Mining Influential Factors and Spatio-Temporal Patterns of Travel Intention Based on Social Media Data

Hang Zhao *

* Corresponding author: zhaohang22@mails.ucas.ac.cn

Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, China;University of Chinese Academy of Sciences, Beijing, China

Abstract: The questionnaire method is becoming obsolete. Social media data contains textual, location and temporal information, making it possible to more effectively discover the factors influencing large-scale public intention to travel and to conduct spatio-temporal and thematic analysis. Based on Sina Weibo data posted by users who wanted to visit Xinjiang throughout 2022, we used latent Dirichlet allocation(LDA) models, term frequency–inverse document frequency(TF-IDF) algorithms, spatial autocorrelation and other theories and techniques to mine the data, and obtained the following conclusions:we got seven topics,and beautiful view was the factor with the highest proportion;tourists were driven by anticipation in the early period (January-May), and by specific things about the tourist destination in the peak period (June-August), and in the late period (September-December) had resistance susceptibility;focus on gourmet food was stronger in the west and south, and beautiful view was stronger in the more economically developed regions.

Keywords:LDA, TF-IDF, travel intention,spatial autocorrelation

1. INTRODUCTION

Past studies have shown that social media big data analysis has a wide application prospect in the field of tourism [1, 2]. However, existing studies have mainly focused on the evaluation of tourist destinations [3], sentiment analysis of travelogues [4] and recommendation of tourist attractions [5], and there is a great lack of in-depth mining of tourists' travel intention based on social media data [6]. Therefore, this study will fill this gap and provide the development of tourism industry by providing new ideas and methods.

Scholars have already tried to excavate and analyze the factors influencing travel intention [7] by using push and pull theory and other means [8]. However, existing studies have problems in utilizing the traditional questionnaire method, which lacks refinement and quantitative analysis [9]. The questionnaire can only obtain the respondent's thoughts at a certain moment, but we do not know whether the thoughts change over time or space [10] and have no idea of What is the proportion of these factors, and are these factors subject to spatial and temporal conditions [11] and other questions are not answered precisely. In the context of the current big data era, data has become the most competitive asset. The public's access to destination information has become more diverse and the means of expressing travel intentions more convenient [12]. Social media has accumulated a huge amount of disordered and ponderous but valuable information

ICEMBDA 2023, October 27-29, Tianjin, People's Republic of China
Copyright © 2024 EAI
DOI 10.4108/eai.27-10-2023.2341916

and real data containing travel intentions [13] that has not yet been fully utilized and tapped [14] and has greater potential value. Social media data has the ternary characteristics of location-time text [15], which makes it possible to conduct spatio-temporal and thematic analysis of tourists' travel intentions. Therefore, this paper hopes to obtain the sensitivity of each region to specific topics [16], the sensitivity of each topic to temporal changes [17], which topics account for the largest proportion at which times, which regions have the largest number of tweets on a certain topic, etc. through the analysis of social media data [18]. This analysis method is different from the neglect of spatio-temporal information in past questionnaires, and can help tourism departments make more effective decisions in response to changing spatio-temporal conditions to make more effective decisions, provide better tourism services, and assess the potential of the tourism industry [19].

2. DATA & METHODS

In this paper, we use Sina Weibo data with geographic location information for the year 2022, and based on the high-frequency words contained in "Xinjiang tourism" and related hyperlinks, we decide to use "want to go to Xinjiang" as the keyword through Sina Weibo API and the written Python code [20] was used to retrieve 16,841 tweets with the keyword "want to go to Xinjiang", including post location, post time, and text.

First, the collected comment data is cleaned and processed. We remove emoticons, numbers, English characters, punctuation marks, special characters and URL links from the text, and use the stuttering word separation system to split the text and eliminate some meaningless stop words (such as "had", "it" "in", etc.). Second, the latent Dirichlet allocation (LDA) topic model was used to analyze the text [21]. LDA topic modeling is an unsupervised machine learning technique that identifies latent topics in a large document collection or corpus and gives the distribution of words with high probability of occurrence (i.e., high relevance to the topic) under each topic. The confirmation of the number of LDA topics follows the elbow rule [22] to calculate the perplexity values, and the lower the perplexity values, the better the performance of the model clustering.While using the elbow method to identify seven topics, the manual annotation method of the literature [23] was used to continue the segmentation and inspection, and finally a total of 12,664 valid tweets containing category information were obtained, and the classification accuracy of each topic reached 94.2%,99.5%, 98.4%, 94.0%, 98.3%, 97.9%, and 93.3%, respectively.

3. RESULTS AND ANALYSIS

3.1. Topic mining results

The clustering results we obtained by writing the crawler along with the description of each category and the feature words obtained by the TF-IDF algorithm (taking the top three) are shown in Table 1.

Table 1. Clustering results and description

clustering result	category description	top 3 of TF-IDF
escape	Including those that want to travel to avoid familiar environments, boring work, certain people or things, etc.	don't want, go to work ,work
personal wish	Including those that want to travel because of their inner desires, such as increasing experience, going to Xinjiang to graze sheep, supporting education, enjoying freedom and so on	go out to play,ski,self driving tour
gourmet food	Including those that want to travel because of the desire to enjoy the delicious food	fruit,rice flour, grape
beautiful view	Including those that want to travel because of the desire to enjoy the beautiful view	grassland,so beautiful,snow mountain
media induced	Including those that want to travel because of the information on the Internet	tourists,Running Man,Dilraba
epidemic	Including those that can't start traveling because of epidemic prevention and control, and the fear of virus infection	epidemic,end, hope
conditions	Including those that want to travel but are limited by factors such as lack of money, time, or friends	no money,time, friend

For convenience, we will refer to the above topics in order as Topic I(escape),Topic II(personal wish),Topic III(gourmet food),Topic IV(beautiful view),Topic V(media induced),Topic VI(epidemic),Topic VII(conditions).

3.2. Temporal pattern mining results and analysis

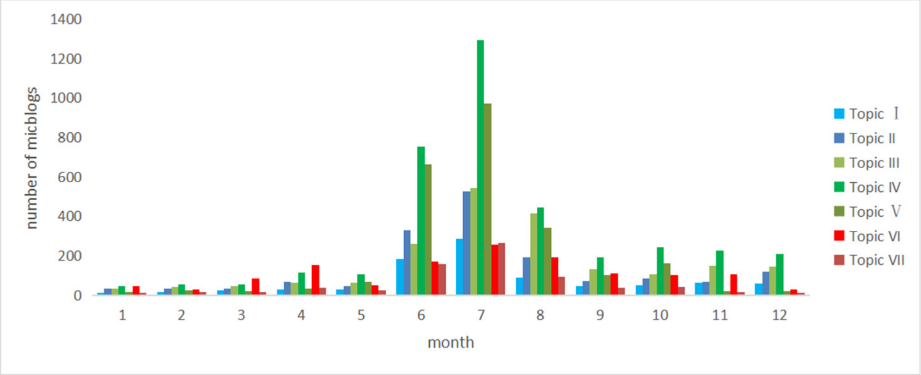

Figure 1: Number of micblogs per month on each topic

It can be seen from Figure 1 that the change of the total number of microblogs presents an inverted "V" shape.According to the above figure, we divided the generation period of travel intention into three periods through breakpoint analysis, and name as the early period (January -May), , the peak period (June -August), and the late period (September -December).Applying the TF-IDF algorithm, keywords of each period and topic are shown in the following table，We hope to summarize the travel intention driving mechanisms of tourists in the three periods through Table 2.

Table 2. Characteristic words for each topic in the three periods

month	Topic I	Topic II	Topic III	Topic IV	Topic V	Topic VI	Topic VII
1 -5	don't want, when, work	skiing, hanging out, hope	rice noodles, food, time	Tibet, look, Yunnan	travel, video, film	epidemic,end, time	graduation, life, work
6-8	go out to play, go to work, work	go out to play, time, place	fruits, grapes, rice noodles	Tibet, Yunnan, grasslands	travel, Dilraba, Running Man	epidemic, travel, end	travel, graduation, Tibet
9 -12	don't want, work, time	ski, hang out, place	travel, delicious food, rice noodles	Tibet, Yunnan, Changbai Mountain	Li Bingbing, Douyin, Travel	epidemic,end, time	travel, life, work

In the early period, generally speaking, the public in this period is mainly looking forward to it, and they are not so strong about traveling. In the peak period, we can see that the intention to travel at this time is driven by some specific things (such as grasslands, rice noodles, and the star effect of Dilraba). And in the this period, it can be seen that despite the resistance factors, most people still focus on going to travel, the intention to travel at this time is more violent, spontaneous and herd-like. In the late period, we can see that the public's intention to travel is still not low, but it is limited by some factors (such as the epidemic situation and work). If these suppressions reduce, the high tourist intention may continue from August.

3.3. Spatial pattern mining results and analysis

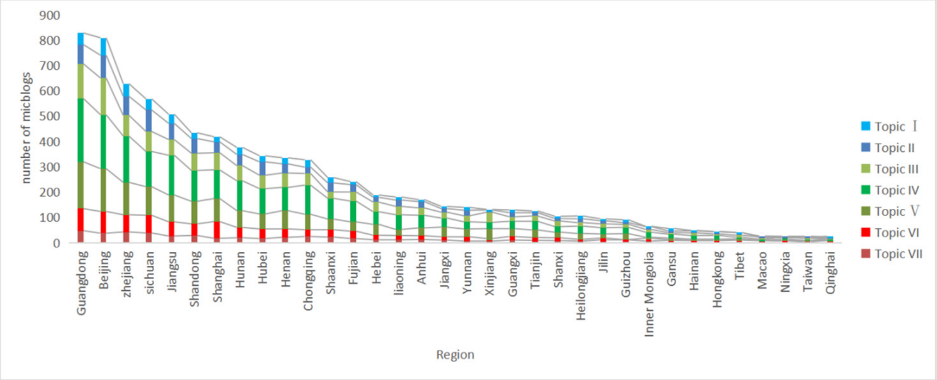

Figure 2: Number of micblogs on each topic by region

The statistics of the total number of microblogs posted and the spatial distribution of the number of topics are shown in Figure 2. The spatial distribution of different topics reflects that the number of microblogs is higher in the economically developed regions in the middle and east than in the central and western regions, and there are similar spatial differences between topics, i.e., coastal regions are generally higher than central regions, and central regions are generally higher than western regions; northeast and northwest regions are generally lower than southwest and southeast regions.

Next, we used spatial autocorrelation analysis and used the Moran's I as a parameter to mine spatial patterns. Since the popularity rate of microblogs and the number of people willing to post travel-related microblogs are not consistent in each region, it is not reasonable to use the value obtained by dividing the number of microblogs on topic i by the number of microblog users to calculate the Moran's I [24]. Therefore, we decided to use the number of topics as a percentage to calculate the Moran's I, and the calculation results are shown in Table 3.

Table 3. Spatial autocorrelation statistics on each topic

	Topic I	**Topic II**	**Topic III**	**Topic IV**	**Topic V**	**Topic VI**	**Topic VII**
Moran's I	-0.220	0.061	0.262	0.348	-0.042	-0.065	-0.021
Z	-1.613	0.813	2.551	3.294	-0.078	-0.273	0.104
q	0.053	0.208	0.005	0.000	0.469	0.393	0.459
Spatial autocorrelation	N	N	Y	Y	N	N	N

In the analysis of the global Moran's I, the original hypothesis was that there was no spatial correlation between the regions, here the p-value for Topic III(gourmet food) and Topic IV(beautiful view) is less than 0.05, thus indicating that the original hypothesis is rejected.Therefore,there is a spatial correlation between these two topics.

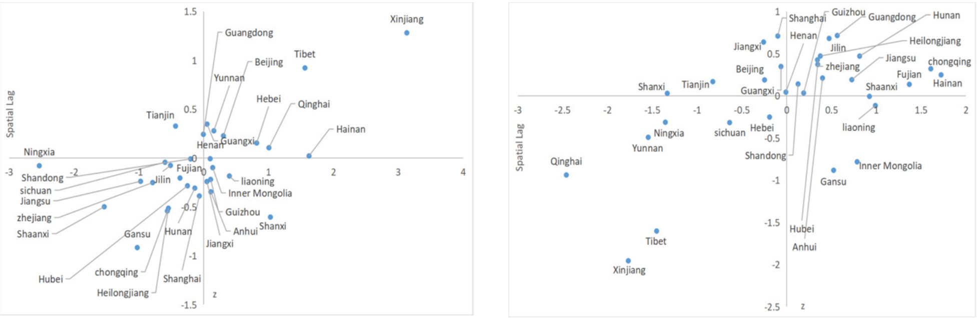

Figure 3: Moran scatter plot(Topic III on the left, Topic IV on the right)

The Moran scatter plot(Figure 3) shows the scatter relationship between the spatial deviation z-value and the spatial lag term Spatial Lag. The scatter plot is divided into four quadrants, with quadrants 1 and 3 being positive spatial correlation and quadrants 2 and 4 being negative spatial correlation.We can get that the regions in figure "gourmet food" and figure "beautiful view" are mostly located in the first and third quadrants, indicating that the regions show positive spatial correlation, among which in "gourmet food In "gourmet food", the first quadrant contains mostly western and southern regions, indicating that these regions pay more attention to gourmet food and have synergy with neighboring regions; while the third quadrant contains mostly central and northern regions, indicating that these regions pay less attention to gourmet food. In the "beautiful view", the first quadrant mostly includes the central, northern, southern, eastern and other economically developed regions, indicating that these regions pay more attention to beautiful view and have synergy with neighboring regions; while the third quadrant mostly includes the western regions, indicating that these regions pay less attention to beautiful view.

4. CONCLUSION

Through social media data mining, we have found that beautiful view is the most important factor. The peak period of travel intention is from June to August, and tourists are driven by expectations in the early period (January -May), while the peak period (June -August) is driven by specific things in the tourist destination. During the peak period, tourists have strong spontaneity and conformity in their travel, and are susceptible to resistance in the late period (September -December). The result of spatial mining is that travel intention is higher in economically developed regions than in less developed regions, with strong attention to gourmet food in the western and southern regions, and weaker in the central and northern regions. However, economically developed regions such as the central, northern, southern, and eastern regions have a stronger focus on beautiful view, while the western regions have a weaker focus.

5. DISCUSSION

Compared to previous studies, this paper has distinctive innovations in the use of technology. Firstly, by applying text mining techniques to potential non-travelling tourists, it is easier to harvest a larger sample and conduct spatio-temporal analysis than the previous use of questionnaires. Secondly, the application of text mining techniques to non-travelling potential tourists is more effective than the application of text mining techniques to reviews of tourist places [23] in uncovering the main attractions of tourist places and the reasons why tourists want to go but are unable to do so. In terms of research implications, this paper presents textual variations of different themes on temporal and spatial scales, enabling easier synergistic analysis with other social factors, and providing data, ideas and methodological support for future deeper analysis of dynamical models and precise regulation between tourism drivers and deterrents and social development.

REFERENCES

[1] Krsak, B., & Kysela, K. (2016). The use of social media and Internet data-mining for the tourist industry. Journal of Tourism and Hospitality, 5(1).

[2] Vu, H. Q., Li, G., Law, R., & Zhang, Y. (2018). Tourist activity analysis by leveraging mobile social media data. Journal of travel research, 57(7), 883-898.

[3] Kaur, A., Chauhan, A., & Medury, Y. (2016). Destination image of Indian tourism destinations: An evaluation using correspondence analysis. Asia Pacific Journal of Marketing and Logistics.

[4] Nawijn, J., & Biran, A. (2019). Negative emotions in tourism: A meaningful analysis. Current Issues in Tourism, 22(19), 2386-2398.

[5] Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I., & Woodward, J. (2013). A context-aware personalized travel recommendation system based on geotagged social media data mining. International Journal of Geographical Information Science, 27(4), 662-684.

[6] Zeng, B., & Gerritsen, R. (2014). What do we know about social media in tourism? A review. Tourism management perspectives, 10, 27-36.

[7] Shapoval, V., Wang, M. C., Hara, T., & Shioya, H. (2018). Data mining in tourism data analysis: inbound visitors to Japan. Journal of Travel Research, 57(3), 310-323.

[8] Regmi, P. R., Waithaka, E., Paudyal, A., Simkhada, P., & van Teijlingen, E. (2016). Guide to the design and application of online questionnaire surveys. Nepal journal of epidemiology, 6(4), 640.

[9] Jing, P., Wang, B., Cai, Y., Wang, B., Huang, J., Yang, C., & Jiang, C. (2023). What is the public really concerned about the AV crash? Insights from a combined analysis of social media and questionnaire survey. Technological Forecasting and Social Change, 189, 122371.

[10] Li, Y., Yang, L., Shen, H., & Wu, Z. (2019). Modeling intra-destination travel behavior of tourists through spatio-temporal analysis. Journal of destination marketing & management, 11, 260-269.

[11] Hall, C. M. (2005). Time, space, tourism and social physics. Tourism Recreation Research, 30(1), 93-98.

[12] Han, J., & Chen, H. (2022). Millennial social media users' intention to travel: the moderating role of social media influencer following behavior. International Hospitality Review, 36(2), 340-357.

[13] Sigala, M., Christou, E., & Gretzel, U. (Eds.). (2012). Social media in travel, tourism and hospitality: Theory, practice and cases. Ashgate Publishing, Ltd..

[14] Hur, K., Kim, T. T., Karatepe, O. M., & Lee, G. (2017). An exploration of the factors influencing social media continuance usage and information sharing intentions among Korean travellers. Tourism Management, 63, 170-178.

[15] Han, X., Wang, J., Zhang, M., & Wang, X. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China. International journal of environmental research and public health, 17(8), 2788.

[16] Häberle, M., Werner, M., & Zhu, X. X. (2019). Geo-spatial text-mining from Twitter–a feature space analysis with a view toward building classification in urban regions. European journal of remote sensing, 52(sup2), 2-11.

[17] Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies. International Journal of Digital Earth, 12(11), 1230-1247.

[18] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

[19] Melian-Gonzalez, A., & García-Falcón, J. M. (2003). Competitive potential of tourism in destinations. Annals of Tourism Research, 30(3), 720-740.

[20] Zhou, Z. (2014). Data crawler for Sina Weibo based on Python. Journal of Computer Applications, 34(11), 3131.

[21] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.

[22] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.

[23] Feng, Z. Q., Peng, X., & Wu, Y. Z. (2022). Tourists Emotion Perception Based on Social Media Data Mining [J]. Geography and Geo-Information Science, 38(01), 31-36.

[24] Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geographical analysis, 39(4), 357-375.