Improving Tourism Prediction Models Using Climate and Social Media Data: A Fine-Grained Approach

Amir Khatibi,[1] Fabiano Belem,[1] Ana P. Silva,[1] Dennis Shasha,[2]
Jussara M. Almeida,[1] Marcos A. Gonçalves[1]
1) Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil
{amirkm, fmuniz, ana.coutosilva, jussara, mgoncalv}@dcc.ufmg.br
2) Department of Computer Science, New York University, New York, USA
shasha@curant.nyu.edu

Abstract

Accurate predictions about future events is essential in many areas, one of them being the Tourism Industry. Usually, countries and cities invest a huge amount of money in planning and preparation in order to welcome (and profit from) tourists. An accurate prediction of the number of visits in the following days or months could help both the economy and tourists. Prior studies in this domain explore forecasting for a whole country rather than for fine-grained areas within a country (e.g., specific touristic attractions). In this work, we suggest that accessible data from online social networks and travel websites, in addition to climate data, can be used to support the inference of visitation count for many touristic attractions. To test our hypothesis we analyze visitation, climate and social media data in more than 70 National Parks in U.S during the last 3 years. The experimental results reveal a high correlation between social media data and tourism demands; in fact, in over 80% of the parks, social media reviews and visitation counts are correlated by more than 50%. Moreover, we assess the effectiveness of employing various prediction techniques, finding that even a simple linear regression model, when fed with social media and climate data as input features, can attain a prediction accuracy of over 80% while a more robust algorithm, such as Support Vector Regression, reaches up to 94% accuracy.

Key words: Tourism demand prediction, time-series analysis, social media and climate data, machine learning

Introduction

The tourism industry has been growing steadily since the 2009 global economic and financial crisis. Thus, decision makers in industries like transportation, accommodation facilities and traveling agencies, all need to have good estimates of future demand. Many factors can interfere with the cyclic and/or trending behavior of visitation rates. For example, exchange rate fluctuations, epidemics, fuel price, crime rates, and even hit movies (Riley and Van Doren 1992) can cause dramatic deviations in tourism demand. Fortunately, most of these factors are reflected quickly in social media (Asur and Huberman 2010). Most prior efforts aimed at forecasting touristic activities (Cankurt and Subasi 2015) have proposed prediction models over an entire country and *not* for specific regions or attractions. Even sites that recommend attractions (e.g., (Borras, Moreno, and Valls 2014)) do not gather attendance statistics. Their main focus is on the users/tourists and not on attraction management. By contrast, our main focus in this paper is on the prediction of visits to the attractions themselves to help attraction managers.

Our main contribution in this paper is to show that social media from TripAdvisor[1] as well as climate data from the U.S National Climate Data Center[2] can accurately forecast tourism demand at the single attraction level. Our prediction techniques ranged from simple to advanced: linear regression, Support Vector Regression (SVR) (Cortes and Vapnik 1995), General Regression Neural Network (GRNN) (Specht 1991), Seasonal ARIMA (Wei 1994) and Seasonal ARIMA with exogenous data (SARIMAX) (Peter and Silvia 2012).

Our experimental results show that by exploiting the social media and climate data, even a simpler linear regression model can attain a prediction accuracy of over 80% while Support Vector Regression (SVR) has a superior result of 94% accuracy.

Related Work

There are plenty of prior efforts to use data from Location Based Social Networks (LBSN) (e.g., Foursquare and Yelp) to study the mobility of tourists and citizens (Li and Chen 2009; Cho, Myers, and Leskovec 2011; Hasan, Zhan, and Ukkusuri 2013; Hossain et al. 2016). For example, in (Georgiev, Noulas, and Mascolo 2014), the authors study the dining and shopping behaviors during the 2012 Summer Olympics in London using Foursquare check-in data.

There is also work on using external information from the Web to estimate future touristic demands, but usually only in a coarse-grained fashion (e.g. country or city-level). For instance in (Cankurt and Subasi 2015), the authors use Multi Layer Perceptron (MLP) regression (Murtagh 1991) and Support Vector Regression (SVR) models in order to make multivariate tourism forecasting for Turkey. The authors use a diverse set of features such as wholesale prices index, US Dollar selling, hotel bed capacity and number of tourism agencies in the country. The use of robust models

[1] http://www.tripadvisor.com/PressCenter-c4-FactSheet.html

[2] https://www.ncdc.noaa.gov/cag/time-series/us/

like SVR and MLP produced accurate predictions, with best results obtained with the SVR model (also our experience).

A few studies focus on analyzing social media data, such as check-ins and comments posted by tourists, to infer the visitation density over time. For instance, in (Spencer A. Wood and Lacayo 2013), the authors use the locations of photographs in Flickr, a famous image hosting website, to estimate visitation counts in some recreational sites around the world. They report the relationship between the empirical estimates of mean annual visitor user-days and those derived from photographs. This is best described by a polynomial function with $R^2 = 0.386$ and that categorizing the recreational parts into more specific profiles could improve correlations. However, they do not address predictions.

In (Fisichelli et al. 2015), the authors analyze the climate and visitation data for U.S. national parks using a third-order polynomial temperature model and argue that it explains 69% of the variation in historical visitation trends. By exploiting a richer feature set, including social media data, we were able to achieve much higher accuracy levels. We show that by exploiting social media as well as climate data, even a simple linear regression model can attain an accuracy of over 80% while a more robust algorithm, Support Vector Regression, produced a superior result of 94% accuracy.

In more recent work (Khadivi and Ramakrishnan 2016), the authors use Wikipedia usage trends in order to forecast the tourism demand for Hawaii. However, they report the accuracy of their prediction results only by RMSE using a autoregressive exogenous model where the external variable is a Wikipedia usage trend time series. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular data and not between datasets, as it is scale-dependent (Hyndman and Koehler 2006). Although there are interesting statements and results in this work, there is no comparison of prediction models to other baselines nor assessment of the results in a comparable manner.

In sum, in comparison to prior work, the novelties of our work are the focus on fine-grained prediction (i.e. attraction-level) of visitation counts and the improvement over prior results by exploiting available social media data. To the best of our knowledge, we are the first to perform such joint analysis. By mixing climate and social media data in order to predict touristic demands, we are able to produce more robust forecasting models while using few features in a simpler linear kernel SVR model.

Problem Statement

Our goal is to predict the visitation count at specific touristic locations (notably US national parks) exploiting social media and climate data. Given a target place, this prediction problem can be formally stated as follows. We first discretize time into a series of equally spaced non-overlapping time windows of given duration (e.g., a month, a week, a day). We define $X = \{X_1, X_2, ..., X_m\}$ as a set of time series, one for each type of social media and climate data used as input (e.g., number of reviews, average temperature). A time series X_i is a sequence $\{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(t)}\}$, where $x_i^{(t)}$ denotes the value of variable X_i measured during time window t for the specific touristic place that is target of prediction. For a specific time series X_i, we denote the sequence of samples between time windows 1 and t by $X_i^{(t)}$, i.e., $X_i^{(t)} = \{x_i^{(1)}, x_i^{(2)}, ..., x_i^{(t)}\}$. Our goal is to forecast $y^{(t)}$, the tourism demand (i.e., visit counts) during time window t at the target touristic place, based on the set of past measurements that are available up until time window $t - k$ (for $k > 0$) (Mourão et al. 2008; Salles et al. 2010), i.e., $\{X_1^{(t-k)}, X_2^{(t-k)}, ..., X_m^{(t-k)}\}$. In other words, we want to develop a forecasting function f such that: $\hat{y}^{(t)} = f(X_1^{(t-k)}, X_2^{(t-k)}, ..., X_m^{(t-k)})$, where $\hat{y}^{(t)}$ is the predicted value for $y^{(t)}$. The forecasting function f is specific for each target place and is learned based on historic (training) data using different machine learning techniques.

Experimental Methodology

In our study, we used datasets[3] collected from three different sources. The first data source is the U.S. National Park Service website. This portal displays official recorded visitation statistics for national parks in the U.S. We downloaded the monthly total numbers of visitors for a number of national parks in the period of January 1996 to February 2016 [4] to use as the ground truth in our study.

We also collected social media data from TripAdvisor, the largest travel website, with more than 570 million reviews and 455 million monthly average unique visitors[5]. We conducted a crawling on the graph of TripAdvisor pages, starting from the page for U.S national parks. We got the reviews and ratings for those U.S national parks with a travel contents page and then we aggregated the results monthly to make it comparable with our ground-truth dataset. At the end, we achieved monthly number of reviews along with the average rating scores of reviewers during the period of January 2011 till September 2016 for a number of parks.

Finally, climate data was collected from the U.S National Climate Data Center[6]. We had to set up a web crawler specific for this case since the climate data is aggregated for climate divisions in U.S states and regions. As a result, for each U.S. national park, we used the climate data associated with the closest climate division considering the earth curvature distance between them. Specifically, we collected the monthly minimum, maximum and average temperatures besides the monthly precipitation. Our climate data covers the period of January 2000 to November 2016 [7].

Model Learning and Parameterization

In our prediction experiments, we perform cross-validation to learn the prediction models as follows. For each national

[3] For reproducibility, all datasets and codes are in https://tinyurl.com/ycsws93j

[4] https://irma.nps.gov/Stats/

[5] TripAdvisor's fact sheet: http://www.tripadvisor.com/factsheet.html

[6] https://www.ncdc.noaa.gov/cag/time-series/us/

[7] We filtered out parks with fewer than 200 reviews in the most recent 3 years (i.e. less than an average of 5 reviews per month), resulting in 76 national parks.

park considered, we first divide each time series into two parts: the training set, consisting of the first 30 months of data, and the test set, consisting of the remaining 4 months of data. The training set is used to "learn" the prediction model, while the test set is used to evaluate the learned model and report accuracy results. Note that a specific model is learned (and later evaluated) for each park, and thus, there is a different parameter choice for each park. For the sake of brevity, the values reported below are averages over all parks.

For both Support Vector Regression (SVR) and General Regression Neural Networks (GRNN), which present tuning parameters, the training set is further split randomly into two parts. One portion, containing 30% of the training data, is used as a validation set for parameter tuning. The other portion is used to build the prediction function.

Prediction Models - comparative results

Here we present the comparison results of the five analyzed prediction methods (Linear Regression, SVR, GRNN, SARIMA and SARIMAX) when applied to our test sets of the final four months of data. We also compare the effectiveness of different combinations of prediction variables, including the past number of visits (#Visits), Climate (i.e., Average, Minimum and Maximum Temperatures, besides Precipitation), and Social Media (i.e., #Reviews and Average Rating) variables.

We evaluate the accuracy of the considered prediction techniques by means of the Mean Absolute Percentage Error (MAPE) which is a widely used measure of forecasting error (Lewis 1982). Table 1, reports results separately for parks with high accuracy (low MAPE), moderate prediction accuracy and low accuracy by ranges of MAPE accordingly ($< 10\%$), (10 to 25%) and ($> 25\%$) and reporting the percentages of parks that fall into each group based on prediction results for each model. As can be seen, the support vector regressino (SVR) model has the best prediction accuracy having a MAPE value $< 25\%$ (meaning a high or good prediction accuracy) for over 94% of the U.S. national parks. The second best method is Linear Regression with a high or good prediction accuracy for over 84% of the parks. This shows that using only Social Media and Climate data, even with a simple linear model, can provide a good prediction accuracy. Figure 2 shows the prediction results by complete feature set, removing climate features and then removing all social media features for the best prediction method, SVR; The figure illustrates the complementarity of social media and climate in achieving high prediction accuracy.

Table 1: Prediction accuracy results for each technique - all values are in percentages (%); **bold** values are showing prediction technique with higher percentage of parks with a good prediction results

MAPE	SVR	SARIMAX	Linear Reg.	GRNN	SARIMA
lower 10%	**22.4**	13.2	15.8	1.3	7.9
10% to 25%	**72.4**	68.4	68.4	34.2	55.3
over 25%	5.26	18.42	15.79	65.79	36.84

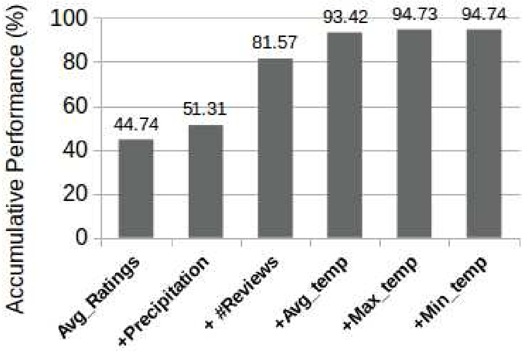

Figure 1: Support Vector Regression (SVR) accumulative features performance as measured by the percentage of parks having low MAPE values as features are added. The reviews vastly improve the performance over the precipitation and ratings alone. Temperature also gives a boost though a lesser one. The test set is the last four months of data.

Feature analysis

We performed two types of analysis: Models with all but one feature and Models with only one feature. Figure 1 presents the accumulative effectiveness of the features prediction, as measured by the percentage of the parks with MAPE values lower than 25%. In order to calculate accumulative results, we started from the worst individual feature and added the others in increasing order of their individual effectiveness. We note that features Avg_Rating and Precipitation, in isolation, do not present good prediction performances (lower than 50% accuracy), which is consistent with the low correlation results of these features with the number of visits. The inclusion of the social media feature #Reviews enormously increases the prediction accuracy. Further improvements are obtained when adding a temperature feature, and the prediction accuracy keeps stable after adding the remaining temperature features.

In addition to training models with climate and social media data, we trained a SARIMA model using only the history of visits in the last 30 month in order to predict the next 4 months visitations. When using all features, Support Vector Regression (SVR) has a MAPE accuracy under 25% for over 94% of the parks while it is 81% for SARIMAX, 84% for Linear Regression and 35% for GRNN. In the case of SARIMA, using only the history of the visits, MAPE accuracy under 25% is just 63%.

General Discussion Our results show that social media in addition to climate data can predict monthly visitation rates much better than either alone for National Parks in the United States. While this data is obviously specific, we suspect that both the value of social media data and its complementarity to climactic information hold widely. After all, most cities share the same climate at all their sites, but some districts are just more attractive. The attractiveness is captured largely by social media. We have shown further that Support Vector Regression is a good way to combine these

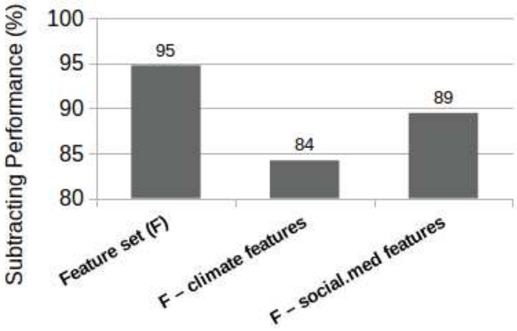

Figure 2: Support Vector Regression (SVR) performs best with the complete feature set. Climate alone or social media alone don't do well. Each complements the other.

two disparate sources of information. In this work, the granularity of time was chosen to be monthly since the official ground-truth data was available and aggregated monthly. However we have reason to believe the same methodology can be used ad weekly, daily and perhaps even hourly granularities.

Conclusion and Future Work

This paper has shown how to use Social Media combined with climate data to forecast touristic demand for specific tourist attractions. Our dataset consisted of visitation data from more than 70 National Parks in the United States. Social media and climactic data each contribute to prediction accuracy. Further, the two are complementary and are well synthesized using Support Vector Regression.

Sites that depend less on weather (such as museums or theaters) may have other important non-social media features (like genre). In future work, we plan to include such features and to apply text analysis and classification techniques on top of tourist reviews in social media.

Acknowledgements

This work is supported by CNPq, CAPES, and Fapemig and INRIA (for Shasha).

References

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. In *WI-IAT*, volume 1, 492–499.

Borras, J.; Moreno, A.; and Valls, A. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications* 41(16):7370–7389.

Cankurt, S., and Subasi, A. 2015. Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components. *Balkan Journal of Eletrical and Computer Engineering* Vol.3, No.1.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *KDD*, 1082–1090.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Fisichelli, N. A.; Schuurman, G. W.; Monahan, W. B.; and Ziesler, P. S. 2015. Protected area tourism in a changing climate: Will visitation at us national parks warm up or overheat? *PLoS ONE* 10(6).

Georgiev, P.; Noulas, A.; and Mascolo, C. 2014. Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. ICWSM.

Hasan, S.; Zhan, X.; and Ukkusuri, S. V. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Urbicomp*, 6.

Hossain, N.; Hu, T.; Feizi, R.; White, A. M.; Luo, J.; and Kautz, H. 2016. Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. *arXiv preprint arXiv:1603.03181*.

Hyndman, R. J., and Koehler, A. B. 2006. Another look at measures of forecast accuracy. *Int. J. Forecasting* 22(4):679–688.

Khadivi, P., and Ramakrishnan, N. 2016. Wikipedia in the tourism industry: Forecasting demand and modeling usage behavior. In *ICWSM*, 4016–4021.

Lewis, C. D. 1982. *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Li, N., and Chen, G. 2009. Analysis of a location-based social network. In *Comp. Sci. & Eng.*, volume 4, 263–270.

Mourão, F.; da Rocha, L. C.; Araújo, R. B.; Couto, T.; Gonçalves, M. A.; and Jr., W. M. 2008. Understanding temporal aspects in document classification. In *WSDM'08*, 159–170.

Murtagh, F. 1991. Multilayer perceptrons for classification and regression. *Neurocomputing* 2(5):183–197.

Peter, Ď., and Silvia, P. 2012. Arima vs. arimax–which approach is better to analyze and forecast macroeconomic time series. In *Math. Methods in Economics*, 136–140.

Riley, R. W., and Van Doren, C. S. 1992. Movies as tourism promotion: A pull factor in a push location. *Tourism management* 13(3):267–274.

Salles, T.; da Rocha, L. C.; Pappa, G. L.; Mourão, F.; Jr., W. M.; and Gonçalves, M. A. 2010. Temporally-aware algorithms for document classification. In *SIGIR'10*, 307–314.

Specht, D. F. 1991. A general regression neural network. *IEEE transactions on neural networks* 2(6):568–576.

Spencer A. Wood, Anne D. Guerry, J. M. S., and Lacayo, M. 2013. Using social media to quantify nature-based tourism and recreation. *Scientific Report* 3.

Wei, W. W. S. 1994. *Time Series Analysis Univariate and Multivariate Methods*, volume 2. Pearson Addison Wesley.