A Random Forest based Learning Framework for Tourism Demand Forecasting with Search Queries

Xin Li Dr.
Beijing Union University, Beijing, China

A Random Forest based Learning Framework for Tourism Demand Forecasting with Search Queries

Tony Chen, Ph.D. Candidate
School of Economics and Management
Beihang University, Beijing 100191, China
Telephone: 86-156-003-36797
E-mail: tonychen1989@gmail.com

and

Xin Li (corresponding author), Ph.D.
Institute of Tourism
Beijing Union University, Beijing 100101, China
Telephone: 86-159-011-60410
E-mail: leexin111@163.com

Abstract

This study proposes a novel framework for tourism demand forecasting, which combines search queries generated on the Internet, advanced feature selection methods, and machine learning based forecasting technique. This new methodology is applied to forecast tourism demand in two popular destinations in China: Beijing and Haikou. The study evaluates the performances of various feature selection approaches in tourism forecasting. We further show that the random forest feature selection and support vector regression with radial basis function can be extremely useful for the accurate forecasts of tourist volumes. This research highlights the advantage of big data and machine learning algorithms in the tourism forecasting.

Keywords: random forest, feature selection, machine learning, tourism demand forecasting

Introduction

Search queries have emerged as popular and significant sources for tourism demand forecasting, and they have been documented to predict tourist volumes pretty accurately (Choi & Varian, 2012; Bangwayo-Skeete & Skeete, 2015; Yang et al., 2015). In general, search queries have high dimensions because they are generated by the public on the Internet (Wu & Brynjolfsson, 2013; Cho & Tomkins, 2007). To improve the forecasting accuracy, researchers need to select appropriate search queries, and model them with optimal approaches. Machine learning based methods can model large data sets with relatively superiority compared with econometric models (Liao, Chu, & Hsiao, 2012; Pai, Huang, & Lin, 2014). To the best of our knowledge, single models are unlikely to perform well, especially with large data sets. Therefore, a key challenge to bridge the gap is to propose an integrated methodology that makes the best of machine learning based approaches.

The goal of this study was to provide an integrated framework for tourism demand forecasting, which combines search queries, feature selection process, and machine learning approaches.

Search queries are considered as features, and feature selection methods can eliminate irrelevant features and keep the ones that are helpful for the forecasting. Then, machine learning methods are applied to modeling these selected search queries. Compared to econometric models, machine learning based models can process and modeling large date sets with higher accuracy. Moreover, empirical findings suggest that random forest based feature selection performs best among other filter and wrapper based methods. Our study extends the existing studies about tourism demand forecasting with search queries by using machine learning framework.

Literature Review

Tourism demand forecasting with search queries is first reviewed. Then, the commonly used feature selection methods including both filter-based and wrapper-based ones are illustrated. Afterwards, machine learning based approaches and their advantages are briefly reviewed. Finally, the research gaps are addressed at the end of the section.

Search queries reflect how people show interests and attention on specific topics on the Internet. Existing studies have demonstrated that these data can predict the future trends (Choi and Varian 2012). Researchers have incorporated these unique data sources into the forecasting in the fields of tourism and hospitality (Pan, Wu, & Song, 2012; Yang, et al., 2015). However, the approaches used to model the search queries are relatively simple, such as autoregressive model, vector autoregressive model, and autoregressive distributed lag models (Ettredge, Gerdes, & Karuga, 2005; Song & Witt, 2006; Chu, 2009; McLaren, 2011; Vosen & Schmidt, 2012; Akın, 2015; Hassani, Webster & Heravi, 2015). These econometric models are superior to depict the linear correlation between the dependent and independent variables, especially when the dimensions of search queries are low. However, these models may fail to process large, nonlinear data sets with exceptional performances. Compared to econometric models, machine learning based approaches have advantages of modeling large data sets. In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression (Cortes & Vapnik, 1995; Chen, Chiang & Storey, 2012).

From the perspective of machine learning, each search query can be considered as a feature, so the major problem is to select most appropriate features. In general, there are mainly three types of feature selection methods: filter-based, wrapper-based ones and embedded based ones (Kohavi et al., 1997; Guyon & Elisseff, 2003; Tong & Mintram, 2010). In particular, Guyon and Elisseff (2003) suggested that feature selection can improve the prediction performances, provide faster and more cost-effective predictors, and better understand the data generation process. A filter-based and three wrapper-base feature selection methods are illustrated as well. In particular, random forest combines several binary decision tress that are from learning samples and randomly selected variables (Genuer, Poggi, & Tuleau-Malot, 2010).

To date, scarce literature adopted machine learning based approaches to forecast tourism demand with search queries. It is necessary to propose a new methodology to improve the forecasting accuracy of tourism demand. In this study, different feature selection and support vector machines are combined to model search queries.

Methodology

An integrated machine learning based forecasting framework is proposed, which combines random forest feature selection and support vector machine. To evaluate the performances of different feature selection methods, we use the following methods, including filter based feature selection (FBFS), backwards feature selection (BFS), genetic algorithm feature selection (GAFS), and random forest feature selection (RFFS). Then, a support vector machine with radial basis function is used to train and test the forecasting model.

This framework has four integrated components: search queries collection, feature selection, machine learning based modeling, and forecasting evaluation. First, weekly search queries are collected from Baidu trends following specific search terms. Then, the feature selection is used to eliminate the most irrelevant features. Afterwards, the selected series are modeled with support vector machines. Finally, the performances of the models are evaluated.

Results

We conduct rigorous experiments to verify the effectiveness of the proposed framework in the forecasting of Beijing and Haikou tourism volumes. Different feature selection methods generate various subsets of features, which are shown in Table 1.

Table 1. Selected Features

Beijing: Feature Selection Methods	Number of Selected Features
FBFS	Top 11 of 46 features
BFS	Top 20 of 46 features
GAFS	Top 13 of 46 features
RFFS	Top 9 of 46 features
Haikou: Feature Selection Methods	**Number of Selected Features**
FBFS	Top 7 of 20 features
BFS	Top 6 of 20 features
GAFS	Top 4 of 20 features
RFFS	Top 7 of 20 features

The selected feature sets using different methods, are shown in the above table. As shown in Table 1, 9 out of 46 features of Beijing tourism and 7 out of 20 features of Haikou tourism are selected using random forest feature selections.

Then, for the robustness of the empirical study, the samples are randomly split into three-folds, two samples are used to train and the other one is used to test the model. The average forecasting accuracy with support vector machines are also computed. Table 2 presents the average forecasting accuracy of Beijing and Haikou tourist volumes.

Table 2. Average Forecasting Accuracy of Beijing and Haikou Tourist Volumes

City: Beijing		Average forecasting accuracy
Machine learning	Feature selection	Mean Absolute Percentage Error (MAPE)
SVM with Radial Basis Function	FBFS	0.18
	BFS	0.22
	GAFS	0.25
	RFFS	0.17
	All features	0.29
City: Haikou		Average forecasting accuracy
Machine learning	Feature selection	Mean Absolute Percentage Error (MAPE)
SVM with Radial Basis Function	FBFS	0.09
	BFS	0.08
	GAFS	0.09
	RFFS	0.08
	All features	0.12

Interesting findings are shown from Table 2. First, the models with feature selection are superior those without feature selection process. The first four SVMs are constructed with the selected features sets, which are obtained with a filter based and three wrapper based feature selection methods. The last SVM model handles all features without the selection process. From the whole forecasting accuracy of Beijing tourism demand, SVM with the selected features perform better than the model without any feature selection process.

Second, random forest based feature selection has the lowest MAPEs, compared to other feature selection methods. Filter based feature selection method also has exceptional accuracy in this forecasting task. Random forest feature selection improves the forecasting accuracy by 5.56%, 22.73%, 32%, and 41.38% compared to filter, backwards, genetic algorithm based feature selection and all features.

In addition, from the empirical study of Haikou forecasts, random forest feature selection and support vector machine perform best, with the lowest MAPE. The worst performance is the combination of all features and support vector machine. In particular, features using random forest improve 33.33% forecasting accuracy of tourism demand compared to all features.

In average, wrapper based feature selection methods outperform filter based ones in the forecasting of tourist volumes, with the reduction of forecasting error being 18.52%. Feature selection can eliminate unimportant search queries and significantly improve the forecasting accuracy of tourism demand. Therefore, it is convinced that the forecasting framework that combines search queries, feature selection, and machine learning approaches can significantly improve the tourism demand forecasting accuracy.

Conclusion and Discussion

Search queries generated on the Internet reflect tourists' attention on the destinations. By incorporating search queries into the forecasting of tourism demand, researchers can update their

forecasts timely and accurately. We propose an integrated methodology that effectively analyze and modeling large search queries. This study considers search queries as features, and aims to apply machine learning based approaches into the forecasting of tourism demand. We hope to provide new insights for the timely and accurate forecasts of tourism demand in the Big Data era.

Empirical results show that random forest based feature selection and support vector machines have the superiority to improve the forecasting accuracy of tourism demand. The proposed methodology is useful for analyzing large search query sets, because it can select most important variables for the forecasting. In the future research, we need to examine whether this methodology can accurately forecast tourist volumes to other destinations. In particular, the future research should emphasize on the adoption of machine learning based approaches so as to analyze more user-generated, multi-types, and big data.

Acknowledgements

The authors would like to thank the Chair, co-chairs and reviewers of ttra conference for these valuable help, insightful suggestions and comments, which has dramatically improved the quality of this research. This research is supported by grants from the National Natural Science Foundation of China (NSFC No. 71373023), National Key Technology Support Program (No. 2012BAZ03744), Beijing Higher Education Young Elite Teacher Project (No. YETP1750) and support from Talents Program in Beijing Union University (RK100201509).

References

Akın, M. (2015). "A novel approach to model selection in tourism demand modeling." Tourism management 48: 64-72.

Bangwayo-Skeete, P. F. and R. W. Skeete (2015). "Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach." Tourism management 46: 454-464.

Chen, H., R. H. L. Chiang., and V. C. Storey (2012). "Business intelligence and analytics: From big data to big impact." MIS Quarterly 36(4).

Cho, J. and A. Tomkins (2007). "Social media and search." IEEE Internet Computing 11(6): 13-15.

Choi, H. and H. Varian (2012). "Predicting the present with google trends." Economic Record 88(s1): 2-9.

Chu, F. L. (2009). "Forecasting tourism demand with ARMA-based methods." Tourism management 30(5): 740-751.

Cortes, C. and V. Vapnik (1995). "Support-vector networks". Machine Learning 20(3): 273. doi:10.1007/BF00994018.

Ettredge, M., J. Gerdes, and G. Karuga (2005). "Using web-based search data to predict macroeconomic statistics." Communications of the ACM 48(11): 87-92.

Hassani, H., A. Webster, E. S. Silva, and S. Heravi (2015). "Forecasting U.S. tourist arrivals using optimal Singular Spectrum Analysis." Tourism management 46: 322-335.

Genuer, R., J. M. Poggi, and C. Tuleau-Malot (2010). “Variable selection using Random Forests”. Pattern Recognition Letters, 31 (14): 2225-2236.

Kohavi, R., and G.H. John (1997). “Wrappers for feature subset selection”. Artificial Intelligence. 97(1-2): 273-324.

Guyon, I., and A. Elisseff (2003). “An introduction to variable and feature selection”. Journal of Machine Learning Research 3: 1157-1182.

Liao, S. H., Chu, P, H., and Hsiao, P. Y (2012). "Data mining techniques and applications – A decade review from 2000 to 2011." Expert Systems with Applications 39(12): 11303-11311.

McLaren, N. (2011). "Using Internet search data as economic indicators." Bank of England Quarterly Bulletin 51(2).

Pai, P. F., K. C. Huang, and K. P. Lin (2014). "Tourism demand forecasting using novel hybrid system." Expert Systems with Applications 41(8): 3691-3702.

Pan, B., C. Wu., & H. Song (2012). "Forecasting hotel room demand using search engine data." Journal of Hospitality and Tourism Technology 3(3): 196-210.

Song, H. and S. F. Witt (2006). "Forecasting international tourist flows to Macau." Tourism management 27(2): 214-224.

Tong, D. L. and R. Mintram (2010). "Genetic algorithm-neural network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection." International Journal of Machine Learning and Cybernetics 1(1-4): 75-87.

Vosen, S. and T. Schmidt (2012). "A monthly consumption indicator for Germany based on Internet search query data." Applied Economics Letters 19(7): 683-687.

Wu, L. and E. Brynjolfsson (2013). “The future of prediction: How Google searches foreshadow housing prices and sales”. Economics of Digitization, University of Chicago Press.