Analysis of Tourism Review Information Based on Data Mining Technology

Fengyu Bi

Corresponding author: bifengyu2021@126.com

School of Tourism and Urban Planning, Zhejiang Gongshang University Hangzhou, Zhejiang, 310000, China

Abstract: In recent years, with the rapid development of data mining technology, it has gradually begun to analyze and apply complex data in various fields. Tourism industry is one of the industries with the broadest application prospect of data mining technology. It can excavate and analyze tourist preference and provide marketing decision basis for tourism management department. This paper takes the online tourism review data of 5A scenic spots in Zhejiang Province as the research object to conduct in-depth text mining and analysis. First, the python program is used to capture relevant online review texts on the Ctrip, and the tourism review data set is constructed. By extracting high-frequency words, the preliminary mining and analysis are carried out. Secondly, the supervised learning algorithm based on machine learning is used for further emotion classification and quantification. Then, the LDA topic model and visualization method are used to extract the subject words of the text data, and the subject classification is carried out. Finally, reasonable suggestions are put forward according to the experimental results.

Keywords: Data mining, LDA model, big data analysis, travel review

1. INTRODUCTION

As an industry highly dependent on information data, tourism is one of the industries with the broadest application prospects of big data, which mainly includes UGC data, equipment data and transaction data [1], among which UGC data, which usually explicitly or implicitly records the habits and preferences of tourists, has become one of the main forms of tourism big data [2]. How to dig out the potential value of tourism UGC data has become a hot topic in current research. In terms of mining content, scholars build tourist portraits based on text mining technology [3], tourist satisfaction [4-5], analyze destination perception image [6-7] and other aspects. Relevant research fields mainly focus on small and medium-sized areas such as single scenic spot and single city tourist destination, and lack of research on multiple regional review texts. In terms of mining methods, most researches only focus on content analysis and social network analysis, focusing on word frequency analysis and semantic network analysis, involving ROST CM, DiVoMiner, Gephi, Ucinet and other technical tools [8-10]. The text analysis method is relatively simple and the analysis conclusion is relatively rough. In general, relevant academic research lags behind the practical application of data mining technology in text analysis.

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334275

Therefore, this paper takes the evaluation data of all 5A scenic spots in Zhejiang Province as samples to expand the sample size. At the same time, combined with high-frequency word analysis, emotion analysis and topic extraction, it explores the application of more complex data mining technology in the analysis of tourism review information. In the aspect of high-frequency word analysis, we use the word cloud map of review data to preliminarily understand the attention tendency of tourists. In terms of emotion analysis, this study analyzes the emotional tendency of online reviews of tourist destinations in Zhejiang Province based on emotion dictionary. Through the analysis of positive emotion and negative emotion tendency, the overall satisfaction of tourists to tourist destinations in Zhejiang Province can be further quantified effectively. In the aspect of topic extraction, this paper tries to introduce a machine learning model. Based on subject classification technology in machine learning, irregular UGC data can be classified according to specific topics, thus improving the reliability of subject extraction results of travel review texts [11]. In this paper, Latent Dirichlet Allocation (LDA) topic model is selected. This model is one of the potential topic models in machine learning. It can convert unstructured review text into structured topic propensity value without constructing training data manually. This paper applies the data mining technology to the analysis of tourism reviews, which can explore the concerns and demands of tourists on Zhejiang tourist attractions, and provide some reference suggestions for the improvement and promotion of relevant facilities and services, as well as the formulation of accurate marketing strategies.

2. REVIEW DATA ACQUISITION AND PREPROCESSING

2.1 Comment data acquisition

This paper chooses Ctrip online travel platform as the data source. After comprehensive consideration, 11 5A tourist attractions in Zhejiang Province are selected as the research object, and their user text comments are taken as the data set. In order to ensure the timeliness of comments and better reflect the current situation of scenic spots in recent years, this study selected the user evaluation data in the past three years. This paper uses python software to process the comments on travel websites. The original review data totaled 32990.

2.2 Text data preprocessing

Text preprocessing is carried out by text de-duplication, deletion of invalid text, Chinese word segmentation, removal of deactivation and other measures. Firstly, 32990 crawled comments are processed to remove duplicates and invalid texts, and finally, 13580 valid tourism comment texts are reserved. Secondly, in order to facilitate the subsequent text mining work, the text comments are divided into Chinese words. After text de-duplication and word segmentation, there are still a lot of meaningless and frequent words in the text comment data, which are called "stop words", including quantifiers, conjunctions, modal particles and other types of words, for example: "a", "and", "bar". These words are filtered to speed up the model analysis and improve the accuracy of the model.

2.3 Analysis of high frequency words

In this paper, the keyword is extracted by calling jieba. Using the Word cloud library of Python program, the word cloud of visitors' online comments is made to provide a more intuitive reflection of the results (figure 1). The most obvious words in the picture include: "the scenic spot","the ancient town ", "attractions","place ", "the scenery". It shows that the most important thing for tourists to go out to play is scenic spots, not other factors such as food and accommodation. In addition, words such as "worth","good "and" very"also appear to reflect the feelings of tourists. It is also verified that certain emotional preferences of tourists can be excavated by means of data cleaning, de-noising and normalization. Moreover, and according to these high-frequency words, it probably reflects that tourists have a good overall impression of 5A scenic spots in Zhejiang. Through the production of Word cloud map, it can be roughly seen that tourists' general concerns and attitudes towards 5A scenic spots in Zhejiang Province. In order to dig out more information, then we quantified the emotion.

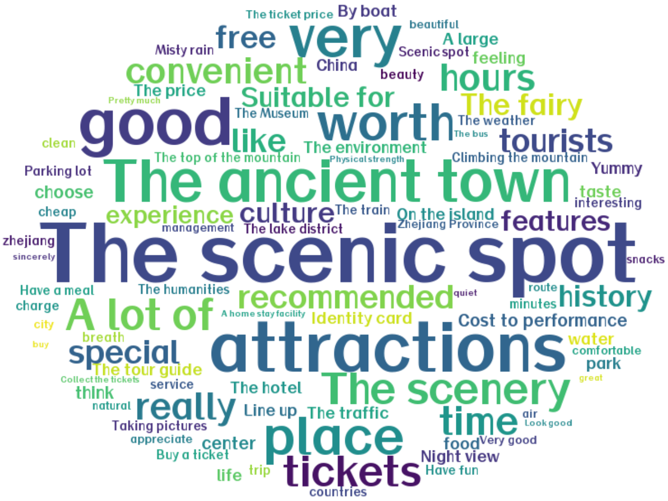

Figure 1: Comment data word cloud

3. SATISFACTION MEASUREMENT BASED ON SENTIMENT ANALYSIS

3.1 SnowNLP

The words and sentences posted by tourists usually have certain emotional tendencies to express their feelings. In this paper, the SnowNLP sentiment dictionary was loaded with python by constructing the sentiment dictionary, and each comment was matched with the sentiment words in the SnowNLP sentiment dictionary. Then, the weighted matching values were added to calculate the sentiment score of each comment and get its specific emotional tendency.

Among them, the process of using SnowNLP library to conduct sentiment analysis on comment data is divided into three steps: The first step is to read the classified negative corpus and positive corpus, and calculate the prior probability of positive and negative polarity through the Naive Bayes model. Naive Bayes basic principle comes from Bayesian theory. The premise of algorithm implementation is to assume that all features are independent from each other, and the weight between features is the same. On this premise, a posteriori probability y is calculated through the joint probability distribution model between features. Naive Bayes algorithm Chinese text b belongs to category A, according to the Bayes definition (formula 1) :

$$p(A|b) = \frac{p(A)p(b|A)}{P(b)} \tag{1}$$

The second step is to calculate the posterior probability of positive and negative polarity of each word for the text data requiring emotional scoring. The third step is to select the category with high probability as the category of positive or negative polarity of the word. Finally, the specific visualization of the total number of tourist comments and emotional rating is shown in the figure below (figure 2), where the probability is [0-0.5], and the judgment is negative; the probability is [0.5-1], and the judgment is positive.

3.2 Emotional quantization result

According to the scatter distribution in the figure, the study found that the distribution of most tourists' emotional tendency is extreme, and the sample size of tourists' positive emotional tendency is much larger than that of tourists' negative emotional tendency, that is, most tourists tend to have a positive evaluation of 5A scenic spots in Zhejiang Province.

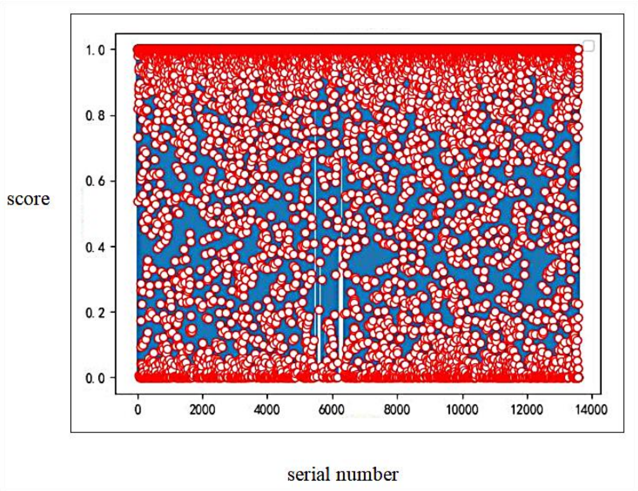

Figure 2: The relationship between the total number of visitors' comments and their emotional scores

The results of relevant statistics on the emotional tendency values are shown in the following table 1:

Table 1： Statistical Results of Emotional Tendency Description

Emotional tendencies	Frequency	Frequency
Positive	11153	82.13%
Negative	2427	17.87%

As can be seen from the above table, in the overall number of tourists' comments, the positive emotional tendency of tourists has the largest number, a total of 11153 comments, accounting for 82.13%; There were 2427 comments with negative emotional tendency, accounting for 17.87% of the total. According to the proportion of 82.13% of the tourists' comments on the positive emotional tendency of Zhejiang Province, most of the tourists are satisfied with the tourist destinations in Zhejiang Province， which indicates that they have a good impression on Zhejiang Province, have high loyalty, have a relatively strong willingness to visit Zhejiang Province, and can generate a good word-of mouth effect on the external publicity of Zhejiang Province. For the 2427 tourists who are lower than the average score of negative emotion, relevant departments should listen to their dissatisfaction with the tourism destinations in Zhejiang Province, analyze and find out the reasons, and take corresponding measures to improve the service, so as to improve their satisfaction with tourism in Zhejiang Province.

4. REVIEW TEXT TOPIC MINING BASED ON LDA MODEL

4.1 LDA topic selection

Because LDA is an unsupervised algorithm, the number of topics needs to be specified manually. In order to help determine the optimal number of topics, this paper first calculates the perplexity of LDA topic model under each topic. Perplexity measures how uncertain the model is about which topic an article belongs to, so the less perplexity, the better the model. As shown in the calculation formula 2:

$$perlexity(\mathrm{D}) = \exp(-\frac{\sum \log \mathrm{p}(w)}{\sum_{d=1}^{M} \mathrm{N}_d}) \qquad (2)$$

Where is the length of the article d, that is, the number of words, and p (w) is the probability of each word, that is, the product of the document topic probability and the topic word probability. In this paper, the number of topics is set from 1 to 20, and the perplexity of the model under each topic number is calculated. The study finds that when the number of topics is more than 9, the perplexity of the model has been decreasing (figure 3), and the model has been fitted. Therefore, the study considers the final number of topics in the 1-9 topic number. And with the help of theme consistency to determine the number of topics.

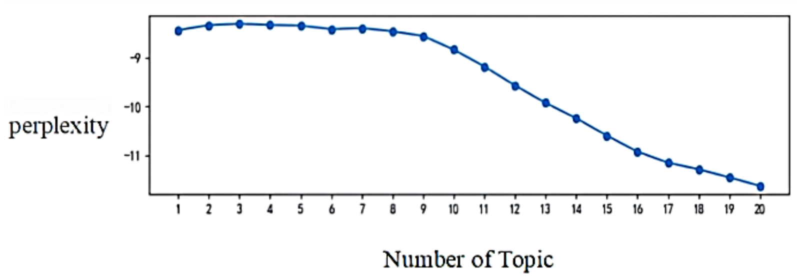

Figure 3: Broken line chart of number of topics and perplexity

Subsequently, the consistency of the themes was tested, and it was found that when the theme was selected as 4, the model scored the highest (figure 4), so the number of themes was selected as 4.

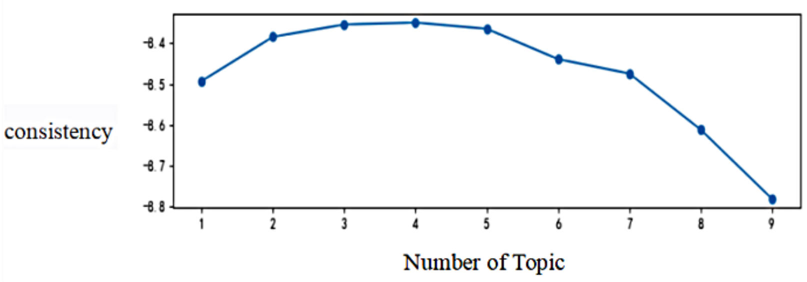

Figure 4: Number of Topics-Consistency Line Chart

4.2 Topic naming and explanation

In this paper, the topic number parameter of the LDA topic model is set to 4, feature extraction is performed on the four selected topics, words that are meaningless to topic judgment are deleted, and the top five main topic feature words of each topic are output. The results are shown in table 2 below:

Table 2: Probability of thematic feature words

Topic 1	Probability	Theme 2	Probability	Theme 3	Probability	Theme 4	Probability
Hour	0.014	Culture	0.008	Feeling	0.008	On the hill	0.008
Ticket	0.008	History	0.006	Convenient	0.006	Temple	0.004
Time	0.008	Characteristics	0.004	Recommend	0.006	Cloud and mist	0.004
Buy	0.006	China	0.004	Like	0.006	Hole	0.003
Queue up	0.005	Zhejiang Province	0.004	Free	0.006	Buddhism	0.003

Each topic can be named and explained based on the high probability feature words in the table above. The study finds that most of the feature words in each topic are related to each other, which proves that the topic mining work is reasonable and effective. Although the theme is condensed according to the same text, the results will be different because of personal subjectivity. However, at present, there is no effective, reasonable and unified way to solve this problem, and the article is temporarily condensed in a manual way the meaning of the theme, and name and explain the theme.

The first five high-frequency feature words in theme 1, including "hour", "ticket", "time", "buy" and "queue up", directly reflect tourists' views on the price of tourist attractions and the importance of cost performance. Theme 1 is named "cost performance"; The words of "culture", "history", "characteristics" ,“zhejiang Province”and "China" in theme 2 are mostly related to internal semantics, which can explain tourists' attention to the characteristics of tourist destinations in scenic spots, so theme 2 is named "scenic spot characteristic"; In theme 3, there are many words expressing tourists' subjective attitudes, such as "feeling", "convenient", "recommend", "like" and "free" and other high-frequency characteristic words, which reflect tourists' experience and feelings of the whole scenic spot when they travel. Therefore, theme 3 is named "overall impression"; High-frequency characteristic words such as "on the hill", "temple", "cloud and mist", "hole" and "Buddhism" appear in theme 4. These words can reflect the basic types of scenic spots in Zhejiang Province. Therefore, theme 4 can be named as "basic type".

4.3 Visual analysis of LDA model mining comments

Based on PyLDAvis visualization tool, visual analysis was conducted on the results of topic mining of LDA model (figure 5). The circles in the left part of the figure represent four themes. Theme 1, theme 2, theme 3 and theme 4 correspond to the cost performance, scenic spot characteristics, overall impression and basic types in table 2. When the mouse moves to topic 1, the feature words of this topic will be displayed on the right in the form of a bar chart list. The length of the red bar chart indicates the word frequency of this topic word in topic 1. For example, "hour", "ticket", "time", "buy", "queue" and so on. Through visual analysis, the general classification of high-frequency feature words and themes can be seen more intuitively.

In addition, the farther away the left circles are, the better the clustering effect of the model is. By observing figure 4, it can be found that the four circles are separated from each other without overlapping, which partially supports the rationality of theme selection and clustering in this study.

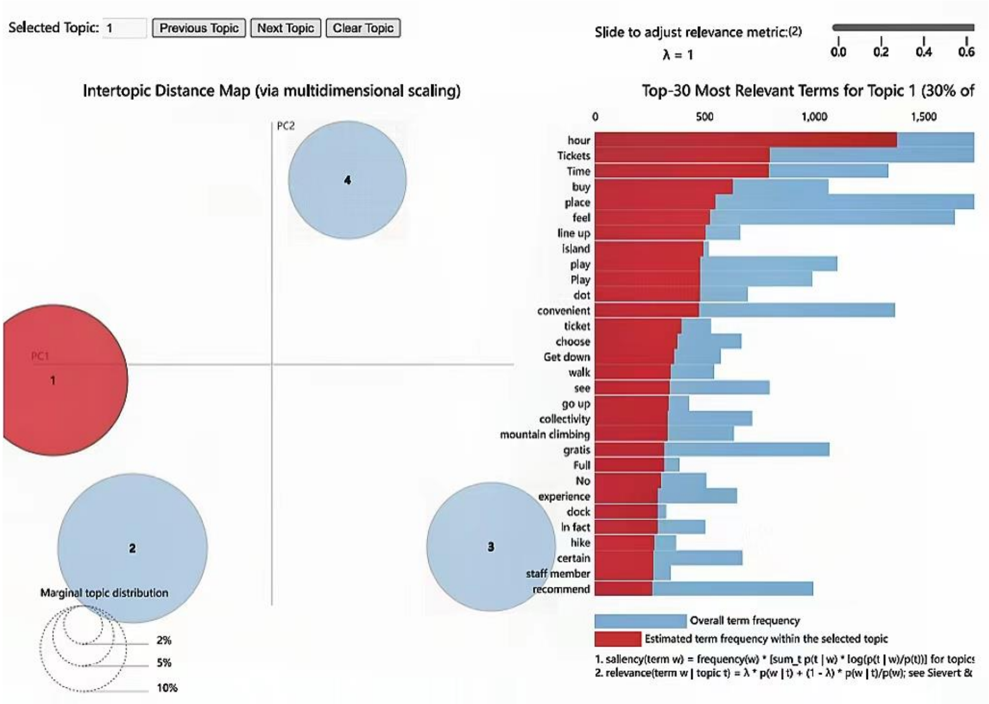
Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (30% of
0
500
1,000
1,500
hour
Tickets
Time
buy
place
feel
line up
island
play
Play
dot
convenient
ticket
choose
Get down
walk
see
go up
collectivity
mountain climbing
gratis
Full
No
experience
dock
In fact
hike
certain
staff member
recommend
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert &

Figure 5: Visualization of Zhejiang Online Review Topic pyLDAvis

5. CONCLUSION

5.1 Conclusion

Aiming at the problems of insufficient depth analysis of tourism reviews in the traditional tourism industry and too small sample size, this paper proposes a review text analysis method combining high-frequency word analysis, emotion analysis and LDA topic clustering to dig deeply into the reviews of Zhejiang tourist attractions. Through the analysis of high-frequency words, the study preliminarily determined the tourists' attention tendency and emotional attitude towards the tourist destinations in Zhejiang Province. Among them, compared with the attention to other factors, the characteristics and quality of scenic spots are crucial, and tourists have a good overall impression on the scenic spots in Zhejiang Province. In addition, by classifying emotions based on supervised learning algorithm in machine learning, this paper further quantifies the emotional tendency of tourists. Through satisfaction measurement of emotion analysis, it is found that the satisfaction of tourists to 5A scenic spots in Zhejiang Province is as high as 82.13%. Finally, the LDA model is used to dig deeply into the review texts, and this result is well presented through the visualization of the LDA model. The experimental results show that tourists pay more attention to four themes during the tour: cost performance, scenic spot characteristics, overall impression and basic type.

5.2 Recommendations

As can be seen from the above results, most tourists' comments on tourist destinations in Zhejiang Province are positive, but there is still room for improvement. This paper suggests that tourism management departments should start from the four themes that tourists pay more attention to, such as cost performance, scenic spot characteristic, overall impression and basic type. On the one hand, further improve management services, standardize ticket price management, pay attention to improve the cost performance of scenic tourism products, focus on adjusting consumer prices in scenic spots, and avoid excessive price rises and overcharging during holidays. On the other hand, we should pay attention to the development of tourism characteristic products. The basic types of 5A scenic spots in Zhejiang Province are mainly natural landscapes. Relevant personnel should carry forward the existing advantages, fully focus on tourism resources and further develop characteristic tourism products on the basis of protecting the regional natural environment and maintaining the advantages of natural scenery. In terms of marketing and publicity, we should pay attention to the positive image building of the destination and improve the overall impression of tourists on the destination. In addition, the relevant administrative departments should do a comprehensive review of the basic types of tourist attractions in Zhejiang Province, design different styles and types of tourism products, and then targeted for the wider population.

REFERENCES

[1] Li J, Xu L, Tang L, et al. Big data in tourism research: A literature review[J]. Tourism Management, 2018, 68: 301-323.

[2] Ye X, Liu W, Li L, et al. Utilizing big data in tourism marketing[C]//3rd International Symposium on Social Science (ISSS 2017). Atlantis Press, 2017: 240-245.

[3] Zhou Jinming, Mao Runze. Research on the construction of tourist portraits in scenic spots based on online travel notes -- taking Shanghai Disneyland as an example [J]. Tourism Forum, 2020,13 (03): 34-45

[4] Geng N N, Shao X Y. Tourist satisfaction of ancient village scenic spot analyzed by fuzzy comprehensive evaluation—A case of House of the Huangcheng Chancellor scenic spot[J]. J. Arid Land Resour. Environ, 2020, 34: 202-208.

[5] Chen Hongling. Research on the satisfaction of tourists in mausoleum scenic spots based on online comments-taking the Ming Tombs in Beijing as an example [J]. Journal of Guangxi University (Philosophy and Social Sciences Edition), 2017,39 (03): 49-53

[6] Zhang G, Li J, Zhang L. A research on tourism destination image perception of Huashan scenic spot: Based on text analysis of weblogs[J]. Tourism science, 2011, 25(4): 87-94.

[7] Lu L, Liao X. Research on image perception of tourism destination based on UGC data: A case study of south mount Heng[J]. Econ. Geogr, 2019, 39: 221-229.

[8] Xu Yonghua, You Xibin, Wang Yanan. Research on tourist satisfaction evaluation based on the ROSTCM method -- taking five domestic terraced scenic spots as examples [J]. Tourism Forum, 2018,11 (05): 22-34

[9] Sujie W, Kaiyi H, Yujie D. A Study on China’s visual tourism images: A social network perspective[J]. Tourism Science, 2018, 32(2): 66-79.

[10] Yan M, Zhao Y. Research of Canal Heritage Tourism Development Based on the Network Text and ASEB Grid Analysis—A Case Study of Qingming Bridge Scenic Area in Wuxi[J]. J. Nanjing Norm. Univ, 2016, 3: 124-129.

[11] Weerdenburg D, Scheider S, Adams B, et al. Where to go and what to do: Extracting leisure activity potentials from web data on urban space [J]. Computers, Environment and Urban Systems, 2019, 73: 143-156.