Analysis and Research of Tourist Appeal Based on Online Review Visualization

Jiaxin Hu[1,a], Mengjia Lei[2,b], Jianfeng Dong[2,*]

[1]College of Computer Science and Engineering, Jishou University, Zhangjiajie, China

[2]College of Tourism, Jishou University, Zhangjiajie, China

[a]953072301@qq.com , [b]2021400764@stu.jsu.edu.cn

*Corresponding author: yuanrele@163.com

Keywords: Online comments; Tourist appeal; LDA topic analysis; Sentiment analysis

Abstract: The demand of tourists is the direct reflection of tourists to the service of tourist scenic spots, and it is also the basis of improving the service of scenic spots. This paper takes Zhangjiajie National Forest Park as the empirical object, and identifies the keywords and emotional tendencies in the comments by deeply mining the online comments of Ctrip users. On this basis, through co-semantic analysis, the review data is transformed into a visual relationship network to reveal the internal logic and interrelation of tourists' demands. Through data visualization analysis, the main characteristics of tourists' demands and their causality are extracted. The research results show that the demands of tourists are mainly reflected in the layout of scenic spots, infrastructure and services, tourism interaction and other aspects, which can be improved by optimizing the layout of scenic spots, strengthening the construction of infrastructure and creating interactive scenes.

1. Introduction

According to the 52nd Statistical Report on the Development of the Internet in China [1], as of June 2023, the number of Internet users in China has reached 1.079 billion. Among them, the scale of online travel users (travel booking) reached 454 million people, accounting for 42.1% of the total number of Internet users, "Internet + tourism" has become an important way of tourism consumption. OTA platform as an important carrier of online travel, its user generated content (UGC) has become a key information source for travel decision-making and management. Various tourist reviews in the OTA platform not only reflect the direct experience of tourists, but also contain the expectation and evaluation of service quality. Especially with the rapid development of domestic tourism, the online travel service platform represented by Ctrip collects a large number of tourist comment data, providing rich data for tourism services and management decisions. Zhangjiajie National Forest Park, as an internationally renowned tourist destination, attracts a large number of domestic and foreign tourists every year. However, the major OTA platforms do not categorize the comments of tourists[2]. On the one hand, tourists cannot obtain effective comment information in a short time. On the other hand, with the increase of user comments, how to optimize the management of scenic spots according to comments and improve tourists' satisfaction has become an urgent problem for tourist destinations.

Therefore, based on the perspective of tourists' comments, this paper intends to establish a visual model of tourists' demands through in-depth analysis of users' online comments, and then build a comprehensive tourist satisfaction evaluation system. It can not only provide decision-making suggestions for tourism destinations to improve the quality of tourism services, but also help meet the growing tourism demand and promote the healthy development of tourism industry.

2. Theories and methodologies

2.1 Theories

2.1.1 Tourism satisfaction theory

The theory of tourism satisfaction refers to the overall evaluation degree of tourists on tourism

  DOI: 10.25236/etmhs.2024.009

experience, including tourists' feelings and evaluation on scenic spot environment, service quality, price rationality and other aspects[3].Research shows that satisfaction is an important factor affecting tourists' willingness to travel again, and it is also the main driving force of tourist word-of-mouth communication. This paper will use the theory of tourism satisfaction to make a visual analysis of tourists' online reviews to understand their satisfaction level and its influencing factors.

2.1.2 Online review analysis theory

Online comment analysis is a process of collecting, sorting, analyzing and mining online comments published by users on the network platform to reveal the information and characteristics contained therein[4]. In the tourism industry, online comments of tourists have become an important source of information, which can reflect the actual experience and needs of tourists, and have important guiding significance for the management and service improvement of scenic spots. According to the Zhangjiajie Statistical Yearbook, Zhangjiajie received 4.3 million tourists in 2023. Among them, OTA platform is the main position for tourists to choose tourism services, but also the gathering area of tourists online comments. Through the analysis of Ctrip, a major OTA platform, there are 22,503 comments about Zhangjiajie National Forest Park and 4,918 relevant answers, which provide a rich source of data for this study.

The above theoretical basis is conducive to the theoretical research and analysis of tourists' online comments, and provides theoretical support and guidance for the subsequent empirical analysis.

2.2 Methodologies

2.2.1 Text word frequency analysis

Word frequency analysis is a text analysis method based on computer text processing technology. Word frequency statistics are carried out on cleaned data through Python program to mine the main content and key features of text[5]. The calculation method of word frequency is divided into absolute word frequency (that is, the number of occurrences of words) and relative word frequency (that is, the ratio of the occurrence of words here to the total number of words in the text). In this paper, the absolute word frequency is calculated and the words are sorted by word frequency, and the word cloud map is used to show the word frequency distribution.

2.2.2 Topic analysis based on LDA

Latent Dirichlet distribution (LDA), as an unsupervised learning algorithm, can effectively mine the latent topic of text data and provide an effective means for text clustering[6]. By introducing potential topic variables to describe the relationship between documents and words, the LDA topic model is constructed on the topic vector of extracted text, and the model is trained to obtain document-topic distribution and subject-word distribution, so that the topics of statistical text are clustered in the form of probability distribution. [7] The document-topic distribution refers to the probability distribution of different topics in each document. Suppose the document is represented by d and θd is represented by the probability distribution of the document topics, i.e. the probability of κ in the document $\theta d, \kappa$ for each topic is[8]. The formula is:

$$\theta d \sim \mathrm{Dir}(\alpha)$$

When θd is obtained (that is, the probability of each topic in each document estimated during model training), the probability of each topic in a particular document is found as:

$$P(z = \kappa | d) = \theta d, \kappa$$

In the word-topic distribution, for any topic κ, its distribution is $\emptyset\kappa$, while for each topic κ and each word ω, its probability distribution can be expressed as$\emptyset\kappa, \omega$, expressed as:

$$\emptyset\kappa \sim \mathrm{Dir}(\beta)$$

For each topic, the probability is normalized, and for a particular topic k, the sum of the probabilities of all words is 1, expressed as:

$$\sum_w \emptyset\kappa, \omega = 1$$

Given the parameters α and β, the joint distribution probability of a set of documents and their topics can be expressed as:

$$P(\theta, z, \omega|\alpha, \beta) = d = 1\prod DP(\theta d|\alpha)(n = 1\prod N_d P(Zd, n|\theta d)P(Wd, n|Zd, n, \beta))$$

On the basis of existing literatures, LDA thematic model is used as the research method, online reviews of Zhangjiajie National Forest Park on Ctrip are used as the data source, and text mining technology is combined to analyze online reviews of Zhangjiajie tourists and analyze consumer portraits of different tourists, so as to deeply understand tourists' demands. This could provide references for precise marketing of Zhangjiajie as a tourist destination and promote high-quality development of tourism in Zhangjiajie.

3. Empirical study

All the tourist comment text about Zhangjiajie National Forest Park on Ctrip platform was selected as the data source, and the data was crawled by Python crawler, including the comment content, user ID and comment time. It received 14,762 comments, with a total of about 820,000 words. In the data preprocessing stage, the data was first de-duplicated, and 12,672 data were obtained after the 14,762 data were de-duplicated. For pre-processed comments, there is still a risk of excessive error caused by direct use, so the Python Jieba package is called, and the data is then stopped for word processing. Here, with the help of HIT's stop word table, some modal words and adverbs are also added to the code as stop words.

3.1 Text word frequency analysis

3.1.1 Word frequency statistics

In order to fully reflect the overall impression of tourists on Zhangjiajie National Forest Park, this section selects the high-frequency keywords in the comments as the analysis object, which can not only avoid the fuzzy results caused by too few words, but also ensure the integrity of the core keywords and avoid the burden of excessive words. Table 1 shows the statistical analysis of some high-frequency words. y1 is the order of frequency words, y2 is the word name, y3 is the word frequency, and y4 is the part of speech. Part of the word frequency statistical analysis is shown in Table 1:

Table 1 Word frequency statistical structure

Serial number	name	Part of speech	frequency Serial number	Serial number	name	Part of speech	frequency Serial number
1	Zhangjiajie	n	2751	11	Elevator	n	1469
2	Scenic area	n	2457	12	Tianzi Mountain	n	1467
3	Scenery	n	2421	13	Yuan Jiajie	n	1313
4	Nice	a	1929	14	Bailong	n	1120
5	Very	d	1800	15	Time	n	1082
6	Beauty	n	1684	16	Without	V	980
7	Attractions	n	1567	17	Senses	n	976
8	Cableway	n	1549	18	Queue	n	957
9	Ladder	n	1524	19	Wulingyuan	n	937
10	Worth	v	1516	20	Convenient	a	912

3.1.2 LDA topic cluster analysis

Latent Dirichlet Allocation (LDA) is a document topic generation model based on Bayesian

distribution. [9] It mainly consists of three layers: word, topic and document. Figure 1 is the labeling diagram of the LDA topic model, where double circles represent measurable variables, single circles represent potential variables, arrows represent dependencies between variables, and rectangular boxes represent repeated sampling.

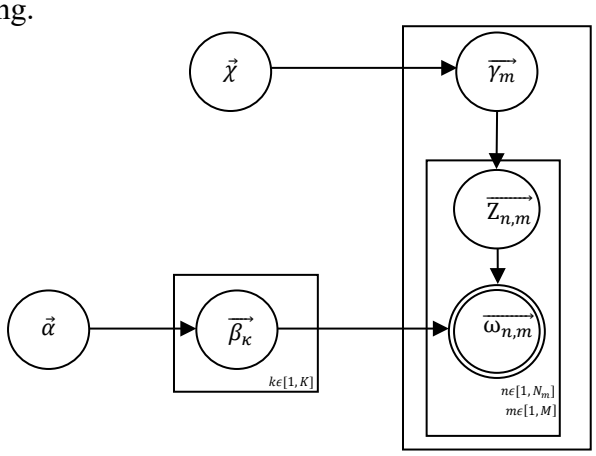

Figure 1. Markup diagram of the LDA topic model

Based on the comments and distribution of different words, the text is divided into five themes: National Forest Park, cableway, elevator, landscape and Zhangjiajie. The distribution of each theme in the text is shown in Figure 2 as follows:

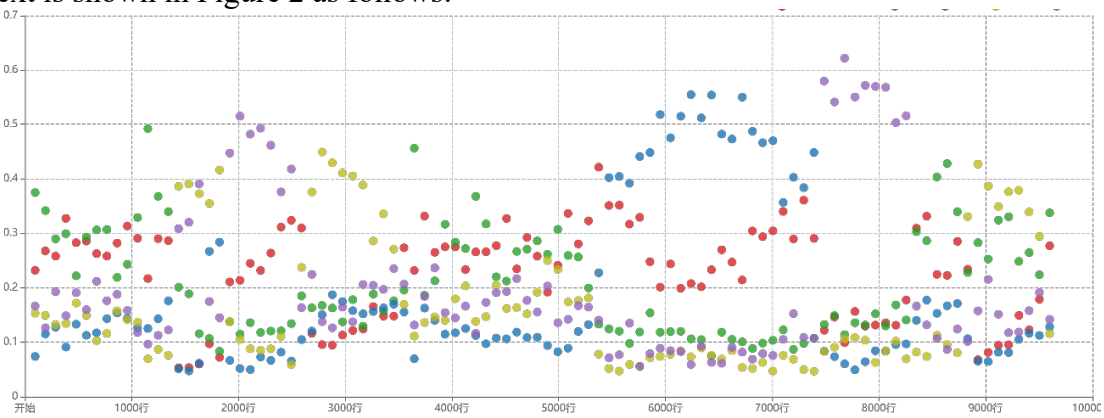

Figure 2. Scatter plot of topic distribution

4. Conclusions and Suggestions

4.1 Research conclusions

By analyzing the online comments of Zhangjiajie National Forest Park, this paper visually analyzes the comments text and effectively identifies and analyzes the demands of tourists. The study found that tourists have certain demands for the information openness of scenic spots. The transparency and openness of scenic spot information directly affect the information acquisition and decision-making process of tourists[10]. In the information age, tourists get information about scenic spots through various channels and make travel plans accordingly. At present, some scenic spots have problems such as opacity and asymmetric information, which makes it difficult for tourists to identify effective information, thus affecting their decision-making.

Secondly, some tourists have certain demands for the improvement of tourism infrastructure and services, especially the setting of rest points, the cleaning of the scenic area environment, and the provision of facilities and services such as dining, shopping and transportation will have a greater impact on the tourism experience. The perfection of scenic area infrastructure is directly related to the quality of tourist experience. Good infrastructure can enhance the comfort and satisfaction of visitors. On the contrary, if the infrastructure is not perfect or improperly maintained, it may cause dissatisfaction and complaints from tourists, thus affecting the reputation and image of the scenic spot.

Finally, as a kind of experiential consumption behavior, tourists have more demands for personalized and interactive tourism experience, and they need scenic spots to provide corresponding services to meet their demands. Personalized and interactive experience has been paid more and more attention, and contemporary tourists pay more and more attention to personalized and interactive travel experience. Tourists want to have a unique and engaging experience while traveling, not just a traditional sightseeing excursion.

4.2 Suggestions and countermeasures

First, we need to broaden the information transmission channels to reduce the asymmetry of tourist information. At present, the information of Zhangjiajie National Forest Park is relatively lagging behind. Relevant departments can establish an instant information service platform to facilitate tourists to obtain real-time and comprehensive information. At the same time, infrastructure construction will be strengthened. Maintenance and upgrading of part of the infrastructure, such as increasing the repair of the walkway and expanding the seating capacity. In addition, modern science and technology are used to strengthen the interaction between scenic spots and tourists[11]. For example, the introduction of virtual reality (VR) technology allows visitors to understand and choose the attractions they want to go to with the help of VR devices before entering the park. Scenic spots could set up VR experience stations at specific locations to provide tourists with immersive experience services, allowing them to understand and feel the geomorphological characteristics and formation reasons of Zhangjiajie. Additionally, it's important to reduce the setting of commercial scenes in these areas. Authorities should check and clean up the existing commercial venues that focus on product sales and vulgarization, decrease the commercialization of scenes spots, and appropriately increase scenes themed on folk culture experience and ecological tourism.

As a representative of the experience economy, tourism provides tourists with diversified and personalized service experience, and also gives birth to diversified demands in the process of tourism. This study only discusses the current situation and relevant countermeasures of tourists' demands in Zhangjiajie National Forest Park from the perspective of visual analysis of online comments, hoping to provide relevant reference for optimizing tourism experience and realizing tourism quality upgrading. Limited to the study samples, this study may have some limitations, which need to be further improved and optimized by researchers in the future.

Acknowledgements

This research is supported by Scientific Research Fund of Hunan Provincial Education Department under grant 22A0361, Open Project of Hunan Nature and Cultural Heritage Research Base under grant ZRYC2212.

References

[1] Wu Dan, Wei Wei. Research on the impact of online reviews on brand word-of-mouth communication from the perspective of self-construction [J]. Business Economics Research, 2022(16):98-101.

[2] Xie K L, Zhang Z, Zhang Z. The business value of online consumer reviews and management response to hotel performance[J]. International Journal of Hospitality Management, 2014, 43: 1-12..

[3] WANG Xue, Dong Qingxing, Zhang Bin. User demand analysis framework and empirical research for online reviews: Based on KANO model [J]. Information Theory and Practice, 2002,45(02):160-167.

[4] Xing Yunfei, Cao Gaohui, Tao Ran. Research on Construction and Visualization of Topic Graph of Online User Reviews: A Case Study of Hotel User Reviews[J]. Information Science, 2021(9):101-109.

[5] Filieri Raffaele. What makes an online consumer review trustworthy?[J]. Annals of Tourism

Research, 2016, 58: 46-64.

[6] Chen Limei, Huang Weidong, Chen Chen. The influence path of online reviews on outbound travel purchase intention: Based on the fine processing possibility model [J]. Reform of Economic System, 2019(05):104-112.

[7] Song S, Kawamura H, Uchida J, et al. Determining tourist satisfaction from travel reviews[J]. Information Technology & Tourism, 2019, 21: 337-367..

[8] Du Hengbo, Wang Shaoyun, Luo Rundong. Research on the export effect of Chinese books based on LDA topic model: A case study of Amazon's overseas readers' reviews [J]. Economic Issues, 2024(02):17-23.

[9] Xu Hong, Zhang Siting, Li Lingfang. Topic discovery and hot spot analysis of rural reading promotion based on LDA model and co-word analysis [J]. Information Science, 2002,40(10):67-73.

[10] Tu Haili, Tang Xiaobo. Tourist sentiment analysis model based on online reviews [J]. Modern Information, 2016,36(04):70-77. (in Chinese)

[11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.