Research on Tourism Network Index Model Based on Baidu Index --- A Case Study of Sanya

Caixia Chen[1], Chun Shi[1] and Jue Chen[2]
[1]*School of Information Science and Technology, Hainan Normal University, Haikou 570206, P.R. China*
[2]*School of Economic Management, Hainan College of Vocation and Technique, Haikou 570105, P.R. China*

Abstract. Tourism index is a "barometer" to reflect the overall development level of tourism. The tourism index compiled by historical data can not reflect the real situation accurately with the increasing influence of network events on tourism In the Internet era. This study collects time series data of tourism network search by Baidu index tool and uses data mining method and principal component analysis method to detect and standardize the stability of the data. The spss system and the weighted analysis method are used to construct the tourism network index model. Finally, the model detection is carried out by comparing the actual tourism data in Sanya. This study is an important supplement to the existing tourism index.

1 Introduction

The tourism network index is a collection of emotions, opinions, attitudes, viewpoints, etc. which are expressed by Internet users around tourism emergencies and hot spots. CNNIC research shows that More than 80% of tourism users obtain tourism information through the Internet. Big data on the Internet has a stronger tendency, immediacy and predictability than the historical data lagging behind. This study collected more than 3 years time series data of tourism related keywords in Sanya by Baidu index tool, detected and standardized the data using data mining method and principal component analysis method, using spss system and adding the right analysis method constructs the tourism network index model, finally carries on the model inspection through the contrast Sanya actual tourism data. This study is an important supplement to the existing tourism index.

2 Keyword selection

Different keywords have different search frequency. The number of keywords must be rich and comprehensive.

(1) Through online and offline questionnaires and the <meta> tag of the top-ranked travel website HTML in recent years, the basic keywords of Sanya tourism were selected: Tianya Haijiao, Weizhizhou Island, Yalong Bay, Nanshan Temple, Luhuitou, Haitang Bay, Coconut Dream Corridor, Penang Valley, Dongtian, Dadong Sea, West Island.

(2) Using the "Station Master's House" tool, 12 basic keywords were extracted and 124 extended keywords were obtained.

(3) Using the SPSS system to analyze and collate the statistical results, the total search index accounted for 79.5% of the total 25 key words, namely: "Sanya Tourism Strategy", "Weizhizhou Island", "Tianya Haijiao", "Yalong Bay", "Sanya Tourism", "Yalong Bay Seabed World", "to Sanya tourism more. "Little money", "Haitang Bay", "Sanya Tourist Attractions", "Nanshan Temple", "Yalong Bay Tropical Paradise Forest Park", "West Island", "Where is the End of the World", "Penang Valley", "Sanya Yalong Bay", "Sanya Begonia Bay", "Coconut Dream Corridor", "Big and Small Cave", "End of the World Tickets"," "Sanya Weizhizhou Island", "Tianya Haijiao Picture", "Luhuitou", "Sanya Nanshan Temple", "Sanya Tianya Haijiao", "Sanya Dadong Sea".

3 Constructing tourism network index model

3.1 Collecting keywords, web search data

(1) Enter the Baidu Index page, through the input of 25 keywords, get the corresponding keywords in 2014-2016 time series data map.

(2) Collecting time series data by Octopus harvester.

3.2 Data detection

The time series data of 25 keywords from 2014 to 2016 were input into SPSS system and tested by KMO and Bartlett. The results showed that KMO = 0.761, 0.7 < KMO = 0.761

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

< 0.8, and the weight could be calculated by principal component analysis.

3.3 Constructing tourism network index model

3.3.1 Principal component analysis

Principal component analysis showed that the characteristic roots of the four principal components of "Sanya Tourism Strategy", "Weizhizhou Island", "Tianya Haijiao" and "Yalong Bay" were more than 1. The cumulative variance contribution rate of the first two principal components was 84.522%, more than 80%. Therefore, the first four principal components can basically reflect the information of all the indices, and can replace the original 25 indices ("Sanya Tourism Strategy", "Weizhizhou Island", "Tianya Haijiao", "Yalong Bay", "Sanya Tourism", "Yalong Bay Seabed World"...).

3.3.2 Correlation coefficient

The number of loads, or factor loads, represents the load of the first variable on the j common factor, reflecting the relative importance of the second variable on the j common factor.

3.3.3 Determine weight

Principal component analysis is used to determine the weight, that is, the index weight equals to the weighted average of the coefficients in the linear combination of the principal components with the variance contribution rate of the principal components.

(1) Coefficient of index in linear combinations of different principal components.

The linear combination of the four principal components is as follows:

F1=0.176 X1+0.166 X2+0.258 X3 +……+0.194 X25
F2=0.258 X1-0.342 X2-0.103 X3 +……+0.299 X25
F3=0.120 X1+0.105 X2-0.094 X3 +……+0.015 X25
F4=0.142 X1-0.097 X2-0.126 X3 +……+0.078 X25

(2) According to the variance contribution rate of principal components, the coefficients of the comprehensive model are obtained.

"Initial eigenvalue" of the "variance%" represents the principal component variance contribution rate, the greater the variance contribution rate, the greater the importance of the principal component.

Variance contribution rate of four principal components

The coefficient of index is the index in the linear combination of the four principal components.

The comprehensive coefficient of "Sanya tourism strategy":

$$\frac{0.176*50.070+0.258*19.902+0.120*8.159+0.142*6.390}{50.070+19.902+8.159+6.390}$$

Similarly, the coefficients of all indicators are calculated. The comprehensive score model is as follows:

$$Y = 0.187x1 + 0.021x2 \ldots + 0.181x25$$

(3) Normalization of data

Because the sum of the weights of all the indexes is 1, the index weights need to be normalized on the basis of the index coefficients in the comprehensive model.

Eg: weight of "Sanya tourism strategy"
=0.187/(0.187+0.021+……+0.146+0.181)=0.063.

(4) Sanya tourism network index model the weights obtained can indicate the relative importance of the corresponding keywords in the whole. The weights can be used to synthesize the keywords comprehensive index. The JZZS is used to represent the Sanya scenic area comprehensive index. The 25 keywords are X1, X2, X3, X4... The formula for JZZS is as follows:

JZZS=0.063*X1+0.007*X2+0.037*X3+0.015*X4+……+0.049*X24+0.067*X25

The comprehensive index of Sanya scenic area for 2014-2016 years is obtained (Tab1).

Table 1. Key words of Sanya scenic spot in 2014-2016.

Mon	2014 (number of people)	2015 (number of people)	2016 (number of people)
1	1023.203	1284.866	1381.13
2	1367.192	1400.364	1521.406
3	1013.036	1314.363	1227.428
4	919.817	1097.39	1077.103
5	902.504	1005.995	1070.073
6	940.215	1010.439	1054.287
7	1035.879	1190.755	1257.51
8	964.438	1131.542	1194.847
9	917.993	922.318	1201.271
10	1038.281	1005.011	1331.536
11	1000.593	1088.016	1206.531
12	1120.28	1237.549	1394.156

4 Actual tourism data of Sanya

According to the statistics provided by Hainan, the number of visitors to Sanya 2014-2016 (monthly) is shown in Table 2.

Table 2. Number of tourists in Sanya in 2014-2016

Mon	2014 (ten thousand	2015 (ten thousand)	2016 (ten thousand)
1	126.32	135	147.64
2	133.2	149.83	162.57

3	122.08	132.29	144.34
4	98.5	109.88	117.73
5	88.2	101.93	109.02
6	80.54	88.49	100.42
7	92.61	114.63	114.63
8	104.56	119.81	134.12
9	86.43	94.47	106.82
10	107.14	116.52	130.53
11	140.93	154.61	171.24
12	172.25	190.78	212.5

5 Empirical analysis

5.1 Correlation analysis

By processing the comprehensive index and comparing it with the actual number of tourists, the graph shows the search volume of combination keywords and the tourist flow graph of March from January 2014 to December 2016. From the graph, we can see the change trend of search volume of combination keywords and the change trend table of tourist flow in March. There is a strong consistency (Fig1).

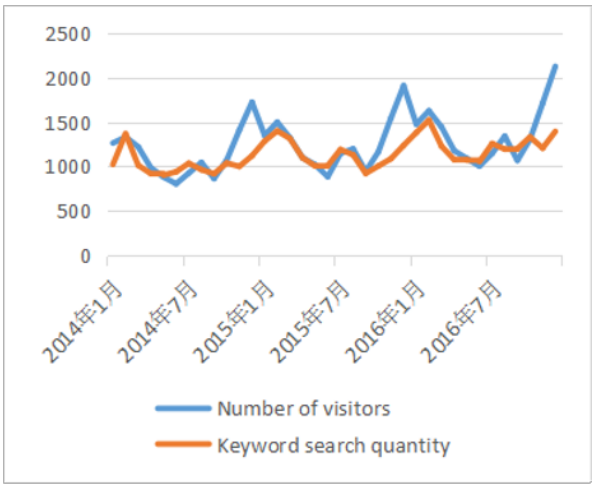

Figure 1. Comparison of keyword search volume and Sanya tourist volume trend

Correlation analysis showed that the correlation coefficient was 0.820, significant P = 0.000 < 0.01, with statistical significance, which proved that there was a correlation between the two. It is proved that the actual traffic volume of Sanya can be reflected through Baidu index data.

5.2 T test

In order to further mine the characteristics of Baidu index and Sanya tourist volume, this paper tests three groups of data from 2014 to 2016. The results show that from 2014 to 2016, the Sanya composite index and the average and standard deviation of the actual number of people showed an increasing trend. The contrast between Sanya's actual tourist season and the peak season is getting bigger and bigger. The sig value is less than The significance level 0.05, and The sig value becomes smaller over time. The correlation coefficient showed an increasing trend and the correlation increased year by year.

5.3 Predictive analysis

Excel is used to deal with the statistics of the actual number of people and the composite index, and the resulting curve is transformed and translated to the following results:

By shifting the composite index one unit forward on the coordinate axis, it can be seen from the graph that after the translation transformation, the two curves of SJRS and ZHZS are more consistent, and the peaks and turning points are more consistent. ZHZS moved forward by a unit, in fact, ZHZS ahead of a month, we can see that ZHZS for the number of visitors to predict a certain premonition, about a month in advance.

6 Conclusions

(1) The search volume of key words in Sanya scenic spot combination is related to the monthly tourist flow in Sanya, and with the passage of time, social progress, information technology developed, this correlation shows a growing trend.

(2) Of all the relevant keywords, the search volume of "Sanya Tourism Strategy" is far ahead, the number of individual tourists exceeds the number of team tourists, and the proportion of individual tourists is increasing year by year.

(3) The dispersion of the actual number of tourists in Sanya is increasing year by year, and the contrast between the actual tourist season in Sanya is getting bigger and bigger, which is very unfavorable to the use of resources. The peak season scenic spot pressure is enormous, for Sanya traffic and other aspects of great challenges, but the off-season is bleak business.

(4) Through the analysis of the comprehensive index and the actual number of people in Sanya, the data of Baidu index can predict the actual number of people in the scenic spot, and has a certain lead time. According to the statistical results of this paper, the lead time is about one month.

Acknowledgments

Thanks to the support by NSF of Hainan Province (No.617110 and No.617121).

References

1. Yan Lihua.A Summary of Research on Tourism Information search behaviour [J]. Tourism overview, 2015 (12).
2. Iresearch. China online Travel booking Industry Development report 2010-2011 [R]. Beijing: iresearch Group, 2011.

3. Wu Zhizhun. The Study of Competitiveness Evaluation and Development Countermeasure in Sanyafs International Tourism Service [D]. Hainan University, 2015.
4. Ginsberg, Mohebbi, Patel, Brammer, Smolinski, Brilliant. Detecting influenza epidemics using search engine query data [J]. Nature, 2009, 457: 1012-1014.
5. Xin Yang, Bing Pan, James A. Evans, Benfu Lv, Forecasting Chinese tourist volume with search engine data [J]. Tourism Management, Volume 46, February 2015.
6. China Internet Network Information Center. 24th Statistical Report on Internet Development in China [R]. Beijing: China Internet Network Information Center, 2013.