ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Management of Tourist Attraction Recommendation and Service Optimization Based on Text Big Data Mining

Xiaopei Zhang[1,*]

[1] College of Business Administration, Zhengzhou University of Science and Technology, Zhengzhou, Henan, 450000, China

SUMMARY: *In order to ensure the satisfaction of tourists with the recommended attractions, a text mining based tourist attraction recommendation algorithm is proposed. The topic model is applied to mine the contents, topics, and keywords of tourist attractions. Meanwhile, the LDA algorithm combining time factor is applied to analyze the personalized needs of users and construct the user model. The cosine similarity is calculated according to the models of tourist attractions and users to help users quickly and accurately find the tourist attractions suitable for their needs from the huge amount of tourist information. The test results show that the algorithm of this paper has a smaller average absolute error than the CF algorithm for different attractions with different numbers of near-neighbor users. When the number of nearest neighbors is 10 and 20, the two algorithms are within 0.6 of each other in prediction. It shows that the algorithm in this paper has high accuracy in recommending tourist attractions, which provides technical support for the subsequent optimization of tourism service management countermeasures.*

KEYWORDS: *text mining; LDA; time factor; cosine similarity; tourist attractions*

1 Introduction

With the continuous expansion of the scale of the tourism market, the growth of people's demand for tourism has exceeded the supply capacity of the traditional tourism model. Under the traditional tourism model, tourists obtain information about attractions through travel agencies, offline publicity, books, friends' recommendations, etc., and such information acquisition channels are severely limited [1]. With the development of Internet technology, the development of social media, professional tourism platforms, etc. has broadened the channels for tourists to obtain attractions as well as the efficiency of access. However, at the same time, in the context of the information age, the synergy between the tourism industry and the Internet has made the tourism industry more and more informatized, and the problem of information overload is becoming more and more prominent, and consumers need to invest a lot of time and energy [2-4]. In addition, some tourist attractions have poor service management, such as homogenization of tourist product services in scenic spots, poor allocation of attraction resources, insufficient ability to respond to the sudden increase in tourist flow, and failure to pay attention to the evaluation and feedback of tourists, which results in a poor tourist experience [5-8]. How to select interested tourist attractions and optimize services will become a more prominent problem. To solve this problem, recommender systems have emerged. So far, the progress of recommender systems has realized a kind of major transformation, which has

*13623824208@163.com
https://doi.org/10.65102/is2026501
Published: 30 April, 2026

shifted from the initial collaborative filtering and matrix decomposition to a brand new research field that nowadays combines multiple disciplines such as big data technology [9]. Combining big data mining with tourist attractions in order to help users select the attractions they are interested in from a large amount of information, personalized tourism planning, tourism traffic prediction and service monitoring are crucial to enhance users' travel experience, and thus have important practical research value and theoretical significance.

Big data technology promotes tourism service enhancement through tourist behavior analysis, tourism market trend prediction, personalized recommendation, and service quality monitoring. Literature [10] analyzes tourism with the help of big data algorithms, proposes an attraction reordering method based on heterogeneous information fusion, and constructs a consumer data analysis system to support smarter travel decisions. Literature [11] used big data technology to analyze tourists' behaviors, analyzed tourists' characteristics, decision-making models and consumption preferences with the help of text mining and network analysis, and proposed targeted tourism marketing, product and management optimization strategies accordingly. Literature [12] proposes an intelligent tourism service decision tree algorithm based on big data technology, which dynamically generates personalized recommendations by analyzing users' location, preferences and external data, accurately matching tourists' changing needs and preferences, and improving service satisfaction and experience. Literature [13] constructed a tourism market trend prediction model based on big data analysis, integrating multi-source data, utilizing the random forest algorithm and parallel computing capability, which is better than the traditional method in terms of prediction accuracy and response speed, and can provide effective support for scientific decision-making. Literature [14] reported that tourism management information systems use big data technology to provide tourists with personalized experiences, optimize resource allocation, deepen the understanding of tourists' behaviors, and promote the sustainable development of the tourism industry by integrating multi-source data. Literature [15] states that big data technology plays a central role in monitoring the quality of tourism services. Through real-time systematic analysis of tourist behavior and operational data, service problems can be accurately assessed and predicted, thus driving continuous improvement and strategy optimization to enhance the overall tourism experience.

In the field of personalized recommendation, literature [16] proposes a hybrid recommendation method combining big data technology and artificial intelligence to build a planning system that goes beyond single attraction recommendation and can tailor a complete itinerary for tourists, in order to promote the development of regional tourism. Literature [17] proposed an enhanced big data analytics model capable of generating more accurate and personalized themed travel recommendations by integrating destination information, user activity reviews and real-time dynamic data through text mining, sentiment analysis and machine learning techniques. Literature [18] focuses on big data mining techniques to optimize the tourist attraction recommendation system, by fusing association rules and collaborative filtering algorithms, and designing dedicated frameworks and page layouts to accurately match users' personalized needs, cope with the challenge of information overload, and achieve intelligent attraction recommendations. Literature [19] applies big data mining technology to improve the project-based collaborative filtering algorithm, and constructs a distributed intelligent recommendation system by analyzing tourists' historical and interactive data in order to optimize the recommendation of tourist attractions and achieve efficient and personalized tourist information services. Nowadays, tourists share tourism information with the help of the Internet, and a variety of textual information contains tourists' needs and evaluations of attractions, from which useful information can be obtained to optimize recommendations and services.

In this paper, we try to add the idea of topic characterization in the content-based recommendation method, the idea of improving the problem of considering the change of user interest, and improve the LDA algorithm and the recommendation method based on text mining considering the change of user interest, which contains two main aspects, one is the content-based recommendation, and the other is the topic analysis. Firstly, the LDA model combined with the time factor is used to mine the user's historical behavior data and analyze the user's behavior for user modeling. Second, the LDA model incorporating the idea of thematic feature analysis is used to mine the text content related to tourist attractions and establish thematic modeling. Again, according to the improved text mining-based tourist attraction recommendation method, the similarity between the tourist attraction topic model and the user interest model is calculated. After sorting the similarity results, the attraction with the largest corresponding similarity is recommended to the user to accomplish the purpose of tourist attraction recommendation.

2 Methodology

2.1 Text mining

Text mining is the process of mining valuable information from textual data and is an emerging method in the field of data mining. Text mining technology provides an effective means for processing, analyzing, integrating and mining text data, which mainly analyzes the characteristics of text content, discovers the potential association patterns existing in text data, and provides users with the required information and knowledge. It is because text mining can mine from text data to discover the implied knowledge, to meet the huge demand for the existence of a large number of unstructured or semi-structured text information. In addition to this, text mining technology also addresses and reduces to a certain extent the human, material and financial resources required for manual processing of text data by human beings. The main techniques involved in the text mining process are shown below.

Compared with commonly used structured data, text data is unstructured and difficult to be recognized and processed by computers. Therefore, it is necessary to convert the text data by preprocessing the text and extract the metadata that can represent its characteristics for structured preservation. Among them, text preprocessing mainly includes word splitting, stop word processing, feature representation and feature extraction.

2.2 LDA Subject Modeling

The LDA topic model is a probabilistic generative model based on a Bayesian model with three levels: a text level, a feature level, and a latent topic level. The structure of the LDA topic model is presented in Fig. 1, with each level associated with its specific parameters. The fundamental assumption of the LDA topic model is that the meaning of words in a text has nothing to do with syntax and their positions in texts, but is dependent only on the number of times each word appears in the text. Topic modeling is primarily used for identifying themes buried in text documents. Topic modeling acts as an analytic tool for extracting meaningful semantic content from huge amounts of text and efficiently processing such texts. Compared with other topic models, LDA exhibits a superior ability to identify hidden topics in text documents. It expresses the thematic structure of each document as a probability distribution over topics, and also describes the linguistic structure of each topic as a probability distribution over words. The LDA topic model can be defined in mathematics as Equation (2), with the explanations for all parameters given in Table 1:

$$P\left(\text{Term}\middle|\text{Document}\right)=P\left(\text{Term}\middle|\text{Subject}\right)P\left(\text{Subject}\middle|\text{Document}\right) \tag{1}$$

To wit:

$$P\left(w\middle|d\right)=P\left(w\middle|t\right)*P\left(t\middle|d\right) \tag{2}$$

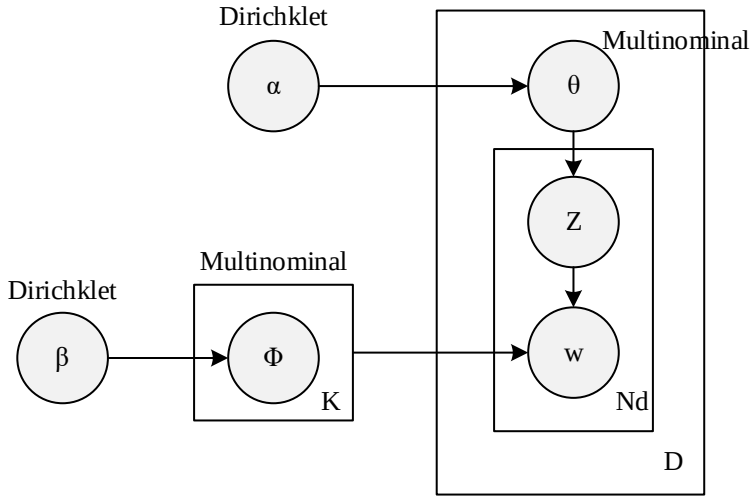


Figure 1: Probabilistic model diagram for the LDA topic model

Table 1: Meaning of LDA theme model parameters

parameter	meaning	parameter	meaning
α	Document - Topic distribution parameters	θ	Document - Topic polynomial distribution
z	The theme of the words in the document	ϕ	Theme - Word Distribution
w	Words in the document	K	Number of themes
D	Number of documents	Nd	The number of words in the document
β	Topic - Word distribution parameters		

Within the LDA topic model, all documents in the text dataset are assumed to share K latent topics. The generative process begins with the formation of a topic distribution for each document and a word distribution for each topic. A topic is then drawn at random in accordance with the document-level topic distribution, and a word is subsequently sampled from the word distribution associated with that selected topic. This procedure is repeated iteratively until a complete document has been generated. The specific steps governing this process are detailed below:

Generate N topic distribution. Let the subject distribution of the ind text be θ_i, which is generated from the Dirichlet distribution with parameter α, so $\theta_i \sim Dirichlet\left(\alpha\right)$. Repeat the operation N times until N subject distributions are generated.

Generate K subject word distribution. Let the distribution of the k nd subject word be ϕ_k, which is generated from the Dirichlet distribution with parameter β, so $\phi_k \sim Dirichlet(\beta)$. Repeat the operation K times to generate K subject word distributions.

Generate the text based on the distribution of topics with the distribution of topic words. The subject Z_{ij} of the j rd word in the i nd text is generated from a polynomial distribution with parameter θ_i, where $Z_{ij} \in Multinomial(\theta_i)$. The j th word $W_{i,j}$ of the i th text is generated from a polynomial distribution with parameter $\phi_{Z_{i,j}}$, where $W_{i,j} \in Multinomial\left(\phi_{Z_{i,j}}\right)$.

The joint probability density function of all variables is:

$$p(\theta,\phi,Z,W|\alpha,\beta)=\prod_{n=1}^{N}p(\theta_n|\alpha)\prod_{k=1}^{K}p(\phi_k|\beta)\prod_{n=1}^{N_m}p(Z_{mn}|\theta_n)p(W_{mn}|Z_{mn},\phi) \tag{3}$$

The text edge probability density function is:

$$p(W|\alpha,\beta)=\prod_{k=1}^{}\int p(\phi_k|\beta)\left[\prod_{n=1}^{N}\int p(\theta_n|\alpha)\prod_{i=1}^{K}\left(\sum_{i=1}^{K}p(Z_{mn}=1|\theta_n)p(W_{mn}|\phi_i)\right)d\theta_n\right]d\phi_k \tag{4}$$

For the parameters in the LDA model, approximate estimates can be obtained by Gibbs sampling method, and the computational steps are as follows:

Step 1: First determine the number of topics K, assign values to hyperparameters α & β, then randomly assign topics to each text word.

Step 2: Calculate the probability of each word in the text using equation (5):

$$p(Z_i|Z_{-i},W,\alpha,\beta)=\frac{\beta_v+n_{k,v}}{\sum_{v=1}^{V}(\beta_v+n_{k,v})}\cdot\frac{\alpha_k+n_{m,k}}{\sum_{k=1}^{K}(\alpha_k+n_{m,k})} \tag{5}$$

where $n_{m,k}$ represents the number of words assigned to the k rd topic in text d_m and $n_{k,v}$ represents the number of words assigned to the k th topic in the text dataset.

Step 3: Repeat step 2 until convergence.

Step 4: Count the subject of each word in the text to get the distribution of the subject of each text θ_{mk} and the distribution of each subject word ϕ_{kv}, where θ_{mk} and ϕ_{kv} are calculated as shown in equation (6):

$$\theta_{mk}=\frac{\alpha_k+n_{m,k}}{\sum_{k=1}^{K}(\alpha_k+n_{m,k})}$$
$$\phi_{kv}=\frac{\beta_v+n_{k,v}}{\sum_{v=1}^{V}(\beta_v+n_{k,v})} \tag{6}$$

2.3 Text Cluster Analysis

The initial problem in carrying out text clustering is that of finding the best way to represent the textual data, or in other words, finding a good way to represent text data in terms of computable entities. The standard technique used in text representation is the vector space model, where each document is mapped into a vector form. Various methods may be used to populate the vector based on the needs of analysis. The two most common ways of weighting the vector are TF and TF-IDF, where each element of the vector is assumed to correspond to a unique word in the vocabulary. If there are no labels in the dataset, clustering becomes necessary.

The core idea of the K-means algorithm is to cluster the dataset $T=\{a_1,a_2,\cdots,a_n\}$ into k categories, and the criterion for judging the superiority or inferiority of each category $class=\{class_1,class_2,\cdots,class_k\}$ is to minimize the squared error and SSE, with the following formula:

$$SSE=\sum_{i=1}^{k}\sum_{a\in c_i}\left\|a-center_i\right\|^2 \tag{7}$$

where, $center_i$ is the cluster center of the i nd cluster category, the formula describes the closeness of the intra-cluster clusters in each cluster to the center of that cluster, if the value is lower it indicates that it is closer within that cluster. In addition to this, the meaning of this formula is to calculate the sum of the Euclidean distances (8) of each cluster class distance to the cluster center of that cluster.

$$dist\left(d_i,d_j\right)=\sqrt{\sum_{k=1}^{K}\left(Z_{ik}-Z_{jk}\right)^2} \tag{8}$$

Specific steps of K-means algorithm:

Step 1: Select K initial center.

Step 2: Compute the Euclidean distance between each sample point and the K initial cluster centers, then assign every point to the class corresponding to the nearest center. Proceed to calculate the distances from the remaining sample points to the k centers and allocate each of them to their respective classes.

Step 3: Compute the mean value of all data points that belong to the same cluster and shift the cluster centroid to the new mean position.

Step 4: Repeat steps 2 and 3 sequentially until the clustering result meets the specified criterion of convergence.

The silhouette coefficient is a frequently used method for evaluating the performance of a clustering algorithm. It acts as a numerical index for quantifying the effectiveness of the clustering operation on determining the boundary values of each cluster. The silhouette coefficient combines both cohesion and separation parameters. Cohesion measures how much a sample point belongs to the other data points within the same cluster, whereas separation indicates the distance between a sample point and other data points from adjacent clusters. The mathematical representation of the silhouette coefficient is shown below:

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{9}$$

where, $a(i)$ represents the degree of cohesion of the sample points and the formula for calculating the degree of cohesion is shown below:

$$a(i) = \frac{1}{n-1}\sum_{j \neq i}^{n} dis\tan ce(i, j) \tag{10}$$

In this expression, j denotes any sample point belonging to the same cluster as sample i, and distance refers to the measured separation between sample point i and sample point j. A smaller value of $a(i)$ signifies that sample i sits more compactly within its assigned cluster. The quantity $b(i)$ is obtained through a procedure analogous to that used for $a(i)$, with the distinction that iteration must be carried out across all remaining clusters to produce a collection of values $\{b_1(i), b_2(i), b_3(i), \cdots, b_m(i)\}$, from which the minimum is selected as the final result. On this basis, $S(i)$ can be formulated as follows:

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & a(i) < b(i) \\ 0 & a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & a(i) > b(i) \end{cases} \tag{11}$$

From the above equation, it can be seen that the contour coefficient $S(i)$ has a value range of $[-1,1]$. The larger the contour coefficient the better the clustering effect.

2.4 Recommendation Algorithm for Tourist Attractions Based on Text Mining

Recommendation approach is the most vital part of a recommendation system since it has a significant impact on the overall performance of such a system. In this case, we use the idea of topic features analysis in the context of content-based recommendations. We can create a model that is based on deep analysis of text topics and relevant content of tourist spots by applying an enhanced LDA model. We will use this model to build a model of interests that will allow us to represent users' interests accurately. An attraction model will be created independently, and then we will measure similarity between the attraction model and the interest model in order to develop the tourist attraction generation model.

Personalized tourist attraction recommendation technology is based on two main principles: content-based recommendation and topic analysis.

The procedure starts with the text mining of the tourist attractions, followed by the construction of topic models using the LDA probabilistic topic model. The second step deals with the modeling of each individual attraction based on their themes, while at the same time analyzing the behavior of users in order to create user models. In the third step, the similarity between the two models mentioned above is computed. Based on the similarity values calculated during the third step, the fourth step involves ranking all the attractions based on their similarity values, and choosing the top k attractions which are most similar to the user model. In this manner, the system recommends the tourist attractions to the users which they will enjoy visiting. The entire process of recommending tourist attractions based on text mining

is illustrated in Fig. 2. The proposed personalized recommendation scheme can be considered based on three key aspects, namely the building of the user interest model, the tourist attraction model and finally the recommendations.

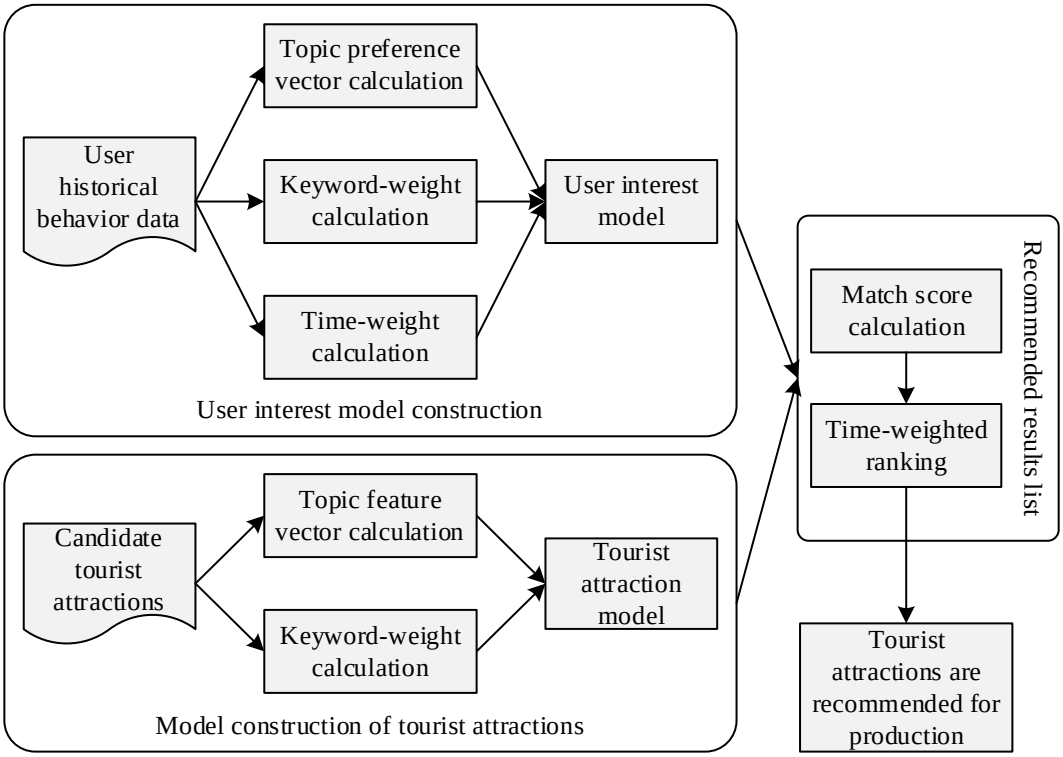


Figure 2: Tourism Record Recommendation Framework Based on Text Mining

(1) Constructing the user model

The user model of the personalized recommendation technique for tourist attractions based on text mining used in this paper can be expressed as Equation (12):

$$F_u = \{T_{ui}, K_{ui}\} \tag{12}$$

T_{ui} and K_{ui} denote the user's topic preference vector and the sequence of keyword weights, respectively, and the keywords are subordinate to the topics.

The user's topic interest preferences are represented by a set of weight vectors as Eq. (13):

$$T_u = \{\omega_{u1}, \omega_{u2}, \cdots, \omega_{um}\} \tag{13}$$

where $\omega_{ui}\left(1 \le i \le m\right)$ denotes the user's interest preference for the i nd topic and m is the total number of topics.

A user's keyword interest is represented by a set of keyword weight sequences as in Equation (14):

$$K_u = \{(k_{u1}, \omega_{u1}), (k_{u2}, \omega_{u2}), \cdots, (k_{un}, \omega_{un})\} \tag{14}$$

where, $k_{uj}\left(1\le j\le n\right)$ denotes the keyword that the user is interested in, and w_{uj} denotes the degree to which the user is interested in keyword K_{uj}.

(2) Constructing tourist attraction model

The tourist attraction model of the improved text mining-based personalized recommendation technique for tourist attractions in this paper can be expressed as in Equation (15):

$$F_s=\left\{T_s',K_s'\right\} \tag{15}$$

The theme feature vector T_s' of the tourist attraction content model is calculated based on the theme model as in Equation (16):

$$T_s'=\left\{\omega_{s1},\omega_{s2},\cdots,\omega_{sm}\right\} \tag{16}$$

Each dimension $\omega_{si}\left(1\le i\le m\right)$ in the vector represents the weight of the tourist attraction in the i nd topic and m is the number of topics.

The keyword weights K_s' of the tourist attraction content model have a set of weight vectors expressed as equation (17):

$$K_s'=\left\{\left(k_{s1},\omega_{s1}\right),\left(k_{s2},\omega_{s2}\right),\cdots,\left(k_{sn},\omega_{sn}\right)\right\} \tag{17}$$

$k_{sj}\left(1\le j\le n\right)$ denotes the j nd keyword of the tourist attraction, w_{sj} corresponds to the weight of keyword k_{sj}, and n is the number of keywords.

The weights of keywords ω are shown by equation (18):

$$\omega(k,t)=\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t}=\frac{n_k^{(t)}\cdot Q_u(t)+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \tag{18}$$

where $n_k^{(t)}$ is shown by Eq. The change of the time influence factor considered by the change parameter, i.e., the decay of the user's interest, the time change factor is shown by Eq. And this time influence factor conforms to the Ebbinghaus forgetting curve.

(3) Tourist attraction recommendation generation

The decision of whether a particular tourist attraction is recommended to a given visitor rests fundamentally on the degree of alignment between that visitor's established preferences and the content profile of the attraction in question.

For the tourists in the system, after the user interest model $F_u=\left\{T_{ui},K_{ui}\right\}$ and the tourist attraction model $F_s=\left\{T_s',K_s'\right\}$ in the system are calculated, the matching degree between them is calculated:

Firstly, the cosine similarity between the two is calculated as in Equation (19) based on the user's topic preference vector T_{ui} and the tourist attraction's topic feature vector T_s':

$$sim\left(T_{ui},T_s'\right)=\frac{T_{ui}\times T_s'}{\left\|T_{ui}\right\|\cdot\left\|T_s'\right\|} \tag{19}$$

Secondly, the keyword weight K_{ui} of the user's historical browsing and the sequence of keyword weights of the tourist attractions in the system are subjected to Jaccard similarity calculation as in Equation (20):

$$sim\left(K_{ui}, K_s'\right) = \left| \frac{K_{ui} \cap K_s'}{K_{ui} \cap K_s'} \right| \tag{20}$$

That is, the degree of similarity between the sequence of keyword weights of user preferences K_{ui} and the sequence of keywords K_s' contained in the text of tourist attractions is compared.

The match between the tourist attraction model and the user interest model is calculated as in Equation (21):

$$sim\left(F_u, F_s\right) = \frac{\sigma \cdot sim\left(T_{ui}, T_s'\right) + \rho \cdot sim\left(K_{ui}, K_s'\right)}{\sqrt{\sigma^2 + \rho^2}} \tag{21}$$

σ, ρ are set scale coefficients for adjusting the similarity of theme features and the similarity of keyword weights sequences. If the result is greater than the specified threshold, the tourist attraction is recommended to the tourist.

To preserve the temporal relevance of the personalized tourist attraction recommendation system, timeliness weights are assigned to each attraction appearing in the user's initial recommendation list. These weights are computed in accordance with Equation (22):

$$W\left(F_u, F_s\right) = \varepsilon \log_2 sim\left(F_u, F_s\right) + \frac{(1-\varepsilon)\Delta t}{2592000} \tag{22}$$

where $sim\left(F_u, F_s\right)$ is the match value between the new tourist attraction located in the user's initial tourist attraction recommendation list and the user as calculated by Equation (20), ε is the pre-set balancing coefficient, and Δt represents the elapsed time between the point at which the tourist attraction information was published and the present moment. The constant 2592000 corresponds to the total number of seconds contained within a standard 30-day month, calculated on the basis of 24 hours per day, and the match value of the two tourist attractions with the user is equal, and if the difference in their release times is 24 hours, then the difference in their timeliness weights is $\left(1-\varepsilon\right)$.

3 Results

3.1 Experimental data

The dataset utilized for this experiment involves travel data extracted by means of web crawling from the Ctrip travel website. Before the implementation of the experimental analysis, raw data collected via the web crawler required some preprocessing operations, including sorting, cleaning, and dividing into sub-datasets.

Crawling of data: in view of the cost, tediousness and missed catches of crawler tools, and the written Python crawler code crawling is convenient and efficient, therefore, the experimental dataset of this thesis uses the written Python crawler code for crawling the data

of tourist attractions, and the crawled tourist website is Ctrip travel website. A total of 1.2 million raw tourist attraction data information was crawled.

Data Sorting: The data concerning tourist attractions derived from the web crawler included a number of irregularities and redundancies that did not contribute to further analysis. In other words, all these pieces of useless data were eliminated, such as invalid user ratings, low-frequency user comments, and duplicated review records. After being sorted out, the dataset involved 16,385 registered users and 10,716 tourist attractions where each of the users has reviewed no less than three tourist attractions. Each row of data includes several items, such as registered usernames, IDs of users, IDs of tourist attractions, names of tourist attractions, numerical ratings of each tourist attraction given by users, and text comments. The summary of the experimental dataset is presented in Table 2.

Table 2: Experimental dataset

Data Source	Number of users	The number of scenic spots	The number of ratings	Number of comments	Sparsity
Tourist cities	16385	10716	17240	19231	99.89%

Data Division: To validate the effectiveness of the proposed algorithm, the preprocessed experimental dataset was randomly divided into two sub-datasets – the training sub-dataset and the testing sub-dataset. The former would be applied to training the model algorithm suggested in this paper, whereas the latter would be applied to evaluating the performance of the competing algorithms. According to the division of 8:2, 80% of the sorted travel data was utilized for training purposes, while the rest 20% was used to assess the performance of different algorithms.

3.2 Descriptive statistical analysis of data

(1) Preprocessing of data

The sample data acquired in this paper are all Chinese text data. The preprocessing operation of the attraction text data mainly includes three parts: attraction text data word division, de-discontinued words, and extraction of feature words.

Taking a scenic spot in X province as an example, the length information of each sample after preprocessing, the distribution of tourist attraction text length is shown in Figure 3. The difference in the number of feature words of each sample is relatively large. The shortest text contains only 164 feature words, while the longest text contains 10683 feature words. Observing the distribution of text length, it can be found that although the texts with the number of feature words exceeding 3000 contain more detailed descriptions, the number of texts in this category is relatively small. The length of texts describing tourist attractions in X province is mainly concentrated in the range of [100, 5000]. This also shows that the text-word matrix is a highly sparse matrix.

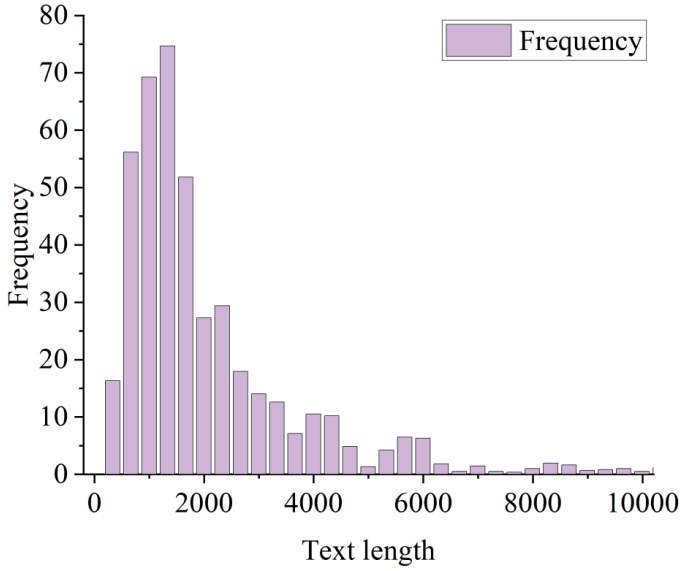


Figure 3: Text length distribution of tourist attractions

(2) Overall Overview of Attractions

Province X is located in the northwestern part of China, and tourism is booming as one of the important economic sources. The website recorded a total of 564 tourist attractions in Province X with grade A and above, and the number of processable attraction texts was 395, excluding those with only the name of the attraction but no specific description.

The distribution of Grade A and above tourist attractions in Province X is shown in Fig. 4, where a~f represent the 10 cities under Province X. The distribution of Grade A attractions in Province X shows a “ladder-like” change. Firstly, city g is located in the first ladder, which not only has more tourist attractions than the other nine cities, but also has more attractions at each level than the other municipalities. This confirms that city g, as the capital city of the province, has a greater advantage in the allocation of tourism resources. On the second rung of the ladder are the four cities b, c, a, and j, which are the neighboring cities of city g. The number of tourist attractions is higher than that of the other cities. The number of tourist attractions is less than that of city g, and the number of tourist attractions at each level is less than that of city g. Located on the third rung of the ladder are the four cities d, h, f, and i. The number of tourist attractions is the next highest. The last rung of the ladder is city e, which has the least number of tourist resources in terms of their allocation.

In addition, analyzing the proportion of various scenic spots can reveal that the one with the highest proportion is "3A-level tourist attractions", accounting for 52.15%. This type of attraction is the most numerous. Following it are "4A-level tourist attractions", "2A-level tourist attractions", "5A-level tourist attractions" and "A-level tourist attractions", with proportions of 37.72%, 6.33%, 3.29% and 0.51% respectively. From the above analysis, it can also be seen that Province X has abundant high-quality tourism resources, but their distribution is uneven.

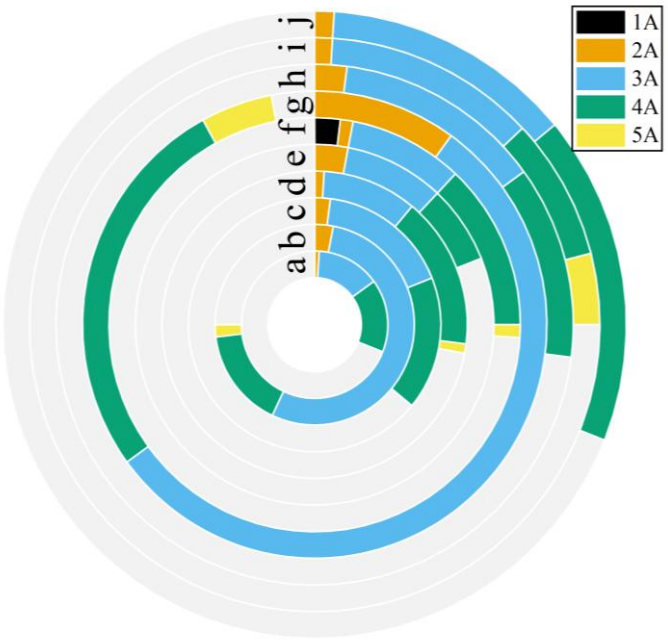


Figure 4: Distribution Map of A-level and above Tourist Attractions in Province X

(3) LDA topic modeling

When doing text analysis, as the number of topics needs to be set in advance, the number of topics set will directly affect the final model results.

The formula for calculating the average cosine similarity of themes is shown in (23) and (24). Where $similarity\left(\phi_i,\phi_j\right)$ denotes the similarity between the i nd topic and the j rd topic, $average-similarity\left(\phi\right)$ denotes the average cosine similarity between all topic vectors, ϕ_i denotes the word distribution vector of the i th topic and $\phi_i=\left(\phi_{i1},\cdots,\phi_{iV}\right)$, K denotes the number of topics, V denotes the number of words. The cosine distance is used in this paper to describe the similarity between two topics because it has a definite range of values and indicates the difference in the direction of the text data. The closer the cosine value is to 0, the greater the angle between the two topics, indicating the greater the difference between the two topics. The closer the cosine value is to 1, the closer the angle between the two themes is to 0, indicating a greater degree of similarity between the two themes. In order to emphasize the representativeness of different themes, it is therefore required that the similarity between different themes is as small as possible. The optimal number of themes can be determined according to the principle that the smaller the average cosine similarity between themes, the more representative the theme is.

$$similarity\left(\phi_i,\phi_j\right)=\frac{\sum_{v=1}^{V}\phi_{iv}\phi_{jv}}{\sqrt{\sum_{v=1}^{V}\phi_{iv}^2\sum_{v=1}^{V}\phi_{jv}^2}} \tag{23}$$

$$average-similarity\left(\phi\right)=\frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}similarity\left(\phi_i,\phi_j\right)}{K\times\left(K-1\right)/2} \tag{24}$$

The mean cosine similarity curve of themes is shown in Fig. 5, in the process of changing the mean cosine similarity curve of themes, when the number of themes changes from 1 to 2, the mean cosine similarity of themes decreases rapidly, and falls to a lower level when the number of themes is 4. At this time, if the number of themes is increased again, the range of variation of the mean cosine similarity of themes is very small. Then according to the variation curve of the average cosine similarity of themes, we determine that the optimal number of themes is 4.

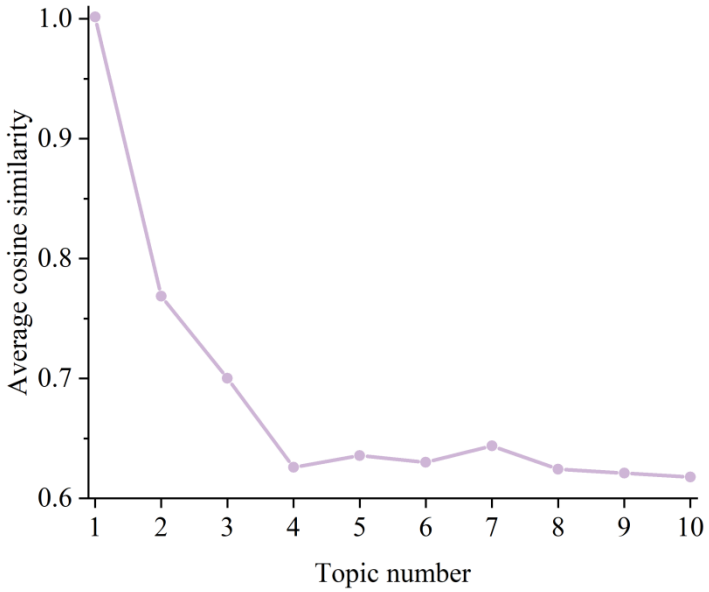


Figure 5: The average cosine similarity curve of the subject

After determining the number of topics, the LDA topic model is created and analyzed. As a result, two types of distributions are obtained: topic distribution of attractions and word distribution in topics. In the topic model, the parameter ϕ determines the distribution of words among individual topics. If the word appears more often in one topic than another, it implies its semantic relationship with the meaning of the topic. Therefore, for each topic, this paper outputs the top 8 associated words with higher probability, and the top8 associated words for each topic are shown in Table 3. The associated words related to theme 1 are all related to war, revolution, and red tourist attractions. The association words related to theme 2 are all related to history, religion and culture. Associated words related to theme 3 are related to natural attractions. Associated words related to theme 4 are mostly related to leisure and entertainment programs.

Table 3: Top8 related words

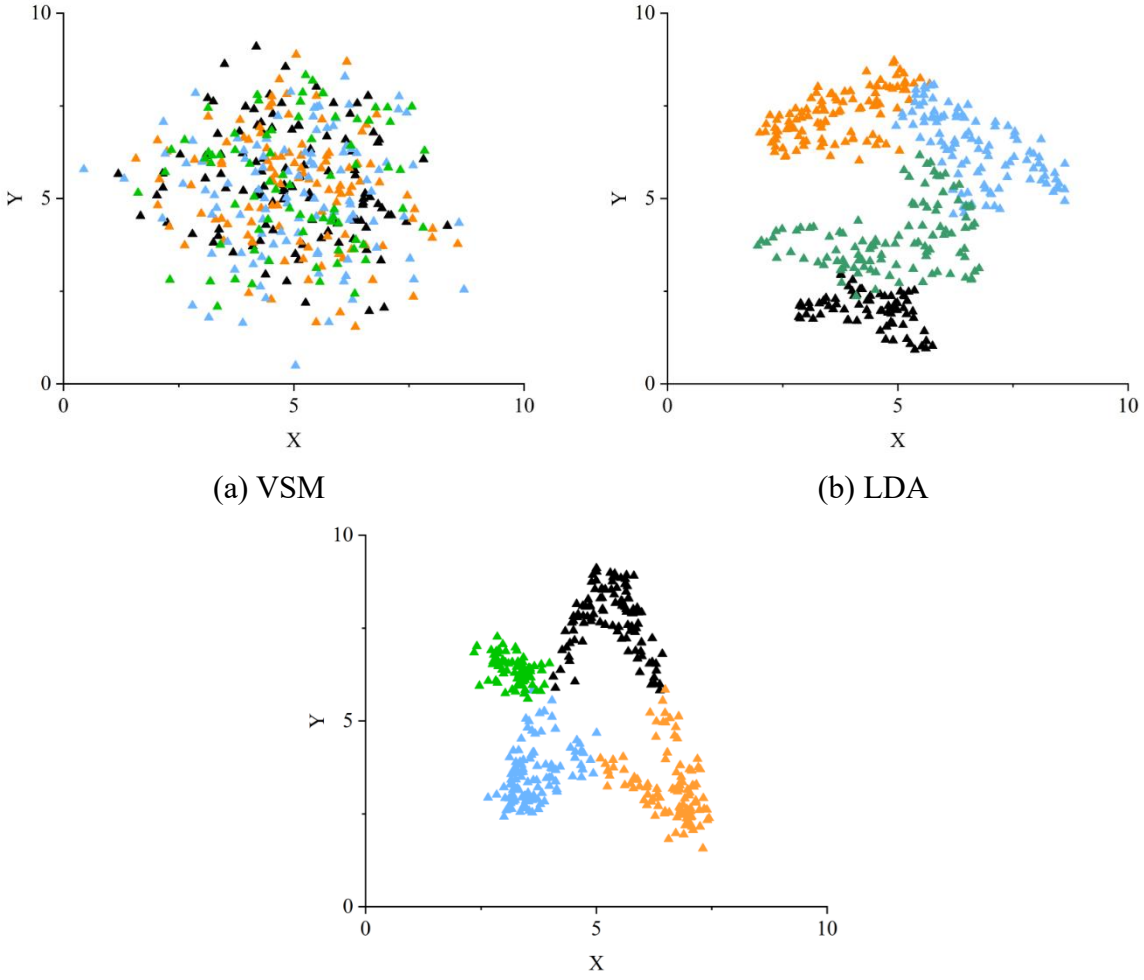
Category	Key words
Theme 1	Revolution, memorial, yan 'an, museum, architecture, China, old site, suidd
Theme 2	Buildings, sites, cultural relics, museums, China, tang dynasty, stone carving and cultural relics protection
Theme 3	Qinling, huashan, waterfalls, forests, xi 'an, altitude, park, natural
Theme 4	Leisure, ecology, project, construction, xi 'an, experience, park and wetland

(4) Comparison of text clustering models

The k-means clustering was performed on the distribution of potential themes of the text, and the 395 attractions in Shaanxi Province were categorized into four classes. The final clustering results are shown in Figure 6. Figure (a) represents the clustering result graph based

on VSM model. Figure (b) shows the clustering result based on LDA theme model. Figure (c) represents the clustering result graph based on LDA topic model after selecting the optimal number of topics based on the average cosine similarity of topics.

From the figure, it is concluded that the clustering effect becomes gradually better from top to bottom. The clustering effect in graph (a) is the worst, and all kinds of attractions are disorganized without any regularity. The clustering effect of the graph in Figure (b) is significantly better than the left graph. At this point, different categories of attractions can be distinguished from each other, but the distance within the same category of attractions is far apart. The clustering effect of graph (c) is the best, the attractions of different categories are well distinguished from each other and are compact within the same category.



(a) VSM (b) LDA

(c) Average cosine similarity + LDA

Figure 6: Comparison diagram of clustering results

The results are compared, and the findings are presented in Table 4. The results show that the clustering based on the VSM model generates the worst outcomes, which include the lowest value of the silhouette coefficient of 0.0058 and the largest within-cluster sum of squares errors of 387.2290. On the other hand, the clustering based on the LDA topic model shows significantly better outcomes in comparison to the previous method. In the LDA clustering model, the silhouette coefficient equals 0.2510, and the within-cluster sum of squares error is 82.4177. The best clustering outcome is generated when the optimal number of topics is found using average cosine similarity, and its intra-cluster error sum of squares is the smallest, which

is 23.0584. And the contour coefficient is the largest, which is 0.6637. So the clustering result of the third way is selected in the subsequent analysis process.

Table 4: Comparison of results of each model

Model	Contour coefficient	Error of the sum of errors in the cluster
Clustering based on the VSM model	0.0058	387.2290
Clustering based on LDA theme model	0.2510	82.4177
Clustering based on LDA theme model (Theme cosine similarity selection theme number)	0.6637	23.0584

3.3 Analysis of experimental results

This experiment compares and analyzes the improved and optimized LDA topic-weighted text mining-based algorithm model and the traditional collaborative filtering algorithm (CF algorithm) model, and adopts the MAE (mean absolute error), which is commonly used in the industry, as the evaluation criterion for the comparison of algorithms.

Since the dataset contains both user ratings and text from the user comments, an evaluation of the performance of the algorithm suggested in this paper relative to the collaborative filtering algorithm with different amounts of nearest users has been performed in order to provide a comprehensive analysis of their performance.

Mean absolute errors calculated using the algorithms with different amounts of nearest users have been provided in Table 5 below. As can be seen from the results, in all cases the mean absolute error of the proposed algorithm is less than that of the algorithm used for comparison, which implies that the accuracy of predictions made using the algorithm presented in this paper is greater. For situations where the amount of user nearest neighbors is low, the proposed algorithm benefits from the information contained in the user comments on attractions, thus producing high-quality predictions when traditional methods struggle.

Table 5: The mean absolute error (MAE) of the two algorithms

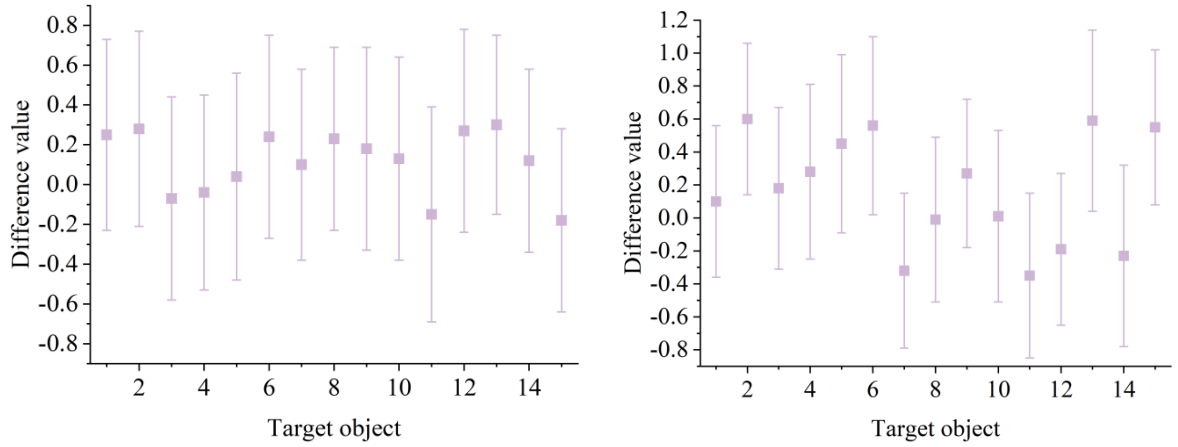
The number of neighboring users	The algorithm in this paper	CF·Algorithm	Improve
3	1.186	1.193	0.007
6	1.179	1.192	0.013
9	1.179	1.189	0.010
12	1.177	1.193	0.016
15	1.175	1.194	0.019
18	1.174	1.191	0.017
21	1.167	1.178	0.011
24	1.165	1.182	0.017
27	1.164	1.171	0.007
30	1.162	1.176	0.014

Similarly, the CF algorithm is used again as the standard basis for comparison for the analysis of attraction ratings and review text data. In order to allow for a comparison between the two algorithms, mean absolute errors are then estimated at various numbers of nearest neighbor attractions. The resulting MAE estimates at various nearest neighbor numbers are then presented in Table 6. From the table, it is observed that the mean absolute errors from the proposed algorithm are smaller compared to those from the CF algorithm, suggesting that the proposed algorithm produces more accurate rating predictions.

Table 6: The average absolute error under different numbers of neighboring scenic spots

The number of neighboring users	The algorithm in this paper	CF Algorithm	Improvement
3	1.162	1.179	0.017
6	1.160	1.179	0.019
9	1.160	1.173	0.013
12	1.158	1.174	0.016
15	1.156	1.176	0.020
18	1.149	1.166	0.017
21	1.145	1.147	0.002
24	1.144	1.151	0.007
27	1.140	1.153	0.013
30	1.140	1.144	0.004

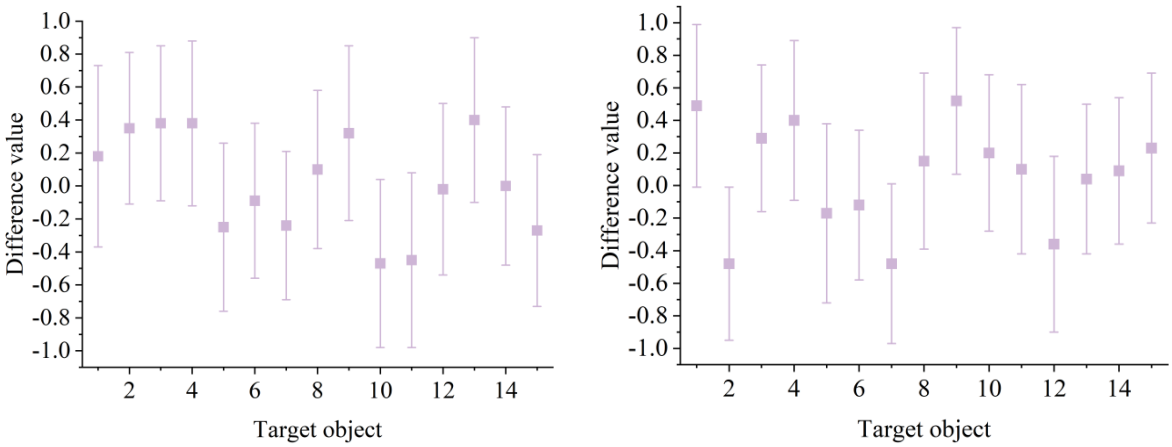
Assuming that both algorithms produce equally accurate rating predictions, a further comparison between the rating predictions outputted by both algorithms are conducted. Here, using attraction ratings and comment text data, and with the nearest neighbor numbers at 10 and 20 respectively, the difference in rating predictions of the top 10 target attractions recommended by both algorithms is calculated. Prediction errors between the two algorithms are then summarized in Fig. 7, where Fig. (a) and Fig. (b) present the results at nearest neighbor numbers of 10 and 20 respectively.



(a)The number of close neighbors is 10 (b)The number of close neighbors is 20

Figure 7: Comparison of error in prediction scores between the two algorithms

In another parallel analysis conducted through the use of attraction rating and review text data, the prediction error between the predicted attractions of the two algorithms is calculated based on the 10 attractions that have the highest recommendation value under the nearest neighbor numbers of 10 and 20. The prediction score error comparison is shown in Fig. 8, with Fig. (a) and Fig. (b) showing the results under nearest neighbor numbers of 10 and 20, respectively.



(a)The number of close neighbors is 10 (b)The number of close neighbors is 20

Figure 8: Comparison of error in prediction scores between the two algorithms

Through an analysis of Figs. 7 and 8, it can be seen that the prediction error differences between the proposed algorithm and the CF algorithm, under both nearest neighbor number of 10 and 20, still remain relatively small, since the highest prediction error is always within 0.6. From this, we can see that the optimized algorithm proposed in this paper is no different from the CF algorithm in predicting tourist attraction recommendations according to the prediction value difference. However, the MAE analysis provides strong proof of the high accuracy in recommendation made by the improved algorithm. All the tests carried out above validate the effectiveness as well as the accuracy of the proposed recommendation algorithm. As seen above, the mean absolute error in the prediction made by the proposed algorithm is smaller than that in the CF algorithm when the nearest neighbor number increases. Besides, there is an inherent advantage of the proposed algorithm compared to the CF algorithm due to the comprehensive use of text data included in user reviews, which helps compensate for the lack of numerical data and improves the recommendation accuracy.

4 Optimizing tourism service management countermeasures

(1) Government level

The government further strongly supports the application of data in tourism. First, improve the infrastructure construction. Further improve the construction of network coverage in tourism areas, provide high-speed and convenient network services for tourists, and provide basic protection for the collection of big data. Second, policy instruments need to be introduced in order to guide and assist in the integration of big data with the physical tourism business, providing institutional support for the fusion of big data technology with the tourism economy to create a synergy capable of propelling the whole industry forward. Fourth, it is necessary to recruit outstanding talents who are familiar with the big data industry in order to establish the base of human resources that will ensure continuous innovation in big data technology.

(2) Enterprise Level

It is incumbent upon tourism enterprises to become involved in the implementation of big data technology. Agencies that offer tours, hotels, eateries, shops selling products for tourists, transportation providers, scenic spots, and amusement parks are all within the domain of this responsibility. They need to work positively with government policies to ensure the effective implementation of the policy directive that encourages the fusion of big data with real-sector companies. More importantly, these tourism enterprises need to appreciate the importance of

big data technology in changing the course of tourism. The best approach to ensuring sustainability in their operations is to utilize big data technology to cut down costs, improve marketing strategies, and offer quality services to tourists.

(3) Tourist Level

Actively guide tourists' cognition and participation in big data practice. Through various ways to guide tourists to participate in the practice of big data application, you can organize various activities through microblogging, WeChat public number and other platforms to encourage the participation of tourists and give them certain rewards. Or through WeChat QR code and other ways to guide tourists to actively participate in the tourism service quality evaluation platform. Through the accumulation of huge amounts of data obtained from tourists' engagement in activities and their feedback, the system is able to analyze and use the data obtained to come up with better strategies for improving service quality and thus improve tourism service quality within Guizhou province.

5 Conclusion

The attraction recommendation algorithm plays a very important role in the technological framework that is used in managing attractions. The algorithm helps in maximizing visitors' entry at the attraction sites while at the same time making sure that the promotional purpose of the attraction sites is fully met and therefore offering better recommendation experience. In this paper, we propose a tourist attraction recommendation model using text mining technologies. In terms of user rating and review text, the mean absolute error that results from the proposed algorithm is smaller than that of the CF algorithm by an average of 0.007~0.019. In terms of attraction rating and review text, the mean absolute error achieved using the proposed algorithm is lower than that of the CF algorithm by 0.007~0.020. This paper's algorithm can achieve the best recommendation effect for the tourists based on the matching of users' preferences and attraction themes, while guaranteeing the satisfaction and accuracy of tourists. While ensuring the satisfaction of tourists and the accuracy rate, it realizes the reliable recommendation of attractions and greatly improves the management level of attractions.

About the Author

Xiaopei Zhang, was born in Yuzhou, Henan, P.R. China, in 1985. She received the Master degree from Southwest Jiaotong University. Now, she works in College of Business Administration, Zhengzhou University of Science and Technology. Her research interests include rural tourism and tourism area plan.

References

[1] Jovicic, D. Z. (2019). From the traditional understanding of tourism destination to the smart tourism destination. Current Issues in Tourism, 22(3), 276-282.

[2] Zhou, X., & Chen, W. (2021). The impact of informatization on the relationship between the tourism industry and regional economic development. Sustainability, 13(16), 9399.

[3] Zhong, D., Wang, Y., Wang, L., Sun, Q., & Wang, M. (2025). Information overload in digital tourism marketing: Challenges and opportunities for enhancing purchase intentions. Information Development, 02666669251334136.

[4] Liu, J. (2025). The Impact of Information Overload on Preferences for Tourism Information Media. Finance & Economics, 1(6).

[5] Rasethuntsa, B. C. (2021). Service delivery in the Lesotho tourist attraction sector: are tourists satisfied with the service provision?. Journal of Business and Management Review, 2(10), 677-691.

[6] Romanyuk, A. V., & Gareev, R. R. (2020). The system of indicators for assessing the effectiveness of the regions in the field of tourist services in Russia: Key problems and solutions. Journal of Environmental Management & Tourism, 11(6), 1347-1367.

[7] Shao, T., Yang, P., Jiang, H., & Shao, Q. (2023). An analysis of public service satisfaction of tourists at scenic spots: The case of xiamen city. Sustainability, 15(3), 2752.

[8] Wu, Y., & Wang, Y. (2024). An empirical study on the tourist cognitive evaluations of tourism public services in Xinjiang province, China. Sustainability, 16(5), 1712.

[9] Sarkar, J. L., Majumder, A., Panigrahi, C. R., Roy, S., & Pati, B. (2023). Tourism recommendation system: A survey and future research directions. Multimedia tools and applications, 82(6), 8983-9027.

[10] Yang, J., Zheng, B., & Chen, Z. (2020). Optimization of tourism information analysis system based on big data algorithm. Complexity, 2020(1), 8841419.

[11] Li, J., & Cao, B. (2022). Study on tourism consumer behavior and countermeasures based on big data. Computational Intelligence and Neuroscience, 2022(1), 6120511.

[12] Hu, H., & Li, C. (2023). Smart tourism products and services design based on user experience under the background of big data. Soft Computing, 27(17), 12711-12724.

[13] Li, T. (2024, December). Construction of Tourism Market Trend Forecasting Model Based on Big Data Analysis. In 2024 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC) (pp. 532-536). IEEE.

[14] Wang, L. (2024). Enhancing tourism management through big data: Design and implementation of an integrated information system. Heliyon, 10(20).

[15] Li, C., & Wen, X. (2025). Big Data-Driven Smart Tourism Service Quality Monitoring and Enhancement Strategies. International Journal of Management Science Research, 8(1), 22-29.

[16] Al Fararni, K., Nafis, F., Aghoutane, B., Yahyaouy, A., Riffi, J., & Sabri, A. (2021). Hybrid recommender system for tourism based on big data and AI: A conceptual framework. Big Data Mining and Analytics, 4(1), 47-55.

[17] Asaithambi, S. P. R., Venkatraman, R., & Venkatraman, S. (2023). A thematic travel recommendation system using an augmented big data analytical model. Technologies, 11(1), 28.

[18] Wan, Y. (2022, October). Tourism intelligent recommendation system based on big data mining. In Proceedings of the 2022 6th International Conference on Electronic

Information Technology and Computer Engineering (pp. 1040-1044).

[19] Longlong, L. (2024, June). Design and research of smart tourism recommendation system based on big data mining technology. In 2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII) (pp. 794-797). IEEE.