 127, 5899-5926.

[14] Chang, V., Islam, M. R., Ahad, A., Ahmed, M. J., & Xu, Q. A. (2024). Machine learning for predicting tourist spots’ preference and analysing future tourism trends in Bangladesh. Enterprise Information Systems, 18(12), 2415568.

[15] TSAKIRIDIS, S., PAPAIOANNOU, N., VRANA, V., & VARSAMIS, D. (2025, March). ADAPTIVE NEURAL NETWORK MODELS FOR TOURISM PREFERENCE PREDICTION: ACase STUDY IN SERRES. In Conference on Organizational Science Development Human Being, Artificial Intelligence and Organization.

[16] Deng, B., Xu, J., & Wei, X. (2021). Tourism destination preference prediction based on edge computing. Mobile Information Systems, 2021(1), 5512008.

[17] Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(2), e1289.

[18] Novita, D., Herlambang, T., Asy’ari, V., Alimudin, A., & Arof, H. (2024). COMPARISON OF K-NEAREST NEIGHBOR AND NEURAL NETWORK FOR PREDICTION INTERNATIONAL VISITOR IN EAST JAVA. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 18(3), 2057-2070

[19] Guo Li.(2020). Research on Web Crawler Technology for User Demand Large Data Collection. BASIC & CLINICAL PHARMACOLOGY & TOXICOLOGY,126,259-259.

[20] TobiasSchack,MaxCoenen & MichaelHaist. (2024). Image-based quality control of fresh concrete based on semantic segmentation algorithms. Civil Engineering Design,6(3),96-105.

[21] Hainan Wang. (2024). Automatic question-answering modeling in English by integrating TF-IDF and segmentation algorithms. Systems and Soft Computing,6,200087-.

[22] Grisha Weintraub,Noam Hadar,Ehud Gudes,Shlomi Dolev & Ohad S Birk. (2024). GeniePool 2.0: advancing variant analysis through CHM13-T2T, AlphaMissense, gnomAD V4 integration, and variant co-occurrence queries. Database : the journal of biological databases and curation,2024,

[23] Yoschanin Sasiwat,Dujdow Buranapanichkit & Apidet Booranawong. (2024). Implementation and test of a Device-Free localization system with a modified desync network protocol and a weighted k-nearest neighbor algorithm. Egyptian Informatics Journal,27,100532-100532.