Predicting tourism trends with Google Insights

Evan Gawlik, Hardik Kabaria and Shagandeep Kaur

December 15, 2011

Abstract

Choi and Varian [5] propose the use of publicly available web search volume histories from Google Trends together with time-series regression analysis to predict, among other things, tourism rates in specific countries. We improve upon their approach by using query-specific search data, rather than aggregate search data, to predict visitor arrival statistics. We use feature selection to determine the most relevant search queries and evaluate the performance of the algorithm using k-fold cross-validation. We test our approach on a data set consisting of Hong Kong's monthly visitor arrival statistics for the years 2005-2010. In contrast to Choi and Varian, we demonstrate that our algorithm not only exhibits low training error, but also exhibits low test set error. Our results indicate that web search volume histories provide a useful predictor of tourism rates.

1 Introduction

Several authors have advocated the use of web search volume histories to make predictions [1, 2, 3, 4]. Choi and Varian [5] present one of the first of such studies, using Google Trends (a precursor to Google Insights [6]) data to predict, among other things, tourism rates in specific countries. The authors use time-series regression analysis to predict visitor arrival statistics, incorporating aggregate search volume indices as regression data.

We present an improved approach to the problem of predicting tourism rates, whereby we use query-specific search data and implement a feature selection algorithm to choose the most relevant queries. We evaluate the algorithm using k-fold cross-validation and demonstrate not only low training error as with Choi and Varian's algorithm, but also low test set error.

Predicting tourism trends can potentially help a variety of industries that rely, directly as well as indirectly on tourism, and may enable better facilities management for the correct number of tourists. Based on analysis of the origins of tourists to a particular destination, the study may help strategize a targeted campaign for attracting tourists from areas that do not appear to a good source of tourists. There is an immense potential in the kind of uses this study can be put to, including applications beyond tourism itself.

2 Data

We collected tourism statistics from the Hong Kong Tourism Board's Research Statistics web page [7], which provides data on Hong Kong's monthly visitor rates and the geographic origins of its visitors. Let $y^{(i)}, i = 1, 2, \ldots, m$ denote the logarithm of the total number of visitors to Hong Kong in month i, starting in the month of January 2005 ($i = 1$) and ending in December 2010 ($i = m = 72$).

Since the Hong Kong visitor data provides information on the geographic origins of its visitors, we also tested our algorithm's ability to predict the number of visitors to Hong Kong originating from a subset of the geographic regions reported, namely "The Americas"; "Europe, Africa and the Middle East"; and "Australia, New Zealand and the South Pacific." For lack of a better descriptor, we refer to these regions collectively as Western nations in the subsequent discussion.

We collected search volume histories from Google Insights for a variety of search queries related to Hong Kong tourism. The data provided on Google Insights consists of normalized search volume indices that indicate the number of searches made on a given day by internet users for a particular query on Google.

Table 1 contains a listing of the search queries considered in the present study. Note that the list contains not only English queries, but also Chinese queries. We denote by q_j the j^{th} query under consideration, and we let n_q denote the number of queries considered. (In our case, $n_q = 16$).

We preprocessed the search volume data to obtain monthly search volume index averages $z_j^{(i)}, i = 1, 2, \ldots, m$ for each query q_j. The data are, by Google's convention, scaled to range from 0 to 100, with 100 denoting the maximum search volume observed over the period of inquiry.

3 Methodology

Our aim is to predict, for a given temporal index i, the value of the visitor arrival statistic $y^{(i)}$ using knowledge of past visitor arrival statistics $y^{(1)}, y^{(2)}, \ldots, y^{(i-1)}$ as well as past search volume indices $z_j^{(1)}, z_j^{(2)}, \ldots, z_j^{(i-1)}$, $j = 1, 2, \ldots, n_q$.

To accomplish this task, we model $y^{(i)}$ as a linear combination of a subset of the past visitor arrival statistics and search volume indices. Namely,

$$y_{pred}^{(i)} = \sum_{l \in L} \alpha_l y^{(l)} + \sum_{k \in K} \sum_{j \in J_k} \beta_{jk} z_j^{(k)}, \tag{1}$$

with $L, K \subseteq \{1, 2, \ldots, i-1\}$ and $J_k \subseteq \{1, 2, \ldots, n_q\}$ for each k.

Writing equation (1) for all i in a training set leads to a linear system of the form

$$X\theta = Y, \tag{2}$$

with θ a vector containing the parameters α_k and β_{jk}. Let $\hat{\theta}$ denote the least-squares solution to this system. For test data (X_{test}, Y_{test}), let ε denote the relative error in the predicted value of Y_{test}:

$$\varepsilon = \frac{||X_{test}\hat{\theta} - Y_{test}||_2}{||Y_{test}||_2} \tag{3}$$

We have also investigated the use of locally weighted linear regression for predicting visitor arrival statistics, whereby (2) is replaced with a least-squares system $\min_\theta \sum_i w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$. We used Gaussian weights $w^{(i)} = \exp(-||x - x^{(i)}||^2/(2\tau^2))$ with bandwidth parameter $\tau = 11$ for predicting global visitor arrival statistics, and $\tau = 8$ for predicting visitor arrivals from Western nations.

To estimate the generalization error of our algorithms, we used k-fold cross validation with $k = 5$. In particular, for each year in the range 2006-2010, we trained our algorithms on four of the years and tested on the remaining year. We then set our estimate of the generalization error equal to the average test error observed in the five scenarios.

To select features for our models, we used a modification of forward search that allows for the addition of higher-cardinality subsets of features when the addition of singletons ceases to provide improvement. More precisely, given a default feature set $\mathcal{F}_0$ and a set of candidate features $\mathcal{C}$, we perform the following iteration:

1. Initialize $\mathcal{F} = \mathcal{F}_0$ and $s = 1$.
2. While $s \leq s_{max}$,

   Set $\varepsilon_0 = \varepsilon(\mathcal{F})$.

   Let $\mathcal{S}^* = \arg \min_{\mathcal{S} \subseteq \mathcal{C}, |\mathcal{S}| = s} \varepsilon(\mathcal{F} \cup \mathcal{S})$.

   If $\varepsilon(\mathcal{F} \cup \mathcal{S}^*) > \varepsilon_0$, set $s := s + 1$. Else set $\mathcal{F} := \mathcal{F} \cup \mathcal{S}$ and $\mathcal{C} := \mathcal{C} \backslash \mathcal{S}$.

In our studies, we used $s_{max} = 2$, and we used a candidate set $\mathcal{C}$ consisting of the search volume indices of all n_q search queries over the four months $i-1, i-2, i-3, i-4$ preceding each month of inquiry i. Our default feature set $\mathcal{F}_0$ consisted of the visitor arrival statistics from the months $i-1, i-2, i-12, i-13$ preceding each month of inquiry i.

4 Results

Table 1 lists the estimated generalization errors obtained from our weighted and unweighted linear regression algorithms, as well as the corresponding search queries selected by our feature selection algorithm. Overall, the locally weighted algorithm performed best, exhibiting generalization errors of 0.047 and 0.063 on the Western and global arrival predictions, respectively. Both algorithms performed better when predicting arrivals of visitors originating from Western nations compared to visitors from all nations.

Interestingly, both English and Chinese search queries appear to be relevant predictors of Hong Kong visitor arrival statistics, even when only considering visitors originating from Western nations.

In Fig. 1, we show training and test set visitor arrival predictions for visitors originating from Western nations, as computed with locally weighted linear regression. The closeness of the fits lends credence to the predictive power of search volume indices in the context of tourism forecasting.

5 Conclusion and Future Work

We have presented and evaluated an improved approach to the problem of predicting tourism rates. Query-specific search data was used, instead of aggregate search data, and a feature selection algorithm was implemented to choose the most relevant queries. Hong Kong's monthly visitor arrival statistics for the years 2005-2010 were used as the data set for testing. The algorithm was evaluated using k-fold cross-validation. Along with exhibiting a low training error as with Choi and Varian's algorithm, our algorithm was found to exhibit a low test set error. The results indicate that web search volume histories provide a useful predictor of tourism rates. Future directions for the work include using advanced machine learning techniques and more features to get an even more accurate estimation of the tourism rates. Along with tourism, this study is relevant to a wide variety of other industries that can benefit from analysis of web search volume histories to predict useful trends.

	Unweighted								Locally Weighted							
	Western Nations				All Nations				Western Nations				All Nations			
Query Lag (Months)	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Hong Kong Flights	✓	·	✓	·	·	·	·	·	·	·	·	·	·	·	·	·
Hong Kong Map	✓	✓	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Hong Kong Tourism	✓	·	·	·	·	·	·	·	·	✓	✓	·	·	·	·	·
Hong Kong Disneyland	✓	✓	·	·	·	·	·	·	·	·	·	·	·	·	✓	✓
Hong Kong Hotels	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Hong Kong Visa	·	✓	·	·	·	·	·	·	·	✓	·	·	·	·	·	·
Hong Kong Airlines	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Hong Kong Travel	✓	✓	·	·	·	·	·	·	·	·	·	·	·	·	·	·
香港航班	·	✓	·	✓	·	·	·	·	·	·	✓	✓	·	·	·	·
香港机票	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
香港	✓	·	·	·	·	·	✓	·	·	·	·	·	·	·	·	·
香港机场	·	·	✓	·	·	·	·	·	·	·	·	·	·	·	·	·
香港酒店	✓	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
香港旅游	·	·	·	·	·	·	·	·	·	·	·	·	✓	·	·	✓
香港地图	·	·	✓	·	·	·	·	·	·	·	·	·	·	·	·	·
香港迪士尼乐园	·	✓	·	✓	·	·	✓	·	✓	·	·	·	·	·	·	·
Generalization error	0.053				0.070				0.047				0.063			

Table 1: Search queries and corresponding lag times selected by our feature selection algorithm.

References

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, et al. Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014 (2009).

[2] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting Consumer Behavior with Web Search. *Proceedings of the National Academy of Sciences* 107(41): 17486-17490 (2010).

[3] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, Antii Ukkonen, and I. Weber. Web Search Queries Can Predict Stock Market Volumes (2011). http://arxiv.org/abs/1110.4784

[4] B. M. Althouse, Y. Y. Ng, and D. A. T. Cummings. Prediction of Dengue Incidence Using Search Query Surveillance. *Public Library of Science* 5(8): 1-7 (2011).

[5] Choi, Hyunyoung and Varian, Hal R. Predicting the Present with Google Trends. *Google Research Blog* (2009). http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html.

[6] Google Insights, http://www.google.com/insights/search/

[7] Hong Kong Tourism Board Research Statistics, http://partnernet.hktourismboard.com/

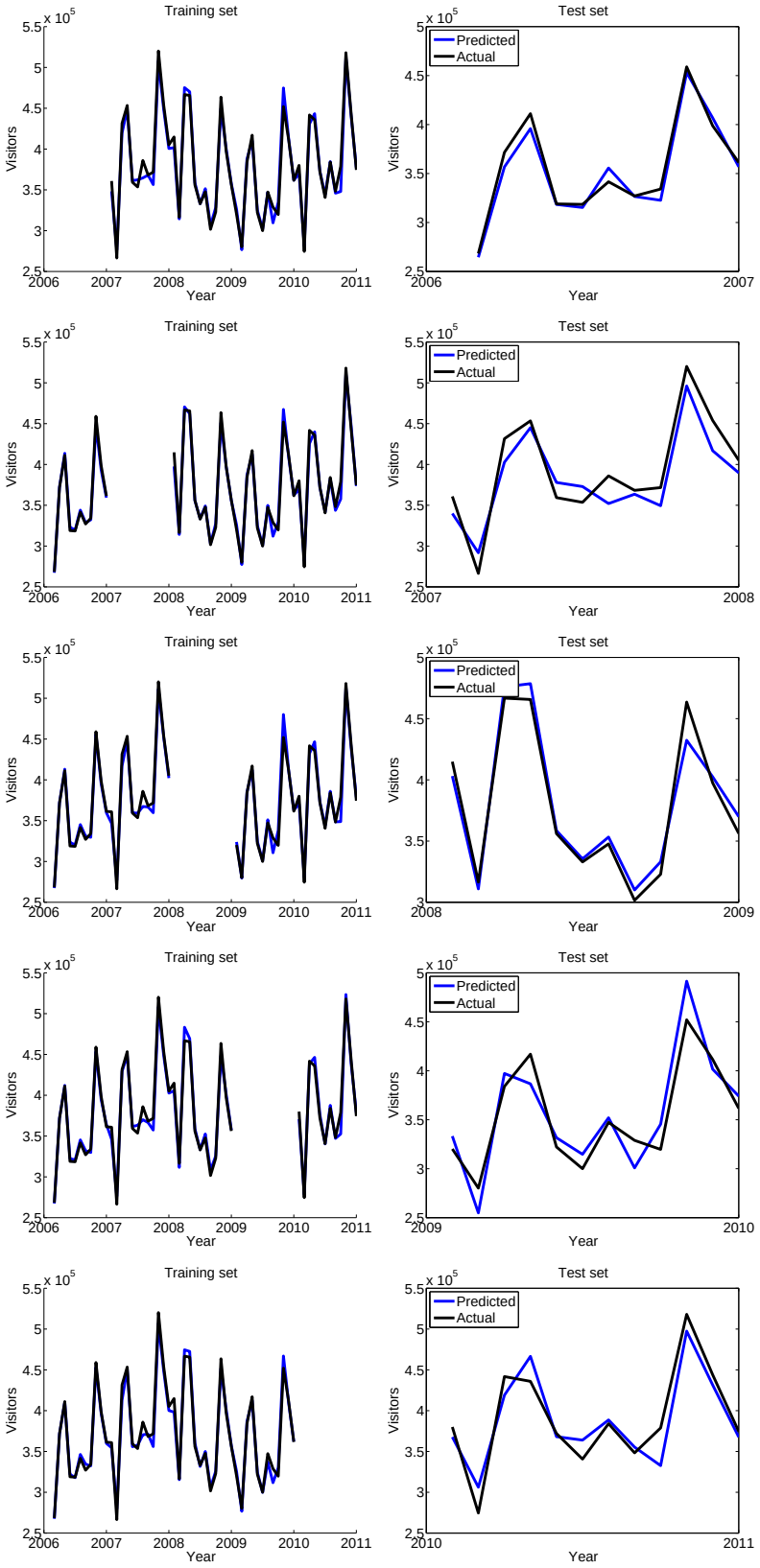

Figure 1: Predicted and actual visitor arrival statistics over the training set/test set time periods, restricted to visitors originating from Western nations, as computed with locally weighted linear regression. Left column: Training set. Right column: Test set.