Research on Tourists Characteristics Based on Big Data Analysis in Cultural Tourism

Siwei Dong and Shan Lu(✉)

Xi'an International Studies University, Xi'an, China
2357741763@qq.com

Abstract. Due to the fierce competition in the cultural tourism market, it is very necessary for tourism enterprises to accurately grasp the characteristics of tourists. In technological contexts, the continued development of information and communication technologies has enabled travel enterprise to gain in-depth knowledge about their consumers. This value is generated from collecting and analyzing user generated content what is termed 'big data'. Based on the multi-dimensional characteristics of online travel agency users, text mining and multinomial logistic regression model are used in this paper to construct tourist portraits in different groups. The result shows that the tourists are mainly divided into four groups, and the differences in the characteristics between various groups are obvious. According to the key user's characteristics, suggestions related to online travel agency user management and tourist attractions promotion are put forward.

Keywords: cultural tourism · big data · online travel notes · tourist portrait

1 Introduction

Due to the rapid development of the tourism industry of China, cultural tourism of has become popular and plays an important role in inspiring national pride, establishes a positive national image, and inherits excellent cultural traditions [1]. However, at present, the competition in Chinese cultural tourism market is fierce. Scenic spots are facing with various problems such as lack of tourists' source variety, poor professionalism in tourism services, insufficient use of tourism brand advantages, and less willingness of consumers to participate in cultural tourism [2]. Therefore, cultural tourism scenic spots urgently need to accurately catch the tourists' needs, enhance consumers' willingness in cultural tourism, improve the quality of tourism service and expand the tourist market.

With the rise of social media and the popularity of mobile devices, Online travel notes are widely published and can provide a reliable information for mining tourists' demands [3]. There have been some practical explorations in the study on tourists' online travel notes which focus on the image of the travel destination, service quality, and the satisfaction of tourists [4–6]. But few studies extract the demands of tourists from the perspective of different consumer group characteristics based on group division. User portrait was first proposed by Cooper A and defined as virtual representation based on real user information [7]. User portraits can effectively classify user groups with

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 181–187, 2023.
https://doi.org/10.2991/978-94-6463-064-0_21

different characteristics and are widely used in mining user demands [8]. However, the current studies in the tourism industry mainly construct tourist portraits based on user basic information [9–11]. The dimension of the tourist portrait constructed by the current studies lacks diversity and it is not able to fully consider the users' basic information and preferences. The construction and application of a multi-dimensional and comprehensive tourist portrait model need to be further improved.

In order to solve the above problems, this paper divides tourist groups based on user preferences. Then constructs multi-dimensional tourist portraits combined with user information. Finally, suggestions are provided for online tourism agencies and cultural scenic spot managers to improve management ability and maintain core users.

2 Method

User portrait is used to describes user characteristics from multiple dimensions [12]. In this paper, four steps are included in the construction of the tourist portrait:

2.1 User Data Preprocessing

Due to the large amount of acquired text data, it is necessary to reduce the noise. First, remove duplicate reviews and meaningless reviews irrelevant to the study. Second, apply Jieba to segment online reviews [13]. Finally, the stop word dictionary is applied to improve the quality of text features [14].

2.2 Topic Extraction

Latent Dirichlet Allocation (LDA) is an unsupervised learning model that can discover latent topics from a large amount of unstructured text contents [15]. Before topic extraction, it is necessary to determine the optimal number of topics. If the number of topics is too less, it will not be able to fucus; if there are too many topics, it will lead to a lack of consistency. In this paper, the perplexity is used to determine the optimal number of topics [16]. Then we use the Latent Dirichlet Allocation package in Sklearn to implement the topic extraction model based on LDA.

2.3 User Group Division

The user similarity calculation can be calculated based on the topic preference distribution probability. Euclidean distance can measure calculate user similarity, which is shown in formula (1).

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

2.4 Extract Group Characteristics and Construct Tourist Portraits

Multinomial logistic regression model is used to identify the typical characteristics of different user groups. And then the tourist portraits of different groups of tourists can be constructed [17]. We suppose that the number of dependent variables is k, a regression equation can be made for each of the $k-1$ variable. Let the logistic regression model of the i-th variable be formula (2):

$$ln\left(\frac{p_i}{1-p_i}\right) = \alpha_{i0} + \sum_{p=1}^{m} \beta_{ip} x_p$$
$$i = 1, 2 \cdots, k-1 \tag{2}$$

3 Results and Discussion

3.1 Data Collection and Preprocessing

Jinggangshan scenic spot is a typical cultural tourism area of China. The data in this paper comes from Ctrip, which is a leading online travel agency in China. We set the time period from June 1th, 2021 to June 1th, 2022 and collected 100 travel notes consisting of 165,785 words and 100 users' information, including each user's gender, the number of followers and follows, the registration time, the number of published travel notes, and the number of travel comments. Then Jieba segmentation tool is used to segment the text content of travel notes.

3.2 Topic Extraction

Determining the optimal number of topics is an important issue before building an LDA model. Perplexity is used to determine the optimal number of topics in this paper. According to the calculation, the optimal number of LDA model topics is 4. Top 8 key words of each topic are shown in Table 1.

Table 1. Top 8 key words for each topic model

Topics	Key words
Topic1	history, assemble, spirit, photo, education, learning, warrior, story
Topic2	rhododendron, Longtan, Shaokou, Bijia Hill, cableway, bamboo forest, telpher, Dujuan Hill
Topic3	hotel, tour guide, airport, train, driver, expressway, bus, hour
Topic4	revolutionary base, former residence, martyr, memorial hall, photography, red flag, cemetery, display

3.3 Division of User Groups

This paper calculates user similarity according to formula (1), and then performs multi-dimensional scaling. According to the classification of the group to which each user belongs, the central user of the group is calculated. By further analyzing the theme participation of users in the center, we can learn about the main topics of different groups in Table 2.

3.4 Tourist Portrait Construction

By significance test, as shown in Table 3, the number of travel comments cannot pass the significance test and should be eliminated.

Different characteristic attribute combinations under the condition of maximum probability of each group are obtained, the predicted probability p of the four groups is greater than 70%, indicating that the accuracy of the model is high. According to the key characteristics of group members, the most likely combination of characteristics of different tourist groups is shown in Table 4.

Table 2. Topic preferences of center tourist

Topic Distribution	Central users			
	Group1	Group 2	Group 3	Group 4
	X97	X94	X49	X44
Topic 1	0.0054673	**0.7509171**	0.0500230	0.1935926
Topic 2	0.0191355	—	0.2882584	**0.6926061**
Topic 3	0.0027336	0.1433441	**0.8470171**	0.0069052
Topic 4	**0.9416636**	0.0080531	0.0502832	—

Table 3. Coefficient significance test

Variables	**Chisquare**	**Significance test**
User Gender (Sex)	4.082	0.046
The number of user's follow (Follow)	12.734	0.005
The number of user's follower (Follower)	6.971	0.033
The number of user's travel notes (Note)	15.565	0.001
The number of user's travel comments (Comment)	4.152	0.246
User's registration time (Time)	58.123	0.000

Table 4. The most possible combination of the same user group

	Group 1	**Group 2**	**Group 3**	**Group 4**
Sex	Male	Male	Female	Male
Follow	The number of follows users is very small, less than 10	The number of follows users is very small, less than 10	The number of follows is more than 50	The number of follows is small, between 10 and 20
Follower	The number of followers is large, more than 500	The number of followers is large, between 100 and 500	The number of followers is small, between 10 and 100	The number of followers is large, more than 500
Note	The number of travel notes is large, more than 100	The number of travel notes is large, between 5 and 20	The number of travel notes is small, less than 5	The number of travel notes is large, more than 100
Time	Join *Ctrip* for more than 15 years, a loyal user of the website	Join *Ctrip* for more than 15 years, a loyal user of the website	Join *Ctrip* for between 10 to 15 years	New users who joined *Ctrip* within 5 years

4 Analysis and Discussion

The users of group 1 are mainly male. These users usually follow a small number of other users but has more followers in *Ctrip*'s online community. They often publish travel notes to share their travel experience, and they joined *Ctrip* for a long time. While traveling this group pays more attention to the historical stories that took place in the scenic area, as well as the spiritual connotation and educational value conveyed by spots. It can be inferred that such users often play an important role as key opinion leaders in the followers group. Online travel agency should pay attention to maintain those users, such as setting travel notes to increase their attention. For cultural tourism managers, it is necessary to fully recognize the commercial value of this group.

The gender of group 2 users is mainly male, and users of this group usually follow a small number of other users. Although the number of travel notes they publish is not large, they have accumulated a certain number of followers in the community, and they have joined *Ctrip* for a long time. While they are traveling, they often pay attention to the artificial landscape of the scenic spot and some facilities of the scenic spot. It can be inferred that these users are producers of high-quality content. For this type of users, *Ctrip* should give them enough space for development, continue to encourage them to produce high-quality content and expand their influence through drainage mechanisms. For scenic spot managers, it is necessary to make full use of the influence of such groups and the high-quality content to expand promotion.

In group 3, the gender of users is mainly female. This group has been in the community for a long time and tents to followed lots of users but publish few travel notes and own few followers. The travel notes they published often more about the arrangements

for hotels, tour guides, and transportation. It is obvious that such users will be fully prepared according to online evaluation information before traveling. For this group, cultural scenic spots managers can register accounts on online travel platforms such as *Ctrip*, actively participate in community, answer questions about cultural tourism and publish scenic photos.

The users of group 4 are mainly male. Although this group has not been in the community for a long time, and the number of users followed is not many, it has published a large number of travel notes and accumulated a large number of followers. They pay more attention to the cultural landscape of cultural tourist attractions. The online platform should encourage them to produce high-quality content. Scenic spots managers can take the initiative to contact such users to give travel packages, hotel discounts, to attract them to participate in cultural tourism.

5 Conclusion and Future Work

In the big date era, effectively use big data technology is of great importance to understand the demands of tourists. By combining the basic characteristics and preferences of tourists, this paper comprehensively constructs the tourist portraits in cultural tourism, which also makes a certain contribution to the related studies on tourism. In future research, the author will consider obtaining more information from online travel agency to further enrich and improve the study.

References

1. Rezaei N. Resident perceptions toward tourism impacts in historic center of Yazd, Iran[J]. Tourism Geographies, 2017, 19(5): 734-755.
2. RUAN Xiao-qing. Research on the development and utilization of red cultural resources from the perspective of inheriting and developing Chinese excellent traditional culture. Leading Journal of Ideological & Theoretical Education, 2017(06): 143-147.
3. TAN Hong-ri, LIU Pei-lin, LI Bo-hua. Perception of tourism destination image in Dalian based on network text analysis. Economic Geography, 2021, 41(03): 231-239.
4. Li X, Geng S, Liu S. Social Network Analysis on Tourists' Perceived Image of Tropical Forest Park: Implications for Niche Tourism[J]. SAGE Open, 2022, 12(1): 21582440211067243.
5. Agostino D, Brambilla M, Pavanetto S, et al. The Contribution of Online Reviews for Quality Evaluation of Cultural Tourism Offers: The Experience of Italian Museums[J]. Sustainability, 2021, 13(23): 13340.
6. Guerrero-Rodriguez R, Álvarez-Carmona M Á, Aranda R, et al. Studying Online Travel Reviews related to tourist attractions using NLP methods: the case of Guanajuato, Mexico [J]. Current Issues in Tourism, 2021: 1–16.
7. Cooper A. The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity. US: Sams Publishing, 2004.
8. ZHANG Han. Research on digital library precision recommendation service based on user portrait. Jilin University. 2019.
9. Nilashi M, bin Ibrahim O, Ithnin N, et al. A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS. Electronic Commerce Research and Applications, 2015, 14(6): 542–562.

10. LIU Yi, CHEN Xin-nuo, BAO Ji-gang, TAN Ke-xin. Tourists' emotional evaluation between natural and cultural attractions. Tourism Tribune, 2019, 34(10): 21-31.
11. LI Qin, LI Shao-bo, HU Jie. Construction and analysis of tourist profile based on joint sentiment-topic analysis. Computer Engineering,2021,1–20
12. ZHOU Li. Research on mobile library user portrait data management strategy based on chain technology. Library Work and Study, 2021(07): 49-57.
13. YOU Zhong-xi, HUA Wei-na, PAN Xue-lian. Matching Book Reviews and Essential Sentiment Lexicons with Chinese Word Segmenters [J]. Data Analysis and Knowledge Discovery, 2019, 3(07): 23-33.
14. Rani R, Lobiyal D K. Performance evaluation of text-mining models with Hindi stopwords lists [J]. Journal of King Saud University-Computer and Information Sciences, 2020.
15. Nikolenko S I, Koltcov S, Koltsova O. Topic modelling for qualitative studies. Journal of Information Science, 2017, 43(1): 88-102.
16. Huang L, Ma J, Chen C. Topic detection from microblogs using T-LDA and perplexity [C]//2017 24th Asia-Pacific Software Engineering Conference Workshops (APSECW). IEEE, 2017: 71–77.
17. LIN Yan-xia, XIE Xiang-sheng. User portrait of diversified groups in micro-blog based on social identity theory. Information Studies: Theory & Application, 2018, 41(03): 142-148.