Advances in Computer Science Research, volume 62
2nd Joint International Information Technology, Mechanical and Electronic Engineering Conference (JIMEC 2017)

Scenic Spot Tourists Flow Prediction Research Based On Web Search Items

Fu Tian , Wang Zhen (✉corresponding authors) , Xun Song Ming
Qionghai， Hainan，China Hainan College Of Software Technology 571400
fu-tian@163.com, w.rainie@163.com

Keywords: BigData; Search Index; Support Vector Machines; Data Test; Regression.

Abstract

In the context of "smart tourism" for large data applications, the Internet search engine records a large number of people searching for data. Compared with the official statistics, the search data has the characteristics of high efficiency and low cost. In this paper, we will explore and analyze the relationship between network search terms and scenic tourist numbers, analyze the theory of support vector machine in time series forecasting, and propose a support vector machine (SVM) algorithm for the forecast of scenic area passenger flow. The forecasting of the tourist flow rate of the local tourist area is carried out by the method of controlling the number of support vectors to reduce the calculation amount of the algorithm. Finally, the results show that the model has good prediction precision and can be used to predict the tourist attractions' scenic passenger flow.

1 Introduction

With the rapid development of China's economy, more and more people try to enhance the level of spiritual needs through tourism.So in tourist season some scenic spots are often overcrowded, tourism resources can not meet the demand of tourists which leads to the decline of scenic tourism services, but tourism resources are idle in some scenic spots. Therefore, how to predict the tourist quantity scientifically and accurately, combined with the carrying capacity of the scenic spot, scientific and rational distribution of tourism resources and the development of response measures, to optimize service processes and improve the quality of scenic spots and promote the sustainable development of scenic resources is particularly urgent and important. With the rapid development of the Internet, people are increasingly relying on Internet search engines, media platforms to make travel plans. While the Internet social media provide the relevant travel information for tourists, the Internet search engine, online forums and social media have become the focus of attention for tourists.

Tens of millions of travel search keywords and browsing records are accumulated in the Internet search engine, online forums and social media platform. These data not only have the characteristics of large capacity, Variety, Velocity and Veracity[1-2], but also easy to obtain data and low cost. We can excavate and analyze these data, select the appropriate sample data to train the appropriate forecasting model which can be a good prediction of the changes of numbers of tourists in the tourist area. With the use of statistical data to predict, it has incomparably superiority.

2 The Research Status in China and Abroad

At present, with the formation of large data, Internet research shows that through the analysis of accumulation of keyword on the Internet, browsing trace data, it can effectively predict the development trend of social behavior. For example, in 2009, Ginsberg et al. successfully predicted the trend of influenza outbreak by using the network search term data, which was 2 weeks earlier than the traditional prediction method. It proved that the network search term data had the ability to forecast the flu outbreak. Yang Shuxin, Dong Jichang, and others through the Google search keywords and the relationship between housing sales price index research, it shows that the housing sales price index and the first five months' search for housing sales index has the greatest relevance[3]. Askitas, Zimmermann and others use Google's Web search data to explore the correlation between the network search data and social unemployment and so on[4-5].

In the current study of tourist flow forecast, relevant literatures in china and abroad have been combined with the traditional time series, linear regression, exponential smoothing, artificial neural network and other models to predict the tourism flow forecast. It can get better macro-prediction results, but these models also have their own flaws that can not be overcome. Therefore, this paper will focus on the use of network search data as a sample data, and establish a new forecasting model based on support vector machine (SVM) to improve the accuracy and timeliness of the forecast. Finally, it uses the official statistical data to verify the predictive effects of the Model.

3 Empirical Analysis

3.1 Keyword selection

Since more and more tourists obtain scenic spot information through Internet, the selection of the web search keywords is directly related to the analysis of the core research on the economic behavior development tendency of the tourists. In this paper, , the basic keywords regarding the travel destination, such as name of the city or scenic spot,

travel route, weather, food and hotel, are selected according to the analysis of the travel information contents concerned by tourists upon the travel destination (scenic spot) before travel. Firstly, "Qionghai Tourism", "Qionghai Hotel Rates", "Qionghai Map", "Qionghai Hotel", "Qionghai Weather", "Qionghai Food", "Boao", "Tan Men", "Long Shou Yang" and "Bai Shi Ling" closely related to the local are selected as the basic keywords; then, relevant tools are adopted to further screen these basic keywords; finally, the six terms --- "Qionghai Tourism", "Qionghai Map", "Qionghai Weather", "Qionghai Food", "Boao" and "Tan Men" which are highly repeated and massively searched are determined as the keywords.

3.2 Data verification

In order to ensure the close correlation between the keywords collected thereby and the number of tourists as well as the stability and good prediction of the keyword data, the keywords selected in this paper are verified from the aspects of correlativity, stability and causality[6].

In this paper, Eviews tool is adopted to calculate the correlation coefficient between the keywords and the number of tourists in the scenic spot. Specifically, the keyword correlation is as shown in Tab. I.

Tab.I Correlation Coefficient between Keywords and Tourists Flow

Keyword	Correlation Coefficient	Keyword	Correlation Coefficient
Qionghai Weather	0.891	Qionghai Food	0.725
Qionghai Tourism	0.838	Tanmen	0.662
Qionghai Map	0.736	Boao Forum for Asia	0.829

3.3 Stability verification

In this paper, ADF test method is adopted for the stability verification of the six keyword variables ---- "Qionghai Tourism", "Qionghai Map", "Qionghai Weather", "Qionghai Food", "Boao" and "Tan Men". If various variable sequences are within the integrated first-order range, then it is indicated that the six keywords conform to the precondition for co-integration analysis. Specifically, the stability of the six keywords is as shown in Tab.II.

Tab.II Stability Test of Keyword Sequence

Keyword	ADF Test	1% Critical Value	5% Critical Value	10% Critical Value	ADF Conclusion
Qionghai Weather	-9.83501	-4.77351	-3.16382	-2.91637	Second-order Stability
Qionghai Tourism	-8.93631	-4.17349	-3.75316	-2.75817	Second-order Stability
Qionghai Map	-9.16748	-3.95637	-3.75521	-2.65133	Second-order Stability
Qionghai Food	-9.56238	-3.46692	-3.65924	-2.76946	Second-order Stability
Tan Men	-8.68892	-4.26569	-3.85931	-2.12644	Second-order Stability
Boao Forum for Asia	-7.67846	-3.59677	-2.65855	-2.34767	Second-order Stability

3.4 Granger causality test

In order to select the independent variables with good prediction capability, it is necessary to carry out Granger causality test for such independent variable as "Qionghai Weather", "Qionghai Tourism" and "Qionghai Map". In this paper, Granger causality test is carried out between the number of tourists in the scenic spot and the six keywords ---- "Qionghai Weather", "Qionghai Tourism", "Qionghai Map", "Qionghai Food", "Tan Men" and "Boao Forum for Asia".

According to Tab.II, the keywords ---- "Qionghai Weather" and "Qionghai Food" have unidirectional causal relationship with the actual tourist flow of the scenic spots in Qionghai. The probability for "Qionghai Weather" and "Qionghai Food" to become the Granger causality of the number of tourists in Qionghai is 94.81%. This indicates: when the three keyword search indexes are changed, the number of tourists in Qionghai will be also changed. Additionally, the keywords ---- "Tan Men" and "Boao Forum for Asia" have bi-directional causal relationship with the number of tourists in Qionghai, the probability of the Granger causality between the two is approximate to 97%. This indicates: when the two keyword search indexes are changed, the number of tourists in Qionghai will be also changed; but when the statistical number of tourists in Qionghai is changed, the two keyword search indexes will be also changed.

4 v-SVR Prediction Model Structuring

4.1 Basic principle and method

The combination of support vector machine (SVM) and regression algorithm is called as support vector regression (SVR) [7-13]. SVR is a support vector machine used for regression analysis, and the algorithm implementation process is as follows: for a given sample set $\{x_i, y_i\}_{i=1}^{k}$, wherein xi is

the input value and y_i is the corresponding predicted value, the original problem is mapped to the linear problem in a certain high-order characteristic space through nonlinear mapping. Specifically, the linear regression function in the multi-dimensional space is as follows:

$$f(x)=\omega\bullet\Phi(x)+b \quad (1)$$

Where w is the weight vector and b is the deviation. The linear regression in the high-dimensional characteristic space is corresponding to the nonlinear regression in the low-dimensional characteristic space to eliminate the dot product calculation of w and φ in the high-dimensional characteristic space. The regression problem is functionally minimized as $\sum_{i=1}^{k}C(e_k)+0.5\|\omega\|^2$, wherein $e_k=f(x_k)-y_k$ is true and $c(\bullet)$ is the loss function. In order to ensure that the loss function has good sparse characteristic, the loss function is selected as $\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{k}(\xi_i^*+\xi_i)$.

During SVM structuring, in allusion to the nonlinear problem, the data points will be corresponding to the N-dimensional characteristic space through a nonlinear function in order to calculate the optimal hyperplane in the new space. Specifically, $\Phi(x)$ is assumed as the correspondence function, and x is placed in the N-dimensional characteristic space for relevant calculation. Formula $y_i(w\cdot x_i+b)\geq 1$ for calculating the optimal hyperplane is converted as follows:

$$w(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_j\Phi(x_i)\Phi(x_j) \quad (2)$$

If $K(x_i,x_j)=\Phi(x_i)\bullet\Phi(x_j)$ is true, the following formula can be obtained through calculation:

$$w(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_jK(x_i,x_j) \quad (3)$$

Where $K(x_i,x_j)$ is the kernel function; when the radial-basis kernel function is adopted as the kernel function, $K(x_i,x_j)=\exp\left[-|x-x_i|^2/(2\sigma^2)\right]$ is true, wherein x is the sample data and $\sigma>0$ is true (σ is the parameter of the radial-basis kernel function). According to Formula (3), SVR regression function is as follows:

$$f(x)=\omega\bullet\Phi(x)+b^*=\sum_{i=1}^{n}(a_i-a_i^*)k(x_i,x)+b^* \quad (4)$$

4.2 Tourist flow prediction model

v-SVR regression model which can automatically calculate ε is adopted in this paper. In v-SVR, parameters C, v and parameter σ of radial-basis kernel function can directly determine the performance of v-SVR. In order to further improve the accuracy of the prediction model, the model should be structured as follows: firstly, the original sample data should be normalized in order to normalize the input and the output into [0.1]; then, the radial-basis kernel function should be selected as the kernel function; finally, penalty factor C and v should be selected to solve quadratic programming problem $s.t\begin{cases}\sum_{i=1}^{k}(a_i-a_i^*)=0\\ \sum_{i=1}^{k}(a_i+a_i^*)\leq C\bullet v\\ 0\leq a_i,a_i^*\leq C/k\end{cases}$. Accordingly, mean square error between the actual tourist flow and the predicted value can be calculated as $\frac{1}{n}\sum_{i=1}^{n}\left[f(x)-y_i\right]^2$, and the number of vectors can be controlled according to C, σ and $v\in[0,1]$ determined thereby, and then parameter v can be calculated through the analysis of the prediction error and the number of support vector machines.

4.3 Result analysis

In this paper, the tourist flow monitoring data of a certain scenic spot of the city for the three festivals ---- the Dragon Boat Festival, the Mid-autumn Festival and the National Day are taken as the training sample set, and the tourist data for the New Year's Day are taken as the test sample set. LIBSVM software is adopted to construct v-SRV model.

The tourist flow monitoring data of a certain scenic spot of the city are adopted in this paper for simulation. Firstly, v is set as v=1, and the nonlinear programming optimization algorithm is adopted for optimal parameter training to obtain the optimal $[C,\sigma]=[86.4374,0.02471]$; then, v is calculated as v=0.05 to obtain good prediction accuracy. In order to verify the prediction effect of the algorithm, Matlab simulation experiment is carried out in this paper to group the tourist flow data for recent two years according to festivals, wherein the records for the Dragon Boat Festival, the National Day and the New Year's Day are adopted. Firstly, the tourist records of the scenic spot for 2013 and 2014 are adopted for training, as shown in Tab.III. Meanwhile, the tourist records for 2015 are adopted to verify the prediction effect.

Tab. III Tourist Flow Statistics for Two Years

Date	The Dragon BoatFestival	The Mid-autumn Festival	National Day	The New Year's Day
Tourist Flow (2013)	322	660	775	707
Tourist Flow (2014)	338	631	792	612

The analysis of the predicted value and the measured value of the scenic spot for 2015 is as shown in the following table IV:

Tab. IV The Results of Tourism Flow Prediction

Date (2015)	Actual Flow	Predicted Flow	Prediction Error%
The Dragon Boat Festival	401	317	-20.9
The Mid-autumn Festival	760	742	-2.4
The National Day	807	868	-7.5
The New Year's Day	764	701	-9.3

According to Tab.III and Tab.IV, if there are less sample data, the model has large prediction error; if there are more sample data, the model has high prediction accuracy but large calculation workload. Generally speaking, the error mean of the model for the tourist flow of the scenic spot is about 10%, and the maximum error is not more than 21% and the minimum error is 2.4%, so the model has relatively satisfactory prediction accuracy and presents good reference value for the tourist flow prediction.

5 Conclusion

In the context of large data use, this paper takes the local city's scenic passenger flow as an example, uses the stability test and causality test of the statistics and tries to testify the relationship between Baidu search keywords such as "Qionghai tourism" , "Qionghai map", "Qionghai weather", "Qionghai food", "BoAo", "Tan Meng" and the numbers of tourists in tourism attractions during the period from January 2012 to October 2015. The scenic passenger flow forecast model based on SRV is built to predict the passenger flow during the festivals. The forecasting model is trained by using the statistic data of tourist numbers from January, 2012 to October, 2015 as the training sample. Finally, it realizes to predict the passenger flow forecast of the scenic spot in 2015 by the use of Internet search data in 2015, and analyzes the reliability of the model prediction accuracy by comparing with the official statistics. The results show that the forecast model based on SRV has better prediction precision; it can provide the scenic area managers with time of resource allocation plan and the measures for the guarantee of passengers' travelling quality, meet the timeliness required for prediction.

References

[1] Big data Specialized Committee. 2013 data technology white paper [A].2013.12.
[2] Liu Ying, Lv Benfu, and, the ability of the network search data to forecast the stock market; theoretical analysis and empirical test [J], economic management 2011.
[3] Wang Lian, Jia Jianmin. Dynamic characteristics of risk perception of sudden disaster events: evidence from web search [J]. management review, 2014, 26 (5)
[4] Choi H，Varian H．Predicting the Present with Google Trends[J]． The Economic Record (Special Issue)，2012，88(6).
[5] Lin Zhihui, Ma Yaofeng, et al., et al. Analysis of the temporal and spatial distribution of tourist attractions [J]. resources science, 2012.34 (12): 2427
[6] Li Juntie, Yang Min. The Xi'an National Tourist network information search behavior of [J] economic geography, 2010, 30 (7):1212-1216.
[7]Pelckmans K., SuykensJ. A. K., De Moor B. Regularization constants in LS-SVMs: a fast estimate via convex optimization, In Proceedings of 2004 IEEE International Joint Conference on Neural Networks, July 2004, 1:25-29
[8] Koley C., urkait P., Chakravorti S. avelet-Aided SVM Tool for Impulse Fault Identification in Transformers, IEEE Transactions on Power Delivery, July 2006, 21(3): 283 – 290.
[9] Song H.,Li G. Tourism demand modelling and forecasting A review of recent research[J]. Tourism Management,2008,29(2): 203-220.
[10]Claveria O.,Torra S. Forecasting tourism demand to Catalonia: Neural networks vs.time series models [J]. Economic Modelling,2014,36: 220-228.
[11] Long Maoxing,Sun Gennian,Ma Lijun, et al.An analysis on the vanriation between the degree of consumer attention of travel network and tourist flow in regional tourism: A case of Sichuan Province[J].aeal research and development,2011,30(3):93-97.
[12]Sun yi, Lv Benfu. A review of reseaches on the correlation between internet search and economic behaviour[J].Management Review,2011,23:72-77.
[13] Ginsberg J, Mohebbi M H, Patel R S,etc al. Detecing influenza epidemics using search engine query data[J].natutre,2009,(2):1012-1014.