3M4-2

Predicting tourism trends through the data of online communications

Huang Dali[*1] Matsuo Yutaka[*2]

[*1, 2] Department of Technology Management for Innovation

Graduate School of Engineering, University of Tokyo

This paper is focusing on telling the effectiveness of the data from web communities to forecast travel trends in real world. The experiments are proposed to discover the difference of various feature-sets in the progress of forecasting Chinese people going traveling in Japan.

1. Introduction

The status of the tourism market is one of the most important indexes of development of one country. As a result, tourism forecasting is a fervid topic in the economics field. In the past decades, many researches have been done with multiple economic models, including time-series models, the econometric approach as well as some emerging new statistical and non-statistical methods, along with various data resources. In the field of tourism demand forecasting, the researches cover the categories of the latest methodological developments [Xu 2012], forecast competition, combination and integration [Song 2012], tourism cycles [Gouveia 2005], turning points, directional changes and seasonality analysis, events' impact analysis [Eugenio-Martin 2005] and risk forecasting [Prideaux 2003] in addition to some general observations [Song 2007].

With the widely use of machine learning tools, some recent studies tried to apply machine learning algorism to this task and got progresses. Nowadays, the Artificial Neural Network (ANN) is often applied in the relevant studies and the Support Vector Regression (SVR) was first introduced to this area in 2006 [Pai 2006]. But SVR was not largely used in these kinds of researches, as traditionally the data used is obtained annually which normally contains around ten to dozens of years. Some has use the daily booking information from hotels or ticket offices. In these cases, the studies are focusing on a single problem such as a sightseeing place or a city [Xu 2012]. More studies were about the transportation services demand prediction [Totamane 2012].

In this paper, we are taking Japan, the country with golden value in the travel market, as a target place and try to use the support vector regression to analyze the travel population from China. As none of current study use the data from online communities (such as forums, online travel guides, etc.), as well as none of them focusing the context in Chinese. The data from a Chinese travel forum is collected and used as features in this study. Also other data related to tourism market is also collected and used in the experiments. In order to tell the effectiveness of using the analyzing results of the data from online communities to forecast travel trends of people in real world, we propose an experiment to discover the effectiveness of different web data to predict Chinese people go traveling in Japan.

In the next section, we describe the data sets used in the methods, and the theory of the methods will be explained in the third section. The fourth section will cover the experiment and followed by the last summary of the study.

Contact: Huang Dali, Weblab, Department of Technology Management for Innovation, Graduate School of Engineering, The University Of Tokyo. 113-0033, Room 92C1, Faculty of Engineering Bldg.2, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Tel: +81-90-9970-8912, Facsimile number: 03-5841-7718, Email: dali.huang@gmail.com

2. Data Sets

2.1 Data set from the web community

Compare with most of the travel population researches which often use seasonally data and annual data, we use monthly data to apply with our methods. As around 70% of the Chinese travelers stay in Japan less than a week [JNTO 2013], it is for sure that one month could normally cover one's integral trip. In this paper, the real communication data are collected from one of the most popular traveling online forums in China named "QYER" (place.qyer.com). This website is chosen for several reasons. One is that the data from this site dates back to 2005, which makes it the oldest forum among the all. Moreover, it also provides the editing and replying information for us to crawl. In total, 54,870 posts from this forum have been collected, including 6,144,221 Chinese characters from November 2005 to August 2014. Among them, 74 months of data, from July 2008 to August 2014, are consecutive in time for the next process.

To determine the popular characters related to the topics about traveling in Japan from the raw data, the concept of Tf-Idf is used here. "Tf" means term frequency, which shows the number of times a word appears in the document, while "Idf" as the inverse document frequency, gives the inverse frequency of the word in the corpus. It is a numerical statistic method, which could reflect how important a word is to a document in a collection or corpus. In this study, we use Sogou corpus [R&D Center of SOHU 2008], a corpus with 17,991 news articles in Chinese. Using Tf-Idf as the weighting factor here we could get a rank of the popular Chinese words of this forum, so that the top-eight popular words could be discovered. The Tf-Idf result of the top eight words shows in Table 1. Then the numbers of times these characters used by time are used as one set of features for the support vector machine. Moreover, the numbers of times

Table 1. Tf-Idf of the popular words

Words*	Kyoto	Osaka	Tokyo	Japan	Like/Love	Nara	Hakone	Hotel
Tf-Idf	0.027932036	0.024451213	0.01982668	0.018268852	0.01631569	0.015489705	0.013878809	0.010307077

*The original words are in Chinese, here shows the meaning of them in English

users replying and editing related topics online are used as another set of features here for a same progress.

2.2 Data sets related to the tourism market

In study of tourism market, some indexes could have influence on the change of travel population. In the situation of transnational tourism, the currency exchange rate plays a great role in affecting the travel trends. Here we collect the monthly exchange rate between Japanese Yen (JPY) and Chinese Yuan (CNY) from the year of 2008 to the year of 2014 [BOJ 2014]. Since we took travel as a consumer behavior, the consumer price index (CPI) of China also has impact on the tourism activities of Chinese people. As a result, the monthly CPI data is gathered from the National Bureau of Statistics of the People's Republic of China [CPI 2014]. As well as consumer price index, the Gross Domestic Product (GDP) of a city or a country is also a good index for traditional tourism demand forecasting task. However the GDP is seasonally counted by the government, which is not suitable for our experiment. Finally, to the local situation of China, the national holidays, such as "the golden week" or "the spring festival", are taken as one set of features for the experiment. The information comes from the general office of the State Council of China, which publishes the holiday schedule of each year [GO-SCC 2014].

Moreover, the historical data of population traveled in Japan each month since 2008 is collected from the Japan National Tourism Organization [JNTO 2014].

3. Proposed Methods

In this study, the Support Vector Regression (SVR) is used to the estimation. For the RBF kernel of this method, grid search is used to find the proper parameters.

3.1 Support Vector Regression

The support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns [Cortes 1995]. The SVM maps data nonlinearly into a higher-dimensional feature space. Support vector machines were designed for classification problems at first and then developed to regression tasks. In the support vector classification procedure, the SVM looks for one hyper plane, one that could separate the elements of the two named classes with the largest margin, by solving a problem of constrained optimization, where a different constraint is introduced for each of the labeled training set points. To solve this optimization problem, the Lagrange multipliers are introduced to change the problem into a linear combination of some of the training examples, which are called support vector machines.

Support Vector Regression is a regression technique based on support sector machines. Apart from the usual regression models, a loss function is employed by the SVR. Through the Ordinary Least Squares solution, a hyper plane could be found that no other plane with a smaller sum of square errors exists. Moreover, this loss function penalizes only prediction errors that are greater than the distance from the observed data and ignores errors that are less. Comparing to a classification problem, the regression one could be considered as an application of the support vector approach to function estimation.

3.2 Parameter settings for the kernel function

With the adoption of kernel functions, which correspond to a dot product in the feature space, the number of features can be much larger than the number of dimensions of the input space. As the number of features increases, the problem is more likely to be linearly separated. Therefore, thanks to the kernel functions, linear classifiers can be applied to non-linear problems as well.

There are two possible solutions to determine the kernel functions. One is to define a special designed function, which is suitable for a specific data set/problem. The other is to use kernels of more general applications, as the Radial Basis Function (RBF) is one of the most widely adopted kernels.

Since in some cases it is not possible to find a hyper plane perfectly meet the constraint, parameter c is need to be set in SVM tools [Vapnik 1996]. In the actual application of SVR process, the values of parameters c, g and p need to be chosen carefully as they have a massive impact on the regression result of the experiment. In the Libsvm [Chang 2011], a widely adopted SVM tool, the parameters could be chosen by the grid search using the tool gridregression offered by the same author of Libsvm.

4. Experiment and Result

4.1 The modeling description

In order to discover the effectiveness of different data to predict Chinese people go traveling in Japan, the models are used to apply with SVR. Since we are trying to predict the future population in this activity. The data from the present need to be able to show some connections with the result in the future. Therefore, a time gap between the label and the features is needed. The historical travel population from China to Japan is used as the label. And several sets of features are tested in the program. The descriptions of features tested are in the following terms.

- T_n: time gap between labels and features in each row, here we tested with the gap from one month to four months. Here "n" values between 1 to 4
- H: National Holidays
- E: the currency exchange rate between JPY and CNY
- C: the consumer index of China
- 6: the six word counts of the six most popular words used in the forum

- 8: the eight word counts of the eight most popular words used in the forum
- R: the times of users from the forum replying and editing the topics in each month

In applying SVR to our experiments, the following aspects need to be considered.

The large differences in the ranges of values of the features, as the Holidays of each month goes from 0 to at most 6 days while the times users replying and editing on the web could be thousands or more. This will lead to an unwanted effect of giving heavier weight to some characteristics than to the others. To address this issue, a data-preprocessing called normalization step should be applied. Here we normalize the value of the features to the range of -1 to +1. This process is done both with the data for model training and the data for testing.

Also, since the sets of data we prepared for this study are non-linear, the kernel functions should be considered to map a problem into a feature space where the target function consists of a line. Given this condition, the grid search is done for the SVR using RBF kernel. As a result, we could determine the best parameters, c, g and p, for the next model-training step. Of all, the 15 months from June 2013 to August 2014 are used as testing data.

4.2 Regression Result

We select the "HEC", the national holidays, the currency exchange rate between JPY and CNY and the consumer index of China out of the six kinds of features to be the benchmark of our experiment. With the adding of different features as well as the time gap change between the prediction target – the label and the features, we could get a result of the different values in showing different degree of the relevance of these features.

The result showed in Table 2 is the squared correlation coefficient values of each model. Here the squared correlation coefficient means the square of the correlation coefficient between the original data and modeled data values, which could be use to measure the relevance of the features.

Table 2. The squared correlation coefficient result of the SVR

Models	Squared correlation coefficient	Models	Squared correlation coefficient
T_1HEC	0.0542925	T_3HEC	**0.360657**
T_1HECR	0.0597827	T_3HECR	0.104273
T_1HEC6	**0.185166**	T_3HEC6	0.112455
T_1HEC8	0.109621	T_3HEC8	0.0468776
T_1HECR6	0.177765	T_3HECR6	0.222104
T_1HECR8	0.109382	T_3HECR8	0.21184
T_2HEC	0.0232152	T_4HEC	0.0924028
T_2HECR	**0.312547**	T_4HECR	0.0183738
T_2HEC6	0.15239	T_4HEC6	**0.22203**
T_2HEC8	0.117742	T_4HEC8	0.16653
T_2HECR6	0.147087	T_4HECR6	0.0764497
T_2HECR8	0.117462	T_4HECR8	0.0613273

5. Discussion and Conclusion

Tourism market is influenced by lots of factors at the real world. That makes it hard to predict by using web forum data only. With more and more people participate in the trip to Japan, our study could tell some interesting phenomenon from the result. As for a relevantly high consumer activity, the trip to Japan for Chinese people is rather a serous consideration to a random choice. With the result from the experiment, in table 2, some interesting discoveries could be found. For the same sets of features used in the prediction tourism demand, the time gap change leads to the fluctuation of the squared correlation coefficient values. In these six kinds of feature-sets, half of them have a better performance with the time gap of three month. And five of them show as the time gap goes over two months, the outcome does better than the others. This suggests that normally people finally travel to Japan intend to discuss and study about it over a month before they make the decision.

If we take the discussion to each team with the same time gap, the features with six most popular words plays a better part in the procedure, though the out come of them all did not meet the satisfaction of prediction. To the 2-month team, it is really evident that the value of squared correlation coefficient changes a lot from "T_2HEC" to "T_2HECR". This could be considered as the phenomenon of "people prefer to join the communications on the travel forum about two month before the trip to Japan". Compared with the total opposite results between "T_3HEC" and "T_3HECR", this assumption become more evident. So it might be better if the tourism promotion of Japan's sightseeing places starts online at about two month before the best season comes, but not three months before it. For the time gap of four months, same pattern of outcomes appears as the time gap of one month. This could be seen as a character of this travel forum, in which people shares their trip and discussing about their experiences.

However, applying SVR requires setting the features exquisitely since they heavily affect the prediction accuracy. There are no general guidelines available for us to select these features for this problem. In this paper, by using the information from web services, we introduce new features to the prediction of tourism demand and compare the different sets of features. So we could determine the best sets in different conditions. But for the use in the real world, there still much improvement needs to be done. Thus, in the future work, to study the consumer behavior of people through social network information, we could try with a range of potential features. Such as the weather condition or the age range of the tourists. Also the travel market in Japan is also highly influenced by natural disasters such as earthquakes. Moreover, for a study like tourism prediction, it is crucial to get more information online to pursuit the better performance.

References

[Totamane 2012] Totamane, R. ; Dasgupta, A. ; Rao, S. Air Cargo Demand Modeling and Prediction, Systems Journal, IEEE (Volume:8, Issue: 1) 2012

[Song 2007] Haiyan Song, Gang Li, Tourism demand modelling and forecasting—A review of recent research, Tourism Management 29 (2008) 203–220, Elsevier, 2007

[Xu 2012] Xu Yijun, Yu Zhang, Analysis and prediction of the total number of ice-snow tourism in Heilongjiang based on times series: A case study of Harbin, Robotics and Applications (ISRA), 2012 IEEE Symposium on, 2012

[Pai 2006] Pai, P. F., Hong, W. C., Chang, P. T., & Chen, C. T. The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study. International Journal of Management, 23, 375–385. 2006

[R&D Center of SOHU 2008] R&D Center of SOHU, Sogou corpus, http://www.sogou.com/labs/dl/c.html, 2008

[JNTO 2013] 観光庁『訪日外国人消費動向調査（平成 25 年版）』, http://www.jnto.go.jp/jpn/reference/tourism_data/pdf/market_basic_china.pdf , 2013

[CPI 2014] Monthly consumer price index data of China, the National Bureau of Statistics of the People's Republic of China, http://data.stats.gov.cn/search/keywordlist2?keyword=cpi, 2014

[BOJ 2014] Monthly currency exchange rate between Japanese Yen and Chinese Yuan, Bank of Japan, http://www.stat-search.boj.or.jp/ssi/mtshtml/m_en.html, 2014

[GO-SCC 2014] Notices of the holiday arrangements, the general office of the State Council of China, http://www.gov.cn/zhengce/xxgkzl.htm, 2008-2014

[JNTO 2014] Monthly travel population of Chinese (from China to Japan), Japan National Tourism Organization, http://www.jnto.go.jp/jpn/news/data_info_listing/index.html, 2014

[Eugenio-Martin 2005] Eugenio-Martin, J., Sinclair, M. T., & Yeoman, I. Quantifying the effects of tourism crises: An application to Scotland. Journal of Travel & Tourism Marketing, 19, 21–34. 2005

[Prideaux 2003] Prideaux, B., Laws, E., & Faulkner, B. Events in Indonesia: Exploring the limits to formal tourism trends forecasting methods in complex crisis situations. Tourism Management, 24, 475–487. 2003

[Gouveia 2005] Gouveia, P. M. D. C. B., & Rodrigues, P. M. M. Dating and synchronizing tourism growth cycles. Tourism Economics, 11, 501–515. 2005

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

[Vapnik 1996] Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. Support Vector Regression Machines, in Advances in Neural Information Processing Systems 9, NIPS 1996, 155–161, MIT Press. 1996

[Chang 2011] Chang, Chih-Chung; Lin, Chih-Jen . "LIBSVM: A library for support vector machines". ACM Transactions on Intelligent Systems and Technology 2 (3). 2011

[Song 2012] Haiyan Song, Bastian Z. Gao, Vera S. Lin, Combining statistical and judgmental forecasts via a web-based tourism demand forecasting system, Original Research Article International Journal of Forecasting, Volume 29, Issue 2, Pages 295-310, 2012