ISSN: 2688-9323

Perception Analysis of Tourist Comments in the Context of Modern Tourism

Xinhuang Xie, Wenzhong Zhu*, Jiawen Jiang, Qikang Wei

Sichuan University of Science & Engineering, Zigong 643000, China

Abstract

Modern tourism has broken down the information barrier and widened the channels of the communication network. However, the problem that follows is how to find the key points from the vast amount of information in the network world and use the existing information to enhance cognition. This paper selects Chengdu Research Base of Giant Panda Breeding as the research site, selects 3,961 commentary data from Qunari as the sample, uses data visualization technology and ArcGIS secondary development technology to accommodate more data types to tourism geography, carries out geographic information and image representation of data mining results, and analyses tourists' perception from online commentary data of tourists., and enhance researchers' understanding of information. The results show that the monthly travel time of tourists is clustered, which can be initially summarized as holiday factors and family outings factors. The emotional characteristics of tourists are polarized. Most tourists tend to be neutral and positive. The negative emotions mainly come from the heavy traffic of people, the guide directions of scenic spots, and so on. Tourists' play experience mainly includes eating, living, and traveling, but it does not show the characteristics strongly related to the season.

Keywords

Modern Tourism; Data Analysis; Natural Language Processing; ArcGIS.

1. Introduction

Tourism is a comprehensive industry of the national economy. According to the data released by China Tourism Research Institute, the domestic tourism revenue is 2.22 trillion RMB, which is reduced by 87.1%[1]. Reported on Domestic Tourism Renaissance 2020, it is pointed out that the revitalized travel trend has undergone new changes and a new development pattern has emerged in the tourism industry. Unlike traditional tourism, which has a single method in the planning of travel routes, industrial ecology is closed, and information exchange between passengers is slow, modern tourism breaks down information barriers and facilitates the exchange of travel experiences among passengers. Under the background of modern tourism, to have a better travel experience and avoid fatigue travel[2], passengers begin to explore actively and plan new travel routes [5-7], which meet their expectations.

In order to analyze the impact of online reviews on consumers, a lot of research has been done by domestic and foreign scholars from the impact mechanism. The typical impact mechanism research is: (1)For the comment itself, researchers quantify the online comment from the comment and score, conduct emotional polarity and intensity analysis on the comment by information mining, and build a theoretical model of the impact of online comment characteristics on consumer decision-making[8]; (2) For reviewers, researchers evaluate the authenticity of reviewers' comments in terms of their rating, nickname, number of words commented, professionalism and content profoundness[11]. Currently, the combination of (1)

ISSN: 2688-9323

and (2) is generally used to identify the group of false reviews[14], which can be used to define the scope of real reviews by researchers, the impact of consumers, and governance strategies.

Research on online review mainly uses data mining to extract key information from a large amount of data, obtain useful potential variables, and convert them into more understandable graphs. By building a linear model to quantify the number and rating of public comments, Zhang Hongyu[15] verifies the impact of online reputation on consumers' online behavior. With the help of ROST Content Mining as an analysis tool, Wang Yuwen[16] conducts high-frequency word statistics and emotional analysis on network commentary, and discusses the influencing factors of tourists' satisfaction with scenic spots. Based on statistics of word frequency, Qin Haifei[17] combines word analysis and cluster analysis, uses the index of word frequency, word frequency, word frequency weight to reduce the dimension of the data, obtains the characteristics of the data, and lays a foundation for consumer classification, manager decision-making, and intelligent recommendation according to the characteristics of the data. Li Zhaohang[18] proposed an ATF*PDF model to calculate the comprehensive weights of feature vocabulary in a text and then designed a method to extract hot place names from the high-frequency vocabulary of travelogue text. Zhang Yansen[19] proposed a dual attention model, which constructs a text semantic representation Library based on traditional content and an emotional symbol library composed of affective words, degree adverbs, negative words, etc.

Based on the modern tourism background, this paper takes the online assessment of scenic spots as the research object, uses Python web crawler technology to get the online assessment of tourists, uses double attention model and ATF*PDF model to analyze emotional data and extract tourist place names respectively, and uses data visualization technology and ArcGIS secondary development technology to accommodate more data types to tourism geography. Geographic information and image representation are carried out according to the mining results, and some systematic and large-scale inferences are made on the original style characteristics to enhance the researchers' knowledge of information.

2. Data Sources

Select the mainstream tourism portal sites with more users, and combine Alexa's world ranking with Chinaz's ranking to make statistics on the outline information of these sites, including the number of users of the site, point-and-answer, and POI of tourist attractions. The specific data are shown in Table 1, and the tourism portal sites are filtered.

Table 1. Data comparison of domestic tourism websites

Portal Site	POI	Subscribers	Comments	strategy	Alexa Ranking	Chinaz Flow(IP)
Qunar	Complete	600 million	15444	1060	2370	9,423~15,051
Ctrip	Complete	400 million	58419	0	3851	8,707~13,907
Mafengwo	Complete	120 million	5922	1433	6840	344,805~550,795
Tongcheng	Incomplete	152 million	8021	0	59269	36,297 ~ 57,981
Tuniu	Complete	Unknown	15	0	53374	44,953 ~ 71,807
Lvmama	Incomplete	Unknown	0	0	11665	22,996 ~ 36,734

According to Table 1, this experiment selects Qunari as the experimental data source, selects Chengdu Research Base of Giant Panda Breeding as the experimental research object, and uses Python crawling technology to get the related evaluation data of Chengdu Research Base of Giant Panda Breeding from Qunar, including user id, user nickname, time of evaluation, POI of the site, evaluation of the attractions and evaluation content. Data preprocessing includes deletion of empty shared content, lack of textual descriptions for pictures only, irrelevant

ISSN: 2688-9323

content for scenic spots, data with duplicate user IDs, and mutual replies from users, and sorts out 3961 pieces of effective comment data. It is found that the time span of comment data is from 2014 to 2021, and there is a significant difference in the number of comments. The pre-processed data is stored in the MYSQL database to support subsequent data mining and visualization. Figure 1 shows the complete process of data collection and preprocessing.

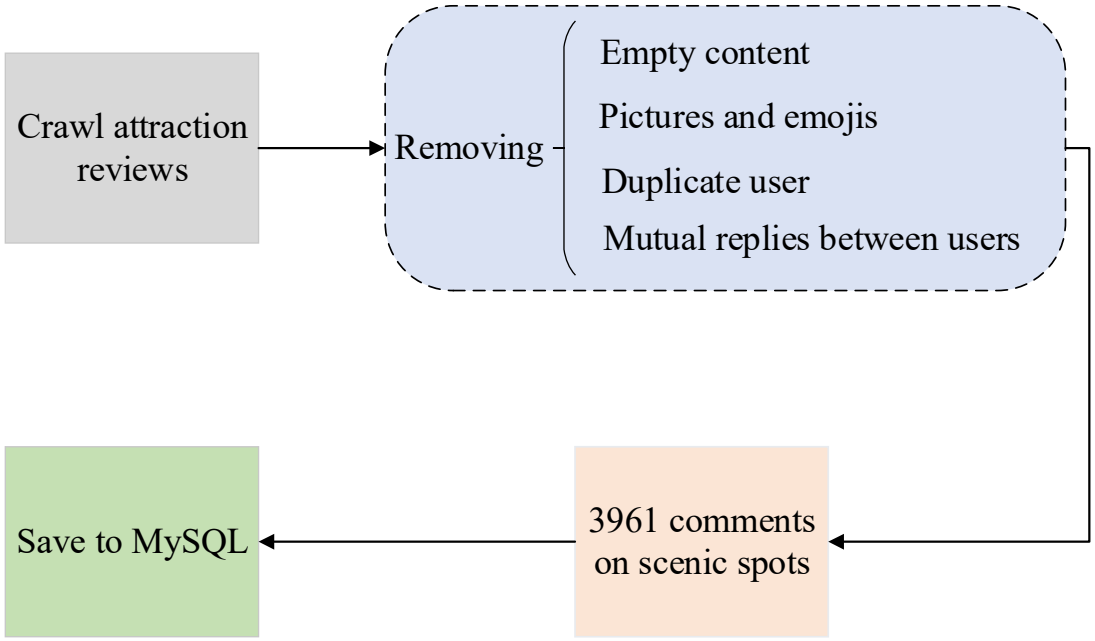

Figure 1. The process of data collection and preprocessing

3. Data Analysis

3.1. Time Attribute Analysis

Firstly, the time attribute of the assessment data is analyzed. The number of the assessment is counted by the time span of the year and the month, and the distribution of the assessment data by the monthly and annual statistics is shown in Fig. 1 and 2, respectively.

By dividing the time span by month, we can get a clearer understanding of the weak and peak seasons of this scenic spot. From Figure 1, the online assessment data of this scenic spot are mainly concentrated in May to August, and the weak season is mainly concentrated in December to March. The distribution of the overall weak and peak season of the scenic spot conforms to the traditional festivals and reality of our country.

The main reasons for the peak season of this scenic spot from May to August include: 1. Travel on traditional festivals; Data volume in May and June is much higher than that in April, and it is the most concentrated in May. There are two factors: on the one hand, the holidays of Labor Day and Youth Day in May are longer than that of Dragon Boat Festival in June; On the other hand, in the traditional festival sense of our country, Dragon Boat Festival means exorcism, avoids evil and seeks peace. 2. Family outings; in light of China's social conditions, there are no statutory holidays in July and August. From the perspective of parents, during the summer holidays, parents have a greater willingness to travel with their children and increase their knowledge. From the perspective of students, only a small number of students have a holiday in July. This part of students often contains two groups: younger children, primary school students, and upcoming students to step into society. In July, students with greater autonomy and the ability to travel with their parents often face academic pressure. To sum up, the monthly comment distribution map can effectively reflect people's actual travel conditions.

ISSN: 2688-9323

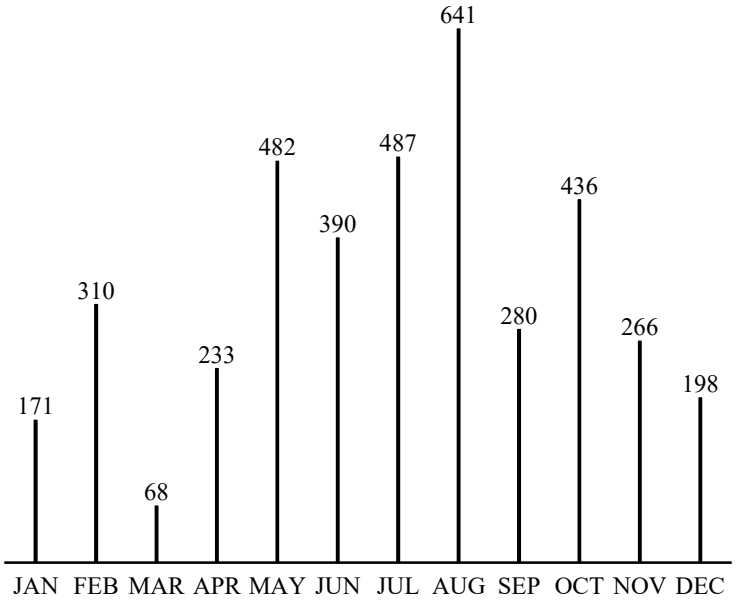

Figure 2. Distribution of monthly comments

By dividing the time span between 2014 and 2017, researchers can more clearly understand the change of the heat trend of the scenic spot. From Figure 2, there are fewer data from 2014 to 2017. The possible reasons for the analysis are as follows: 1. Google retrieved the keywords of “Chengdu Research Base of Giant Panda Breeding”, and selected the time span from 2014 to 2017 to retrieve several panda deaths during this period. Combined with the slogans such as "care for wild animals" publicized by China's animal care organization at that time, the scenic spot may suffer some resistance from the public at that time. 2. The data samples obtained in this paper are only 3961, including fewer data in 2014-2017, and do not cover all the comment data in this period. However, the analysis of the monthly evaluation in Figure 1 shows that the sample data of this experiment is reliable. In Figure 2, the comments peaked in 2018, decreased slightly in 2019, and fell precipitously in 2020. The change in the heat trend in this period is extremely obvious. It is obviously an external factor of the COVID-19, which has led to the weakening of people's willingness to travel, and it has not recovered in 2021. Due to the impact of the COVID-19 as a macro factor, reasonable analysis shows that the heat trends of other scenic spots during this period are also similar.

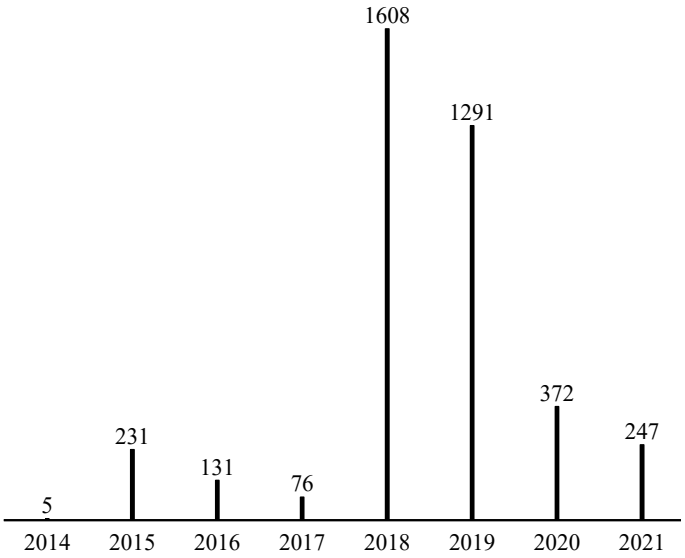

Figure 3. Distribution of annual comments

ISSN: 2688-9323

3.2. Text Sentiment Analysis

The word segmentation system of this experiment uses the Jieba library function in Python to segment text information, remove meaningless stop words, and convert all English into lowercase in the process of word segmentation. After synthesizing the stop word database of Harbin Institute of Technology, the stop word database of the Machine Learning Intelligence Laboratory of Sichuan University and the "Baidu stop word list" and de-duplication, it was selected as the stop word database of this article. The experimental model uses the double attention model proposed by Zhang Yang-sen et al to analyze the emotion of the text. In the current traditional model of affective analysis, affective scores are mainly calculated by traversing the text affective words, degree adverbs, and negative words. The double attention model builds a text semantic representations library mainly based on traditional content, and an emotional symbols library composed of emotional words, degree adverbs, negative words, etc. It not only codes the text as a whole but also pays more attention to the function of emotional symbols in text emotional expression. The model is shown in Formula (1).

$$v = \sum_{t=1}^{T} a_t h_t \tag{1}$$

$$a_t = \frac{\exp\left(e_t^T A\right)}{\sum_{k=1}^{T} \exp\left(e_k^T A\right)} \tag{2}$$

$$\mathrm{e}_t = \mathrm{Tanh}(Wh_t + b) \tag{3}$$

Formula (1), v is a text vector related to the context of the text, weighted by the input state h_t. The weight of h_t is a_t; The values of a_t and e_t are calculated by formulas (2) and (3), respectively. The weight value of a_t is determined by the semantic representation A, the model weight W, and the offset b. From this model, a text vector containing weights for each input state is obtained. With the help of text vectors, the total score is calculated statistically, with a score less than 0, indicating that the text's emotional tendency is negative and a score greater than 0, indicating that the emotional tendency is positive, and the larger the score, the more positive the text's emotional expression.

To sum up, the text affective analysis process is obtained as shown in Figure 3, and the affective analysis score is shown in Figure 4.

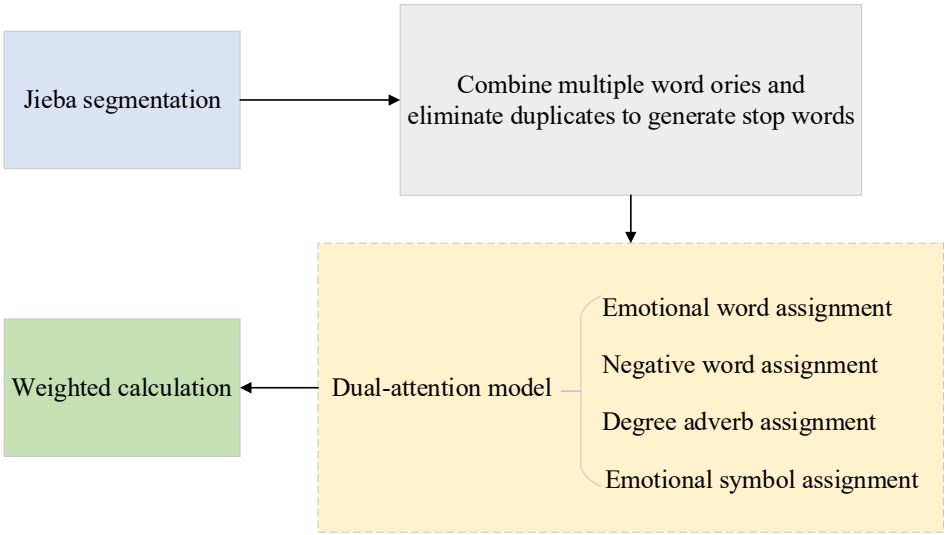

Figure 4. Text emotion analysis process

ISSN: 2688-9323

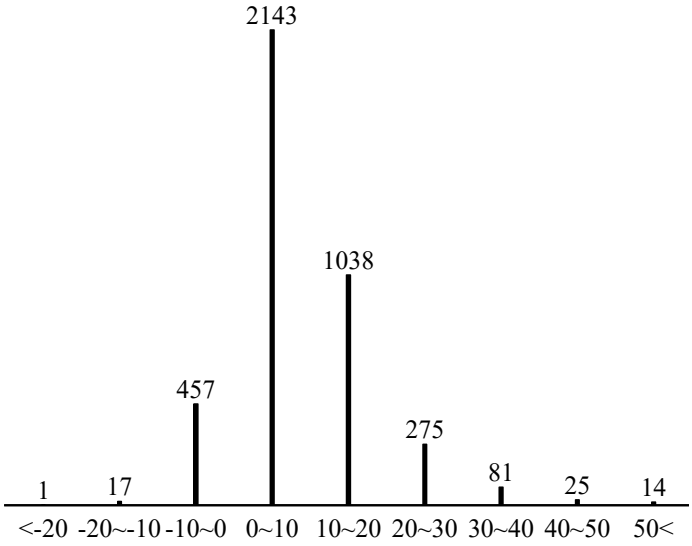

Figure 5. Text emotion score distribution

As shown in Figure 4, the emotional scores of the comment text are mainly concentrated in the two ranges of 0-10 and 10-20. The lowest emotional score is below-20, which has a sample size of 1, while there are 14 samples with a score higher than 50. A large number of data sets are in the middle area, with fewer extreme values at both ends, and the distribution of data from the middle to both sides decreases gradually. The numerical distribution is bell-shaped and satisfies the normal distribution as a whole.

The emotional scores obtained from the above steps can only be used to get the emotional state of the tourists, from which researchers still cannot get the focus of the tourists' emotions. In order to filter a large amount of text information and extract keywords with high frequency, this experiment uses Matplotlib, WordCloud, Numpy, Image library functions in Python to make a word cloud. Here, the Image library function takes the background picture. On the basis of Chinese Jieba word segmentation, the word cloud is generated from the Post-Word commentary data using the WordCloud library function. The Numpy library function is used to perform a matrix operation on the background of the word cloud image and assign the word cloud to the matrix. Finally, Matplotlib is used to draw the word cloud, as shown in Figure 6.

Figure 6. Word Cloud

ISSN: 2688-9323

Combining with Fig. 5, it is observed that there are very few negative words such as "No" and "None", which indicates that some passengers are unhappy in the text commentary. Then, the combination of words with negative emotional tendencies, such as "many" + "people", corresponding to the text emotional score distribution map, constitutes part of the interval population with lower emotional scores than the average. The word cloud consists mainly of positive words such as panda, cute, giant panda, national treasure, good, not bad, and so on. The overall positive attitude is confirmed by Figure 5.

3.3. Peripheral Tourist Route Analysis

By identifying the place names and related vocabulary in online review, it can reflect the degree of tourists' attention to the tourist landscape, and help travelers to have a clear understanding of the route and attention of the scenic area. Due to the irregular and non-standardized naming of places in online text, this section uses ATF*PDF model, combines the two factors of text word frequency and text length, and excludes the results of non-place name vocabulary to get the results of the place names around the site. The ATF*PDF model weights words in the text to reflect their importance in terms of their combined weight in text datasets. The model formula (4) shows.

$$w_i = \frac{\sum_{j=1}^{N} |tf_{ji}|}{N} e^{\frac{n_i}{N}} \tag{4}$$

$$|tf_{ji}| = \frac{tf_{ji}}{\sqrt{\sum_{i=1}^{m_j} tf_{ji}^2}} \tag{5}$$

Formula interpretation: w_i is the weight of word i, and $|tf_{ji}|$ is the normalized frequency of word i in the text. The model formula mainly consists of two parts. In formula (4), $\frac{\sum_{j=1}^{N} |tf_{ji}|}{N}$ is ATF, which is the average frequency of words in the sample text; $e^{\frac{n_i}{N}}$ is PDF, which means to increase the weight of words that appear repeatedly in multiple texts.

Table 2. Word Frequency Statistics

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
Giant panda	283	Loudian	57	Huaxi	29	Restaurant	17
Breed	277	Estern Station	54	Mammy	29	Big hotel	16
Research	275	Jingsha	52	Lan Kwai Fong	28	Paulownia tree	16
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
Hotel	80	Museum	31	South Gate	20	Dujiangyan	13
Estern suburb	78	Bookstore	30	Creativity	18	Dalong	13

Exclude the place name vocabulary and use the Ucinet software as an analysis tool. By comparing the existing quantitative relationships of the vocabulary network with those that may exist in theory, the density of the relationships among network members, that is, the

ISSN: 2688-9323

network density is obtained. In order to judge the correlation between the data, this experiment converts the multi value matrix of the original data into a binary matrix. In the calculation process, each row label is used as a critical point to divide the cell, and the higher than the average value is considered to have a significant relationship, and its weight is set to 1, and vice versa, to 0. Part of the binary matrices are shown in Table 3, and the network density diagram is shown in Figure 7. The degree of local centrality in Figure 7 reflects the degree of concentration of nodes or relationships in the core region. Figure 8 shows the relationship between the central aggregation points.

Table 3. Partial binary matrix

Row Labels	IFS	U37	Museum	Thatched Cottage	Chunxi Road	Daci Temple	Big hotel
IFS	0	0	0	1	1	0	0
U37	0	0	0	1	1	0	0
Museum	0	0	0	1	1	0	0
Thatched Cottage	0	0	0	0	1	0	0
Chunxi Road	0	0	0	1	0	0	0
Daci Temple	0	0	0	1	1	0	0
Big hotel	0	0	0	1	1	0	0
Estern Suburb	0	0	0	1	1	0	0
Estern Railway Station	0	0	0	1	1	0	0
Fangsuo Bookstore	0	0	0	1	1	0	0
Park	0	0	0	1	1	0	0
Heming	0	0	0	1	1	0	0

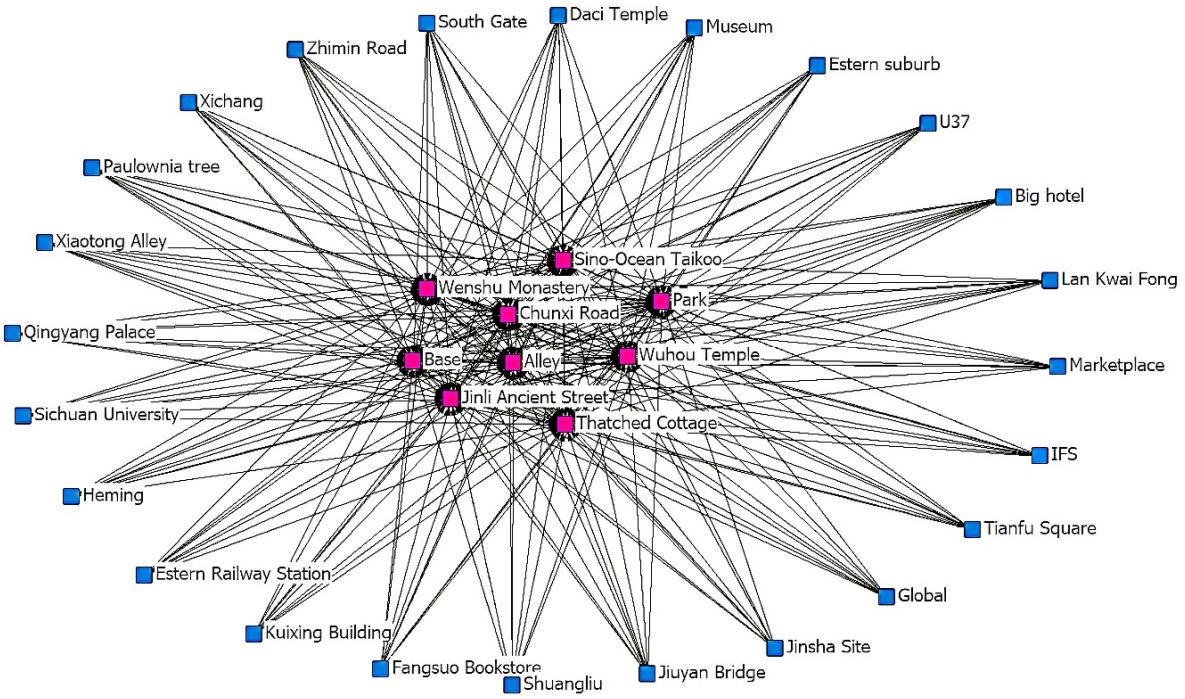

Figure 7. Network density Center

ISSN: 2688-9323

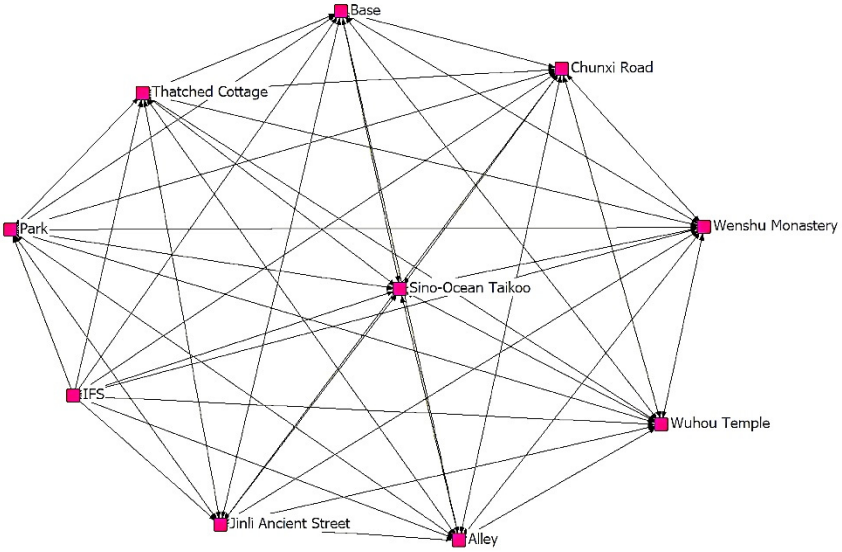

Figure 8. Aggregation point Connection

At the same time, the above experiments provide site samples for the subsequent secondary development of ArcGIS by identifying related place names of tourist destinations. Correctly recording the evolution of tourism destinations is of great significance for the management system and policy-making of each tourism destination. In this section, the geographical information processing shall be carried out for the surrounding place names obtained after processing. The specific implementation includes applying for Baidu map to develop AK key, writing Python program and calling Baidu map API interface, receiving the requested returned data, referring to the returned parameter document, and parsing the data, including longitude and latitude, province, city, county, district, township, etc. Next, the administrative area data of provinces and urban areas that obtain national geographic information from the National Bureau of statistics are imported into ArcGIS software for processing, the elements of Chengdu, Sichuan Province are selected, and a new layer is generated. In the new layer, import the heat place-name coordinate information obtained before, label the features, and get a visual heat place-name map, as shown in Figure 9.

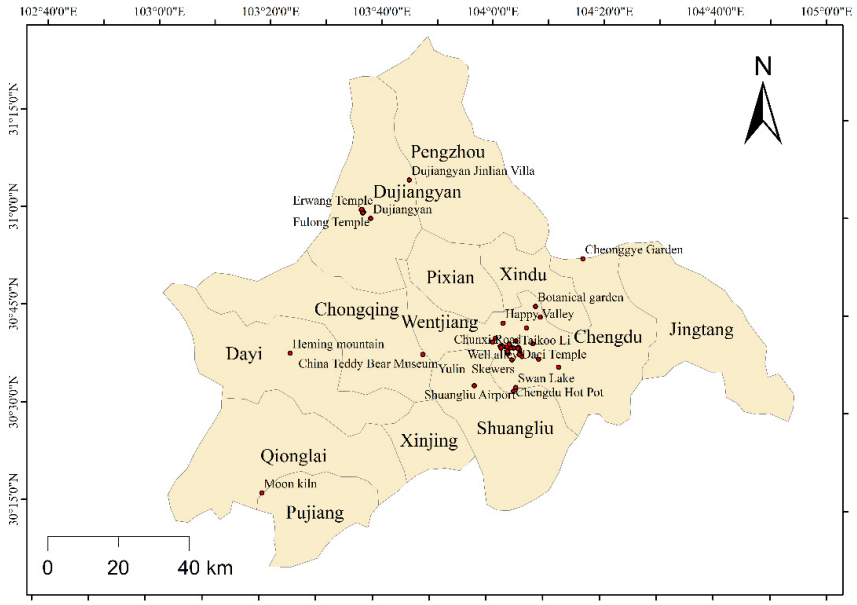

Figure 9. Routes of surrounding tourist hotspots

ISSN: 2688-9323

4. Conclusion

This paper takes the tourists of Chengdu Research Base of Giant Panda Breeding as the object of study. Based on the online review data of qunar scenic spots, Python and ArcGIS are used to analyze the tourists' experience of Chengdu Research Base of Giant Panda Breeding. The main research contents are listed as follows:

(1) There is a clear gathering phenomenon in the month time of the tourists' trip, which can be summarized as holiday factors and family travel factors by subdividing the gathering interval of traffic peak. The year-round flow of scenic spots is concentrated in July and August, which includes both winter and summer holidays, as well as parents' factors to cope with students' academic pressure and travel with their children before school starts.

(2) Tourists' satisfaction with this attraction is relatively good. Most tourists tend to be neutral and positive. Some tourists show very positive emotions about the visit to the scenic spot, while the negative emotions of tourists mainly come from the large flow of people, poor scenic spot guide instructions, and so on.

(3) Tourists' play experience mainly includes three aspects: eating, lodging, and traveling, and it does not show the characteristics of a strong correlation with the season. The "eating" and "touring" of scenic spots are the focus of tourists' play experience, and they are also the driving factors for tourists to recommend surrounding scenic spots.

(4) The tourist routes mainly focus on the core scenic spots such as Breeding base, Kuanzhai Alley, Wuhou Temple, Chunxi Road, and so on. with the core area as the center and extending outward to other scenic spots.

Based on the conclusions of this experiment, in order to improve tourists' play experience, scenic spot management can be improved from the following aspects:

(1) Set up a query button on the ticket purchase page of the scenic spot to query the number of tickets sold in a certain section of the scenic spot, reduce the traffic burden of scenic spot management, and help tourists to choose the time period for travel, and give them to a certain extent foot traffic expectations

(2) Take the initiative to understand the emotional state of tourists, and promptly channel the extreme emotions of tourists. In order to avoid the tourist experience of wandering horses and horses, and to improve the comfort of tourists on the way, the management of attractions should resolutely prohibit the behavior of tour guides forcing tourists to buy souvenirs and joint shops forcing tourists to trade.

Acknowledgments

Technology R & D project of Sichuan smart tourism research base (ZHZJ19-01); Graduate Innovation Fund of Sichuan University of Science & Engineering (y2021090, y2021092).

References

[1] National Bureau of statistics. Statistical bulletin of the people's Republic of China on national economic and social development in 2020 [J]. China Statistics.

[2] Pang Bo. Analysis on the impact of tourist fatigue on the quality of tourism experience [J]. Business Culture, 2021 (08): 142-144.

[3] Mou Lin. Research on the development status, characteristics and existing problems of tourism accommodation standardization in China [J]. Standard Science, 2021 (07): 67-74.

[4] He Biao, Zhang Wen, Zhu Lianxin, Tong Yun. Construction of evaluation index system for governance ability of tourism industry [J / OL]. Journal of Hainan University, 1-9 [2021-08-03].

ISSN: 2688-9323

[5] Zhu Junrong, Huang Ailian. Discussion on the path of high-quality development of Tourism under the impact of epidemic [J]. Journal of Liaoning University of Technology (Social Science), 2021,23 (04): 16-20 + 24.

[6] Yang Yongheng. Road map leading high-quality development of culture and tourism [n]. China Culture Daily, 2021-06-04 (002).

[7] Guo Lanbo, Wang Zhen, Wang Yan, Yan Shuang. Research on tourism route optimization design method based on multi-objective constraints [J]. Surveying and spatial geographic information, 2021, 44 (07): 165-167.

[8] Lu Xianghua, Feng Yue. The value of online word of mouth -- An Empirical Study Based on online restaurant reviews [J]. Management World, 2009 (07): 126-132 + 171.

[9] Wang Hongwei, Song Yuan, Du Zhanqi, Zheng Lijuan, Hua Jin, Zhang Yiwei. Evaluation of express service quality based on Emotional Analysis of online comments [J]. Journal of Beijing University of Technology, 2017,43 (03): 402-412.

[10] Qu Lianzhuang, Guan Yi, Yao Xiaolin. Exploration on the influence mechanism of product comment characteristics on consumer information adoption in e-commerce environment -- from the perspective of information users and system users [J]. Journal of University of Electronic Science and Technology (Social Science), 2021,23 (04): 31-37.

[11] Forman C, Ghose A, Wiesenfeld B. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets[J]. Information Systems Research, 2008, 19 (3): 291-313.

[12] Ngo-Ye T L, Sinha A P. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model[J]. Decision Support Systems,2014,61:47-58.

[13] BAEK H, AHN J, CHOI Y. Helpfulness of online consumer reviews: readers' objectives and review cues[J]. International Journal of Electronic Commerce, 2012, 17 (2) :99-126.

[14] Yuan Lu, Zhu Zhengzhou, Ren Tingyu. Review of false comment recognition [J]. Computer science, 2021,48 (01): 111-118.

[15] Zhang Hongyu, Zhou Tingrui, Yan Huan, Tang Xiaofei. Research on the impact of online word of mouth on consumers' online behavior [J]. Management world, 2014 (03): 178-179.

[16] Wang Yuwen, Luo Peicong, Liu Yingnan, Zheng Wenjuan. Research on tourist satisfaction of Meizhou Island based on online comments [J]. Journal of Fujian Normal University (Natural Science), 2018,34 (05): 83-92.

[17] Qin Haifei, Du Junping. Feature mining of hotel online review data [J]. Journal of intelligent systems, 2018,13 (06): 1006-1014.

[18] Li Zhaohang, Guo Fenghua, Li Renjie, Fu Xueqing, Yan Zhengfeng. Extraction methods and empirical research of hot place names in a large number of online travel notes [J]. Geography and geographic information science, 2015,31 (01): 68-73.

[19] Zhang Yangsen, Zheng Jia, Huang Gaijuan, Jiang Yuru. Microblog emotion analysis method based on dual attention model [J]. Journal of Tsinghua University (Natural Science), 2018,58 (02): 122-130.