DOI: 10.23977/tmte.2023.060304
ISSN 2616-2199 Vol. 6 Num. 3

Research on the Decision Making of Jingyuetan Scenic Spot Based on Social Network Data Processing

Yuan Long[1], Ning Wang[1], Xingmei Xu[2,*]

[1]College of Data Science, Guangzhou Huashang College, Guangzhou, 511300, China
[2]College of Information Technology, Jilin Agricultural University, Changchun, 130118, China
[*]Corresponding author

Keywords: SNS, FDCD-TFIDF, Jingyuetan

Abstract: In the interpretation of the concept of smart scenic spot, this article takes Changchun Jingyuetan scenic spot as an example, and puts forward the accurate suggestion of obtaining smart scenic spot from tourists by using social network (SNS) data. This paper uses a series of data mining technology, through obtaining China more influential tourist communication platform, Meituan, ctrip and weibo comment data, collect Jingyuetan scenic spot tourists accurate ridicule, characteristic word frequency analysis and clustering grid analysis, and summarizes the wisdom of Jingyuetan scenic spot environment, scenic area service, scenic area traffic three aspects put forward the scientific and effective rectification reference Suggestions and empirical research. This research provides a new idea and an important reference for the precise improvement research of smart scenic spots around the world, and is of great significance for the construction of smart scenic spots.

1. Introduction

With the popularization of mobile devices and the rapid development of Internet and Internet technology, the personalized information and data services such as Internet scenic spots and hotels are increasing. Faced with the problem of information and data accumulation, how to find deeper connections and laws from these data resources, rather than superficial connections and laws has become a hot issue for scholars at home and abroad. In the theoretical research, Chen Jianbin [1], Zhang Lingyun [2] and others have analyzed the essence of smart scenic spots from different dimensions, and put forward a general framework. However, in practical application, due to the deviation and lack of understanding and lack of understanding, the phenomenon of "emphasizing hardware over software" is easy to appear in the process of smart tourism construction, resulting in a waste of resources or inefficient [3] service. At present, the relevant research results on smart scenic spots at home and abroad are not comprehensive and in-depth enough.

In the past, the research of tourists in scenic spots has focused on behavior patterns, spatial and space distribution, perceived value, passenger flow analysis, and tourist experience, with very few models extracting information from social networks (SNS). There is no unified and standardized way to measure the evaluation of various facilities and environmental experience in the smart scenic area. This makes some differences in tourist perception of different types of tourist destinations. Thus leading to tourism decision-making mistakes. Therefore, this paper chooses to analyze and study the online comments on various tourism social platforms from the perspective of tourists.

2. Related Reviews

2.1. Overview of Jingyuetan Scenic Area

Jingyuetan Scenic Spot is located in Changchun Jingyuetan Economic Development Zone in the southeast of Changchun city, Jilin Province. It is a national 5A level scenic spot, national scenic area, national forest park, national civilized scenic area demonstration base, national water scenic area, and national outdoor fitness base. The scenic area covers an area of 96.38 square kilometers, of which 5.3 square kilometers and the forest coverage rate reach more than 96%.

Jingyuetan was named after Zheng Yu, the second son of the puppet Manchukuo Prime Minister Zheng, then the head of the puppet Manchukuo Infrastructure Bureau. The forest in the scenic area is artificially built, including a complete forest ecosystem of 30 tree species, which has become a "pure land in the hustle and bustle of the city", and enjoys the reputation of "the largest artificial forest in Asia", "blue sea pearl" and "urban oxygen bar". It is the ecological and green core and city name card of Changchun.

Jingyuetan is not only an ecological leisure center, but also a sports and fitness center. As Changchun summer season and Changchun ice and snow festival of the main venue, Jingyuetan tile sa international ski festival, Jingyuetan forest marathon, Jingyuetan mountain bike marathon, Jingyuetan forest orienteering, Jingyuetan dragon boat race events were held, it is committed to advocating health, fashion and leisure lifestyle, and for the internationally renowned tourism and cultural activities to create a gathering place.

2.2. Research Status of Feature Selection Algorithm in Text Classification

Text classification generally includes several processes of text preprocessing, text model representation, feature selection, classification model training and performance evaluation. The text preprocessing process mainly divides the text data set and removes the stop words. The current main feature extraction algorithm, TFIDF, is analyzed below.

The TFIDF feature weight depends on the contribution of the feature word to the document containing the word. The algorithm because of simple mathematical calculation formula, the overall complexity of the algorithm is low, accurate feature selection become the most common text feature selection method, but in the original algorithm does not consider the difference between text categories and the text vector between class distribution information, so the classification on category bias data set effect cannot get the ideal effect.

In view of the above shortcomings, domestic and foreign scholars have proposed a variety of improved algorithms for the traditional TFIDF algorithm. AtanuDey et al. solve the problem that the traditional TFIDF only use value to evaluate the feature vector of uniary model Unigram or N-gram and makes its classification effect in emotion classification. Improve TFIDF algorithm by constructing N-gram emotion features. The improvement algorithm first extracts emotion words and their enhanced or negative words from the comment data to construct N-gram emotion features, and then combined TFIDF with N-gram emotion features to weight feature words by [4] Manny Rayner et al. proposed a dynamic programming method to optimize the TFIDF algorithm by matching syntactic strings, which improves the classification effect in the data-sparse domain of complex speech [5]. In view of TFIDF ignoring the semantic information contained in text by TFIDF, Yingying L et al. improves TFIDF by introducing the implicit semantic analysis method LSA. The improved algorithm decomposes the feature vector using singular values, and then identifies the similarity between the feature words by calculating the cosine value between the line vectors of the decomposition results, which realizes the feature selection and makes up for the deficiency of TFIDF [6]. Mohamad Irfan et al. improved the traditional TFIDF algorithm by introducing the fuzzy C mean method C-means modified feature weighted algorithm C-TFIDF. The improved algorithm used C-means method to weight the resulting sentences, divides the sentences with high weight and low weight into different groups, and clarifies the contribution degree of each statement in the document, thus improving the accuracy of feature extraction [7]. For the TF-IDF algorithm in a classification system that does not account for class distribution imbalances between features, Hao Jianlin and others proposed a modification. This modification involves constraining word frequencies within the model to improve the TF-IDF algorithm. Specifically, two assumptions were made: the sum of the feature word frequencies equals the total number of documents, and the total number of documents includes the feature word frequencies. Although these adjustments enhance the accuracy of feature extraction, the performance is not satisfactory with text datasets that have a skewed category distribution [8]. Zhao Shenghui et al. optimize the traditional TFIDF algorithm by correcting the distribution difference of feature items among document categories. By introducing category coefficient to represent the category differentiation degree of feature items, this optimization algorithm improves the weight value of feature words with strong classification ability, and reduces the weight value of rare words and noise words [9]. However, the algorithm still does not reflect the difference between text classes, so it has some limitations.

Yuan Na of Wuhan University of Technology et al. introduced the inter-class, intra-class and category distribution factors, which fully solved the imbalance between the inter-class distribution of the TF-IDF feature-weighting algorithm, and the difference between the distribution of the text vector in the classification system [10]. So this paper studies it by introducing this algorithm.

3. Data Processing

3.1. Data Acquisition

In terms of text and information selection, the current mainstream Chinese tourism social networking websites, Tuniu, Qunar, Feizhu, Ctrip, Meituan and Ctrip, divide their main business services and draw conclusions. In addition, the microblog hyperfunction with a large user group, and the search keyword "Jingyuetan", is also the main platform for tourists to express suggestions on the scenic spot.

Tuniu's product business tends to tourism customization; Qunar is mainly related to tourist tickets and hotels, exclusive, visa and group tour; Flying izhu's business is mainly inclined to the introduction and navigation of land scenic spots; Meituan and Ctrip prefer scenic spot evaluation and tickets, and Meituan and Ctrip are more targeted than other websites for studying national scenic spots. Meituan and Ctrip are well-known software in China with a wide range of services. The scenic spot information module provides tourists with comprehensive and representative scenic spot information for communication. As a platform between tourists and tourists, it is the source website for obtaining evaluation data in this study.

The data of the "Jingyuetan" evaluation column in the SNS website is crawled, and the time range is set from January 1, 2019 to June 1, 2021. The content captured included the text information of the comments, the time of publication, the number of comments, etc., and obtained 27,158 pieces of comment data.

Because the comments contain a large number of blank comments, wrong type comments, etc., the initial text data to ensure the feasibility and reliability of text classification, through climbing market review data cleaning, after finishing a total of 13684 valid data, because the two platform evaluation criteria is one star to five star, so the following star statistics, including one star score 6424, two star 558, three star 164,4 star 2836, five star score 3702. Since this experiment is for tourists to improve the scenic spot, only Samsung and Samsung comments are taken as reference. Weibo data content belongs to the one-star content category. Data statistics indicators are shown in Tab.1 below.

Table 1: Data and statistical indicators

Statistical Indicators	Mean	Crest Value	Least Value
Comment on Content Length / Word	36.579	488	1
Additional Comments per Day (from January 1, 2020)	3.805	31	0

3.2. Text Preprocessing

Data preprocessing includes manual annotation of custom dictionary, word segmentation of data, word stopping and stopping processing, etc. Because the word segmentation depends on the computer algorithm, it is inevitable that the word segmentation is not ideal situation, for some regional words, scenic spot specific terms can not be very well divided. By fully reading the review documents, we established a domain thesaurus suitable for Jingyuetan in Jingyuetan, which contains 27 proper nouns related to Jingyuetan. In addition to regional words and proper nouns, there are many numbers, symbols, and connectors like "of" and "and" that have little to do with the content of the text. In order to make the clustering effect better, this article is based on the online famous stop word database, such as "Baidu stop word list", "Harbin stop word bank" and other stop word table, and get a more comprehensive word list, and in the word segmentation found in the process of the text analysis of the words to join the stop table.

For Chinese text segmentation, this paper uses Jieba on Python. Jieba works to identify the word segmentation through the Chinese vocabulary, form a huge corpus in Chinese, determine the correlation probability between Chinese characters through the corpus, and then calculate the probability between Chinese characters, and form the words with large probability to complete the word segmentation. The preprocessing of this article is 40,000 words, and a new column of separate words is used to store each comment. Some of the data are shown in Table 2 below.

3.3. Data Feature Extraction

TF-IDF (Term Frequencyinverse Document Frequency, Word frequency-Inverse document frequency) is a text statistical method for assessing the importance of a word in a file set or corpus in a document. The basic idea is that the importance of a word is directly proportional to the number of times it appears in the current document, and is inversely proportional to how often it appears in the entire file set. The more common a word is, the lower the IDF value is. Multiplying the TF and IDF gives the TF-IDF value, and the higher the value, the higher the word is in this file, and the more likely it will be the key word of the article.

TFIDF is currently the most commonly used feature-weighting processing method nowadays, but it has certain limitations, mainly including the following two points:

1) The imbalance in the distribution between the categories of the datasets was not considered.

Most of the category distribution in real data sets is unbalanced, and different categories often have certain differences. However, the TFIDF algorithm does not take into account this difference, and the calculated feature vector weights are only based on the number of documents.

When the category distribution of the data set is quite different, especially for the weak category distribution, the calculated weight value will be too small, which will affect the classification accuracy, and cannot correctly reflect the distribution difference of the text vector between the various categories in the data set.

Table 2: Shows some data

Order Number	Grade	Comment on the Content	Release Time
1	2.0	The experience is not very good. These several pictures are ones I looked for a long time before finding the angle to take. It gives a feeling of a primitive forest, but there are a few roads artificially built. Many scenic spots are not open, including Figure 2. Rent a bike for 50 an hour, but transportation is inconvenient, so it's not recommended. There are a lot of empty places, with dead branches scattered on the ground. However, the ticket price is still cheap. I haven't gone to the tiger garden, but I might have a chance to go there later.	2021/04/15
2	1.0	The visit was made on 5.2, but the experience was super different. The entire journey of the scenic area is about 20 kilometers. As we had to accommodate the elderly and children, we chose to take a 20 yuan ride in the scenic area's shuttle bus. Although the so-called fixed point allows for random boarding and unlimited rides, the result is that there are fewer vehicles and more people, and we queued for two hours without getting on a car. No one informed me about this lack of capacity before I bought the ticket. After purchasing the ticket, I only managed to get on the shuttle bus at the middle station of the scenic spot at seven o'clock in the evening. I called the complaint line, and they replied that they understood the situation but could not solve it. With the country predicting a surge in tourists on this year's May Day, does this 5A scenic spot not even have an emergency plan? This is too irresponsible! This scenic spot does not meet the 5A standard at all!	2021/05/04
3	1.0	There are few road signs in the scenic area, which is disappointing. Two words can't describe my psychological shadow area. The most impressive part is after leaving the main gate, there's a set of stone steps a distance away. There's no sign, and we all thought going up would offer a great view. To our surprise, the bell tower costs 10 yuan. The staff, when asked for directions, were extremely impatient. They told us the comprehensive route in a way that felt like squeezing toothpaste - only revealing a little bit each time you ask. Furthermore, the attitude of the entire scenic area staff is very bad, even to the point of abusing customers.	2021/05/09
4	2.0	There is too expensive to play, eat is also expensive, a intestine sold to 5 yuan a barrel, a barrel of noodles to buy 10 yuan a barrel, also boring, no longer want to go.	2021/05/06
5	1.0	Before can drive in, now can only do the car of the scenic area, not very good.	2021/03/16
6	1.0	The hardware facilities are particularly not perfect, and the attitude of the staff is very poor. We rented a bicycle, but it broke halfway and we had to return the same way we came, and they didn't inform us of this possibility beforehand. The transportation was not arranged enough, causing us to wait at least two hours on the road.	2020/10/03
7	2.0	Not zha ground, went to three times, what also did not catch up with, the music fountain did not look at, the lotus weeping willow repair, Putuo Temple closed.	2020/10/08
……	……	……	……

2) There is no correct response to the difference in the distribution of the text vector in the classification system.

The inter-class distribution of text vector needs to consider that when the feature item ti has a large word frequency tfij value in class Cj, and the word frequency tfij value in other classes is small, the feature item should well reflect the degree of difference in text category, and should be given high weights.

Low weights are given when present in most categories and with little proportion in the categories. The intra-class distribution of text vectors needs to consider that when the feature term ti is within the Cj class, the feature term with a more consistent within-class distribution should be given high weights. When the feature item appears in only a few documents of the same kind, and rarely appears in such other documents, the feature item may be a special term, and can not reflect the category information of the text well, and should be given low weights. Therefore, this section cites a TFIDF improvement algorithm, FDCD-TFIDF [11], based on the word frequency distribution factor and the category distribution factor.

Interclass distribution factor (Inter-class distribution factor):

Reflects the distribution of feature items between document classes and classes. It can be obtained by calculating the quotient between the number of documents containing feature item ti, aij, in the Cj class and the non-Cj class containing feature item ti, with the following formula:

$$\mathrm{\alpha} = \log\left(2 + \frac{\mathrm{a_{ij}}}{\mathrm{c_{i}+1}}\right) \tag{1}$$

Inclass distribution factor (Intra-class distribution factor):

Reflects the distribution of feature items between document classes and classes. It can be obtained by calculating the quotient between the number of documents containing feature item ti, aij, in the Cj class and the non-Cj class containing feature item ti, with the following formula:

$$\mathrm{\beta} = \log\left(2 + \frac{\mathrm{a_{i,j}}}{\mathrm{b_{j}+1}}\right) \tag{2}$$

Category Distribution Factor (Category distribution factor):

Reflects the distribution information of the various categories of the document. It can be obtained by dividing the total number of documents in the data set by N divided by the quotient of the total number of documents nj contained in the category Cj, defined as follows:

$$\mathrm{\gamma} = \log\left(\frac{\mathrm{N}}{\mathrm{n_j}}\right) \tag{3}$$

So the improved weight calculation formula is as follows:

$$\mathrm{FDCD - TFIDF = tf_{i,j} \times idf_i \times \alpha \times \beta \times \gamma} \tag{4}$$

Comparative experimental design:

The comparison data came from the public Sogou Laboratory text classification corpus, from which nine categories of finance, IT, health, sports, tourism, education, recruitment, culture, and military were selected as the text data set of this experiment, and 2,000 documents were randomly selected from each category for the experiment.

Table 3: Classification results of SVM used in original TFIDF algorithm, FD-TFIDF algorithm and FDCD-TFIDF algorithm

Algorithm	Precision (%)	Recall (%)	F1-score (%)
TFIDF+SVM	86.090	85.569	85.632
FD-TFIDF+SVM	88.730	88.356	88.539
FDCD-TFIDF+SVM	89.930	89.632	89.600

Raw TFIDF algorithm, and the FD-TFIDF algorithm [12] The and FDCD-TFIDF algorithm classifies the category-balanced data sets with text using the SVM and KNN classifiers, respectively, and the classification results of the three algorithms are shown in Tables 2 and 3. The F1-score comparison results of the three algorithms after the classification of 9-category documents in the category-balanced data set are shown in Figures 1 and Figure 2.

Table 4: KNN classification results of original TFIDF algorithm, FD-TFIDF algorithm and FDCD-TFIDF algorithm

Algorithm	Precision (%)	Recall (%)	F1-score (%)
TFIDF+KNN	86.790	86.240	85.247
FD-TFIDF+KNN	89.693	89.176	88.193
FDCD-TFIDF+KNN	90.930	90.654	90.693

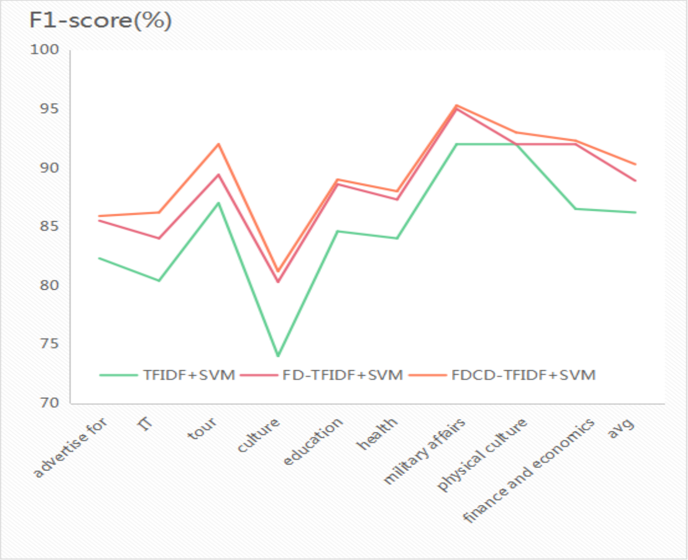

Figure 1: F1 score comparison results of three algorithms using SVM classification

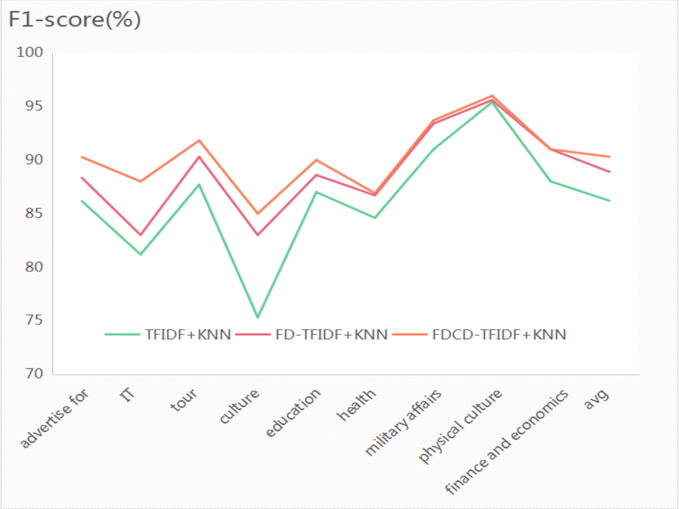

Figure 2: Classification results of KNN used by original TFIDF algorithm, TD-TFIDF algorithm and FDCD-TFIDF algorithm

According to the comparison of the Precision average value in Table 3 and Table 4, the FD-TFIDF algorithm performed the best in the SVM classification, reaching 89.584%, 3.535% higher than that of the TFIDF algorithm, and similar to that of the FD-TFIDF algorithm, and slightly higher than 0.754%. When using the KNN classification, the FDCD-TFIDF algorithm also performed the best, the Precision reached 90.829%, a 4.059% improvement in Precision over the TFIDF algorithm, similar performance to the FD-TFIDF algorithm; From the comparison of the Recall mean values between the two tables, when using the SVM classification, the FDCD-TFIDF algorithm performed the best, the Recall reached 89.218%, the Recall that is higher than the TFIDF algorithm, it performs similar to the FD-TFIDF algorithm. When the KNN classification was adopted, the FDCD-TFIDF algorithm also performed the best, and the Recall is also higher than the TFIDF algorithm and similar to the FD-TFIDF algorithm. According to the F1-score comparison results in Figures 2 and 3, the FDCD algorithm and the FD-TFIDF algorithm and the F D-TFIDF algorithm performed the best, which were much higher than the F1-score of the TFIDF algorithm. With KNN classification, FDCD-TFIDF performed the best, and secondly, FD-TFIDF performed the worst. As can be seen from the F1-score comparison results of each category, the F1-score of SVM and KNN in the FDCD-TFIDF algorithm is higher than the original TFIDF algorithm, which performs similar to the FD-TFIDF algorithm.

According to the above analysis, the FDCD-TFIDF improvement algorithm improves the feature term selection by introducing the word frequency distribution factor, and outperforms the traditional TFIDF algorithm on the data set, and the classification results are similar to the FD-TFIDF algorithm that also improves the word frequency. Meanwhile, improving the algorithm improves the accuracy of SVM and KNN on text classification. The experimental results show that the word frequency distribution factor introduced by the improved algorithm is reasonable and verify the effectiveness of the improved algorithm.

3.4. Text Clustering

There are many methods for text clustering, such as hierarchical clustering and K-means clustering. In this paper, K-means clustering algorithm is chosen, because it is easy to implement, convergence is fast, and only the number of clusters is k.

The focus of the K-means algorithm is on the determination of the k values, and the choice of the k values will greatly affect the performance of the algorithm. Therefore, the elbow method and the contour coefficient method are used to determine the optimal k-value. The core index of elbow method is SSE (Sum of the Squared Errors, sum of error) and SSE is as follows:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2 \tag{5}$$

Where Ci represents the i th cluster, and x is the sample point in Ci, and is the mean of all samples in Ci, the centroid of Ci.m_i.

The SSE is the clustering error for all samples, representing the quality of the clustering results. The elbow method is to determine a range of k-values. As the number of k-clusters increases, the samples are divided more finely, each cluster is more aggregated, and the SSE gradually decreases. When k is less than the actual number of clusters (real the number of clusters), increase k value will significantly increase the aggregation degree of each cluster, SSE will be significantly reduced, and when k reaches the actual number of clusters, increase k worth the aggregation yield will decline rapidly, so SSE decline sharply, and then with the increase of k value is flat. In other words, when the drawn "elbow map" shows an inflection point, the k-value corresponding to this inflection point is the real number of clusters. The drawn "elbow map" is shown in Figure 3:

The figure shows that the sum of squared error within the SSE —— cluster decreases sharply around the k values of 11,20, and 31. Therefore, the 5 k values of k=11, k=12, k=20, k=31, and k=32 are added into the contour coefficient method (Silhouette Coefficient) for calculation. The contour coefficient method is the Peter J. Rousseeuw's 1986 assessment of clustering results combines two factors: cohesion and separation. It can evaluate the influence of different algorithms on the clustering results according to the same raw data. The closer the value obtained by this method is, the more reasonable the clustering of the sample is. After calculation, the contour coefficient obtained when k=32 is the largest, so the cluster k value in this paper is taken as 32.

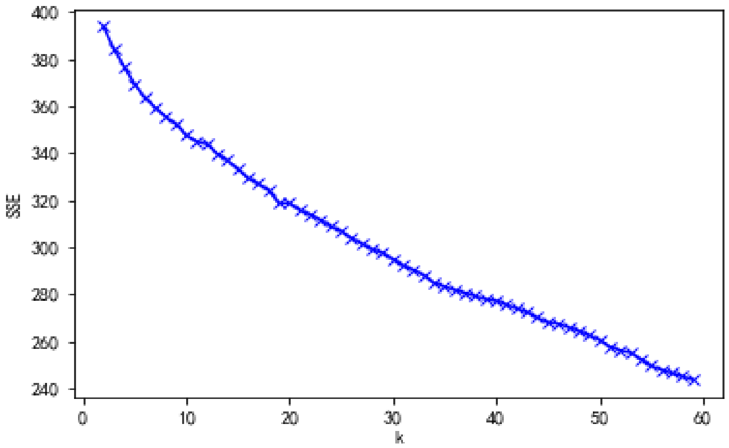

Figure 3: A line plot of within-cluster sum-of-error and cluster number relations

4. Data Analysis

4.1. Semantic Analysis of High-frequency Words

After passing through the FDCD-TFIDF word frequency statistics, Table 5 and Figure 4 are obtained. By table 5 and figure 4, for a given 27158 tourist comments, tourist comment hot topic is service, net moon lake, environment, bicycles, tickets, and signs, it can be speculated that tourists value service this factor, and the environment is clean and tidy and car rental problems will actually affect the experience of tourists, more information needs to further explore.

Table 5: Word frequency statistics

Speech	Word Frequency	Speech	Word Frequency
Serve	26123	Expensive	9283
Net Month Pool	21612	What	9076
Extraordinary	19254	Indicator	8322
Environment	17566	Route or Distance of Travel	7986
Scenery	13579	Arrange	7654
Difference	12123	Child	7388
Bicycle	11589	Wood	7254
Collect Fees	10324	Rubbish	7133
Entrance Ticket	9366	Information about and Appraisal of an Epidemic	6975

It is worth noting that in the single tourist comment words may repeat. For example, the word 'service' appeared 26123 times. You wouldn't think Jingyuetan scenic area has 26123 staff serving tourists. From the negative reviews statistics of the scenic spot SNS, there are 23206 comments including the word 'service'. That leaves nearly 3000 tourist comments. But from the perspective of the word frequency analysis, it doesn't exaggerate the meaning behind. On this basis, further analysis should be done to propose accurate suggestions.

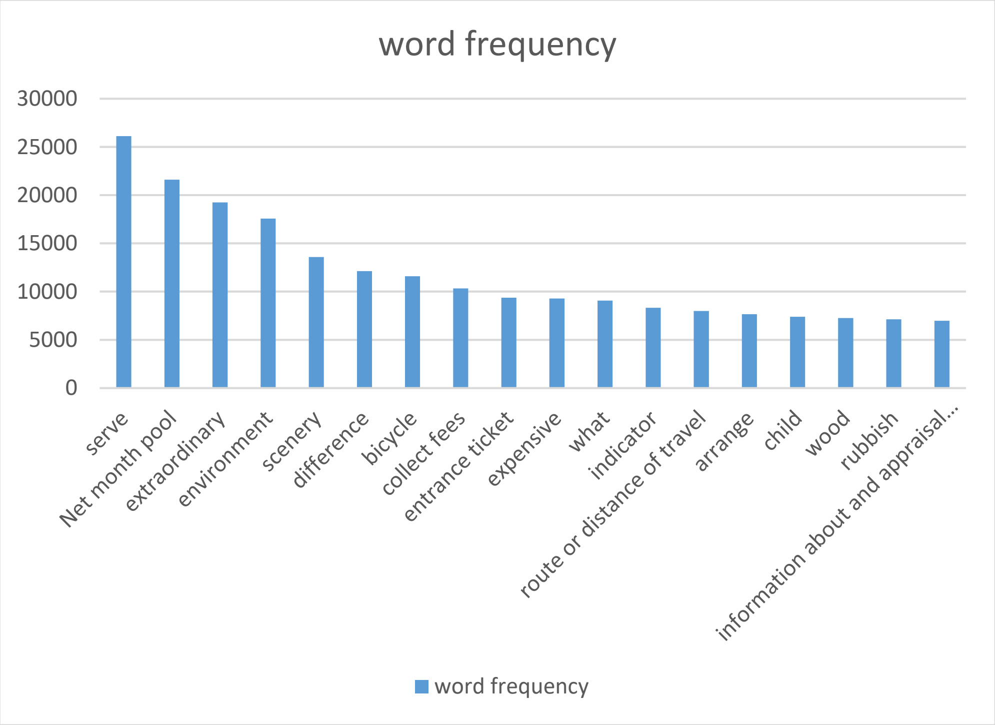

Figure 4: Word frequency statistics bar chart

4.2. Cluster Network Analysis Based on Gephi

The cluster network analysis graph is done by the Gephi tool, because Gephi communicates with other data through csv, only supports the csv format, and the csv document is composed of node data and edge data. So you process the data manually before importing it into Gephi. First, the similarity of each type of cluster results is divided into each document, and the similarity between documents is calculated with gensim (open source tool set based on Python), as the weight of each cluster result. Some document similarity matrix is shown in Tab.6 (keeping three valid digits). Using the permutation combination of 32 cluster clusters and the weights to generate the edge files, importing Gephi generated the cluster network graph, as shown in Figure 5.

As can be seen in Figure 5, this document is divided into 32 categories. After modular statistics, the system automatically divides the 32 categories into 9 categories and expressed in different colors. According to the modular statistics results, purple (C4, C8, C24, C24, C28), blue, C17, C15, C26, C30, C31), light green (C 2, C7, C10, C22, C29) accounted for the largest proportion, accounting for 55.73% of the whole document set. By outputting the top 10 keywords of the above clusters, the purple class is mainly about the content of the scenic spot environment, Including "environment", "snow tools", "taste", "ramp " and other words; The blue category is mainly about the scenic spot services, including "service", "construction", "closed the door", and other words; Light green category is mainly about the traffic content, including "bicycle", "bus", "light rail" and other words; Cyan color (C11, C14, C20) Mainly in terms of degree, including the words "very", "many" and "some"; Rose Red (C5, C19, C23) mainly about the charges, including "expensive", "rent" and other words; Orange color (C1, C6, C13) mainly about the play facilities, Including "tower", "sliding cable", "cable car" and other contents; Light Pink (C12, C21, C25, C32) mainly on the service facilities, including "signage", "slogan" and other words; Gray (C27, C16) mainly focused on epidemic prevention and control, including "epidemic", "green code", "auspicious code" and other words; Coffee-colored (C3, C9) mainly about the flow of people, including the words "queue up," "many people," and so on.

Table 6: Partial document similarity matrix table

D	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	1								
C2	0.098	1							
C3	0.037	0.003	1						
C4	0.024	0.393	0.002	1					
C5	0.069	0.003	0.022	0.003	1				
C6	0.565	0.015	0.106	0.033	0.227	1			
C7	0.003	0.503	0.002	0.257	0.074	0.002	1		
C8	0.003	0.004	0.003	0.266	0.158	0.003	0.156	1	
C9	0.050	0.002	0.768	0.019	0.183	0.109	0.002	0.002	1

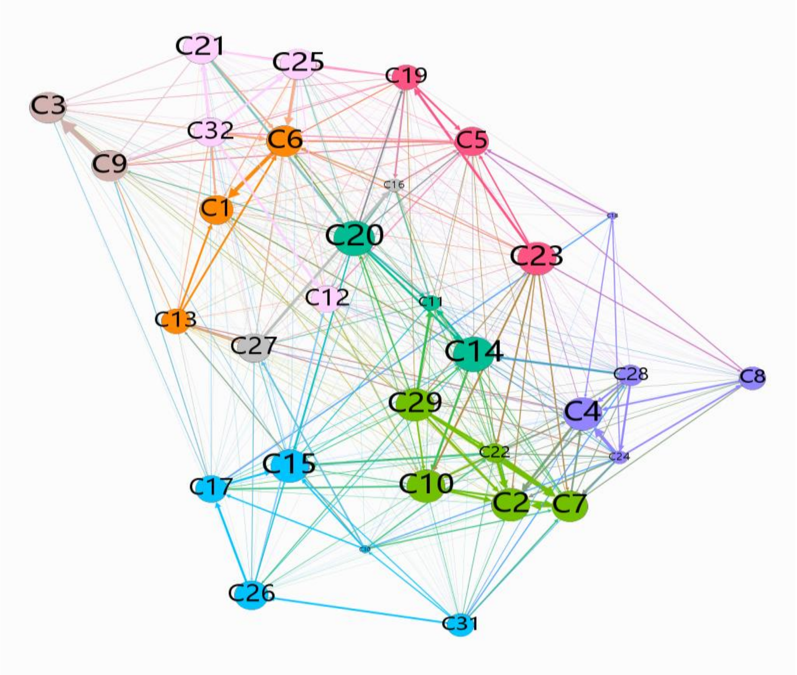

Figure 5: Cluster results network diagram

4.3. Analysis Results

According to the results of the experiment in the last section, normative suggestions are put forward for Jingyuetan scenic spot, mainly including improving the environment of the scenic spot, improving the service quality of the scenic spot, standardizing the car rental service, and regulating the charging chaos.

4.3.1. Improve the environment of the scenic spots

Strengthen the environmental improvement of scenic spots, use remote sensing, telemetry, biology and other modern means, improve the environmental monitoring and management system of scenic spots, realize the comprehensive monitoring of air and chemical pollution, vibration, heat energy, water quality, noise and radioactivity, and establish the environmental monitoring system of scenic spots. On this basis, the cloud computing platform is used to build a smart tourism service platform based on the Internet of Things technology. The platform mainly includes four parts: information acquisition subsystem, data processing and storage module, service release sub-function module and system management. Through the analysis of the return visit data of the scenic spot, monitor the surrounding environment of the scenic spot in real time, give early warning of the excessive data, realize the intelligent scheduling of the scenic spot, and disclose the monitoring results to the relevant management departments and tourists.

4.3.2. Improve the service quality of scenic spots

In addition to the general basic guarantee of hardware, software and other infrastructure, it is also necessary to establish the "smart scenic spot" professional talent training guarantee mechanism, innovate the market investment mechanism, promote the stable and efficient development of smart tourism, and realize the role of improving the service level and enhancing the comprehensive competitiveness of the scenic spots. Considering the factors of the late sustainable development of the scenic spot, the scenic spot needs to cultivate a professional and efficient construction and operation management team, excavate and cultivate talents, in order to adapt to the current development situation of "smart tourism development", and create a stable and good development environment for the construction of Jingyuetan smart scenic spot.

4.3.3. Standardize car rental services and rectify the chaotic fees

Scenic spot staff should test the car rental to play a circle of the required time, flexible time, so that tourists have enough time to play, for the scenic spot sales of goods should be standardized pricing and clearly marked price, reduce the situation of arbitrary charges, excessive charges.

5. Summary and Outlook

In recent years, with the proposal of "smart scenic spots", the state and relevant organizations pay more and more attention to the development of scenic spots. Jingyuetan should closely follow The Times and lead Jilin in the northeast in the new fields of "smart tourism" and "smart scenic spots". Combined with the results of the research experts in related scenic spots and the field research of Jingyuetan, there are indeed related problems within Jingyuetan. It can be seen that the clustering model has practical significance and good effect in text mining. The experiment in this paper is undoubtedly successful.

To sum up, in the bad comments of SNS tourists, the problems existing in the scenic spot are clarified. This article is mainly the jokes and suggestions put forward by tourists to the management center, but the beauty and construction of the scenic spot cannot be separated from the contribution made by each subject. As tourists, we should consciously protect the green mountains and clear waters of the scenic spot.

This paper takes Jingyuetan National Scenic Spot as the research object, and uses the comment data of Ctrip and Meituan network platforms for correlation analysis. However, due to the small sample data, subjective awareness is inevitably affected in the process of vocabulary extraction and word segmentation. Therefore, in the future, the accuracy and timeliness can be further improved through questionnaires and networks.

Acknowledgments

This project is sponsored by the Jilin Provincial Science and Technology Development Program and the Youth Academic Program of Guangzhou Huashang University under grant 20200403176SF, 2022HSXS081. Project Name: Research on hot topics and urban Differences of smart City Construction from the perspective of Big Data. Study on Road Administration Lighting System of Smart City.

References

[1] Chen Jianbin, Zheng Li, Zhang Lingyun. Research on IT Ability Model and Its Core Composition in Smart Scenic Spot. Tourism Science, 2014, 28 (1): 14-21.

[2] Zhang Lingyun, Li Liumin. The basic concept and Theoretical system of Smart tourism. Tourism Journal, 2012, 27 (5): 66-73.

[3] Dang Anrong, Zhang Danming, Ma Qiwei, et al. Discussion on Smart Scenic Spot Management and Service in the Era of Big Data. Western Habitat Environment Journal, 2016, 31 (4): 8-13.

[4] Anchal Gupta, Satish Mahadevan Srinivasan. Constructing a Heterogeneous Training Dataset for Emotion Classification. Procedia Computer Science, 2020, 168.

[5] Dey A, Jenamani M, Thakkar J J. Lexical TF-IDF: An n-gram feature space for cross-domain classification of sentiment reviews. International Conference on Pattern Recognition and Machine Intelligence. Springer, Cham, 2017: 380-386.

[6] Rayner M, Tsourakis N, Gerlach J. Lightweight spoken utterance classification with CFG, tf-idf and dynamic programming. International Conference on Statistical Language and Speech Processing. Springer, Cham, 2017: 143-154.

[7] Li Y, Shen B. Research on sentiment analysis of microblogging based on lsa and tf-idf.3rd IEEE International Conference on Computer and Communications, Chengdu, China. 2017: 2584-2588.

[8] Irfan M, Zulfikar W B. Implementation of Fuzzy C-Means algorithm and TF-IDF on English journal summary. 2nd International Conference on Informatics and Computing, Papua, Indonesia. 2017: 1-5.

[9] Gernein. Hotel Review Emotion Analysis based on Dictionary and Machine Learning. Jiangsu University of Science and Technology, 2019.

[10] Wu Sihui, Chen Shiping. Spam SMS identification of the Self-Attention-Based Bi-LSTM combined with TFIDF. Computer system applications, 2020, 29 (09): 171-177.

[11] Qu Dandan, Yang Tao, Hu Kongfa. Application of NLP in Automatic Collection of Symptoms Information. Software Guide, 2021, 20 (02): 44-48.

[12] Branco Ponomariov. Government-sponsored University-industry Collaboration and the Production of Nanotechnology Patents in US Universities. Journal of Technology Transfer, 2013, 38 (6): 749-767.