Wikipedia in the Tourism Industry: Forecasting Demand and Modeling Usage Behavior

Pejman Khadivi
Discovery Analytics Center
Computer Science Department
Virginia Tech, Blacksburg, Virginia
Email: pejman@cs.vt.edu

Naren Ramakrishnan
Discovery Analytics Center
Computer Science Department
Virginia Tech, Arlington, Virginia
Email: naren@cs.vt.edu

Abstract

Due to the economic and social impacts of tourism, both private and public sectors are interested in precisely forecasting the tourism demand volume in a timely manner. With recent advances in social networks, more people use online resources to plan their future trips. In this paper we explore the application of Wikipedia usage trends (WUTs) in tourism analysis. We propose a framework that deploys WUTs for forecasting the tourism demand of Hawaii. We also propose a data-driven approach, using WUTs, to estimate the behavior of tourists when they plan their trips.

1 Introduction

Tourism is considered as one of the most profitable industries around the world and, hence, tourism demand forecasting is important in various domains such as business planning and assessing economic activity in a region. Various prediction techniques have been used in the literature for demand forecasting (Song and Li 2008). Recently, there have been different studies that suggest that the use of search engine datasets, such as Google Search Trend data can help improve forecasting accuracy (Bangwayo-Skeete and Skeete 2015; Yang and others 2015).

In recent times, online browsing activity has been identified as a precursor to planning travel and visits to far flung destinations. Analyzing such online activity can provide detailed information about places of interest, volume of interest, and spikes/troughs in attention. Our hypothesis in this paper is that, by looking at Wikipedia usage trends (WUT), we may be able to improve the accuracy of tourism demand forecasts.

Wikipedia usage trends have been previously used for forecasting applications in (Hickmann and others 2015) to forecast the Influenza season. Pertaining to the tourism industry, Wikipedia has been used previously for sightseeing recommendation and narrative generation purposes (Fang and others 2015; Hecht and others 2007). However, to the best of our knowledge, there has been no research on using WUTs for tourism demand forecasting. Beside demand forecasting, another important tourism-related issue is to extract the behavior of people when they make decisions about their trips, e.g., how early he/she reads about the National Mall or plans to book his/her flight. The answers to such questions are useful for resource allocation and marketing purposes.

In this paper, we aim to use Wikipedia usage trends for the purpose of tourism analysis. The main contributions of this paper are twofold. First, we propose forecasting algorithms to deploy WUTs and Google Search Trends for tourism demand forecasting. Second, we propose a framework for the extraction of tourists’ behavior in trip planning. We performed various experiments on a Hawaii tourism dataset. Results show that WUTs can improve the accuracy of demand forecasting. Furthermore, the extracted tourist behaviors are consistent with the results of the surveys conducted by tourism departments. Thus, our contributions are:

- Developing a regression approach that deploys WUTs to improve the accuracy of tourism demand forecasting.
- Proposing a novel framework for tourist behavior extraction from WUTs and tourism demand time series.

2 Related Work

In order to forecast tourism demand, different univariate and multivariate regression methods have been deployed in the literature. Song and Witt (2006) used vector autoregressive (VAR) model to forecast tourist arrivals to Macau. In addition to a tourist arrival time series, they also incorporated the costs of living in Macau and the originating country. Gunter and Onder (2015) compared the performance of various forecasting models, such as V AR and ARMA, to predict international tourism demand in Paris. Athanasopoulos and de Silva (2012) deployed multivariate exponential smoothing methods for the forecasting of tourism demand for Australia and New Zealand. Chan and others (2005) used various GARCH models to address variations in tourism demand caused by economical instabilities. The use of gravity models for tourism demand analysis has been addressed in (Morley and others 2014). Song and Li (2008) provide a detailed survey on various forecasting techniques.

A number of tourism demand forecasting approaches are based on usage trend data that are supplied by search engines. Xiang and Pan (2011) performed behavior analysis to understand how tourists use search engines for travel planning. They determined which keywords are more important when tourists use search engines. Choi and Varian (2012)
used Google search trends (GST) to forecast the number of monthly visitors of Hong Kong. They use data from the first two weeks of a month to forecast the total number of visitors in the whole month. In (Bangwayo-Skeete and Skeete 2015), GST data about hotels and flights search terms for popular tourist destinations in the Caribbean are used for forecasting tourism demand. Yang and others (2015) use GST and Baidu trends to predict tourism demand in the Hainan Province of China. In (Artola and others 2015), GST is used to improve the forecasting of tourism demand for Spain.

3 Problem Formulation

In this paper, we use three sets of time series: tourism demand, WUTs, and GSTs. We assume that we have a set of $P$ Wiki-pages related to a specific tourism destination, $\mathcal{W} = \{W_1, \ldots, W_P\}$. The WUT of a specific page, $W_i$, is a time series denoted by $W_i = w_i^1, \ldots, w_i^n$, where $w_i^j$ is the total number of times that $W_i$ has been used at time $t$. We also assume that we have a set of $Q$ Google search terms, $\mathcal{G} = \{G_1, \ldots, G_Q\}$ and that the GST data corresponding to $G_i$ is represented by $G_i = g_i^1, \ldots, g_i^n$. The tourism demand time series is illustrated by $Y = y_1, \ldots, y_t$ where $y_t$ is the total number of monthly visitors at time $t$. In a similar way, $\tilde{Y}^D$ and $\tilde{Y}^F$ are the time series that represent the total number of domestic and international visitors, respectively. For a specific time series $\mathcal{X}$, we denote the sequence of samples between time points 1 and $t_0$ by $\mathcal{X}_{t_0}$, i.e. $\mathcal{X}_{t_0} = [x_1 \ldots x_{t_0}]^T$. In the following subsections we introduce two problems that are studied in the rest of the paper.

3.1 The Demand Forecasting Problem

The demand forecasting problem is to forecast the volume of tourism demand (i.e. tourist arrival) in the future. In the most general setting, forecasting the tourism demand at time $t$ based on time series $\mathcal{D}$ which is available up to time $t - k$, i.e. $\mathcal{D}_{t-k} = d_1, \ldots, d_{t-k}$, can be formulated as follows:

$$\hat{y}_t = \mathcal{F}(\mathcal{D}_{t-k})$$  \hspace{1cm} (1)

where $\mathcal{F}$ is the forecasting function and $\mathcal{D}$ can be any univariate or multivariate time series. As an example, if we aim to forecast tourism demand based on the observed values of $\tilde{Y}$ we will have $\hat{y}_t = \mathcal{F}(\tilde{Y}_{t-k})$.

The problem of tourism demand forecasting using WUT of a Wiki-page, $W_i$, is formulated in the following equation:

$$\hat{y}_t = \mathcal{F}(\tilde{Y}_{t-k}, W_i^t)$$  \hspace{1cm} (2)

where $\mathcal{F}(\cdot)$ is the estimation function. In Eq. 2, the forecasting lead time is $k$ months, i.e. we aim to forecast the demand $k$ months earlier. The forecasting problem using GST and other data sources can be formulated in a similar way.

3.2 The Behavior Extraction Problem

Let us assume that a typical trip planning has $n$ activities, $\mathcal{A} = \{A_1, \ldots, A_n\}$. As an example, these activities may include flight reservation, hotel booking, and planning to visit different attractions. The problem of tourists behavior extraction is to determine how long before the actual trip time, a typical tourist performs each of these activities and is reported as a distribution over time. We use $a_{A_i}^j$ to show the fraction of tourists that perform activity $A_i$ in $j$ months earlier than the actual trip time. Then, we will have:

$$\sum_{j=0}^{\infty} a_{A_i}^j = 1 \quad i = 1, \ldots, n. \hspace{1cm} (3)$$

In reality, a finite upper bound can be considered for $j$. The traditional approach to acquire the $a_{A_i}^j$ values is through conducting surveys (HTA Report 2002-2013). However, since people use online resources for trip planning, we expect that usage trends of online resources can be used to extract average tourist behavior during trip planning. Wiki-pages, based on their contents, can be classified into different categories. As an example, a Wiki-page that is about a national park can be classified as an attraction. This classification can be conducted based on the trip planning activities in $\mathcal{A}$. Then, the problem of tourists behavior extraction using WUTs is to determine how tourists visit the pages in each category prior to their trip and is formulated as follows:

$$a_{A_i}^j = B(\tilde{Y}^D, \tilde{Y}^F, W_{A_i}^j)$$  \hspace{1cm} (4)

where, $B$ is the estimation function and $W_{A_i}^j$ is a set of WUT time series of pages that belong to the category of $A_i$.

4 Dataset Description

In this paper, we mainly use Wikipedia pages and their usage trends for tourism in Hawaii. Monthly tourism demand time series for Hawaii are available through the website of the Department of Business, Economic Development and Tourism of the state of Hawaii\(^1\). We collected Wikipedia pages and their usage trends from English Wikipedia and Wikipedia Trends\(^2\) websites. For this purpose, we conducted a breadth-first-search approach on the graph of Wikipedia pages, beginning from the page for Hawaii\(^3\). We retrieved all the Wiki-pages related to Hawaii by extracting pages that contain a minimum number of mentions of Hawaii. Using this approach we retrieved 3,126 pages. Also, we collected GST time series for Hawaii, Hawaii flight, and Hawaii hotel search terms. Considering the data availability, we focused our study on the time frame between January 2008 and April 2015.

Page Classification: For further analysis, Wiki-pages are classified into the following tourism-related categories:
- Attractions: pages that provide information about various types of attraction such as beaches and museums.
- Flights: pages that provide information about airlines and airports in Hawaii.
- Schools and universities: pages that provide information about schools, colleges, and universities in Hawaii.
- Events: pages that provide information about various events in Hawaii such as festivals and sport events.

\(^1\)dbedt.hawaii.gov/visitor/tourism/
\(^2\)www.wikipediatrends.com
\(^3\)en.wikipedia.org/wiki/Hawaii
5 Forecasting using Wikipedia Usage Trends

In this section, we aim to explore how WUT influences the accuracy of tourism demand forecasts. A well-known approach for tourism demand forecasting is to use autoregressive models, AR(m). Here, we deploy this approach as the building block of other regression methods. The general model of AR(m) is as follows:

\[
y_t^{AR} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j}
\]

where, \( k \) is the forecasting lead time, \( \alpha_j \)'s and \( \beta \) are regression coefficients, and \( m \) is the order of the model. In the following subsections, we extend this model to incorporate WUT and GST time series in the forecasting process.

5.1 Regression with Wikipedia Usage Trends

The general model of tourism demand forecasting using WUT is defined in Eq. 2. Using a linear regression setting for the forecasting function \( F(\cdot) \), the forecasting model based on the usage trend of Wiki-page \( W_i \) is as follows:

\[
y_t^{W_i} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j W_{t-k-j}^{i}\tag{6}
\]

where \( \beta, \alpha_j \)'s, and \( \gamma_j \)'s are regression coefficients. In the rest of the paper, we denote this model by ARW(\( m \)). Note that ARW(\( m \)) is an autoregressive exogenous (ARX) model where the external variable is a WUT time series.

Experiments show that each of the Wiki-pages results in a slightly different estimation and the accuracy of estimation based on each individual Wiki-page changes over time. To overcome this variation and in order to improve the accuracy of the forecasts, we use an ensemble method to aggregate the estimations that have been achieved based on different Wiki-pages. Let us assume that we have \( P \) Wikipedia pages and we aim to forecast the tourism demand at time \( t \), \( y_t \), based on the data that is available up to time \( t-k \). For this purpose, we name the latest \( v \) months of the available data as the validation set, \( V = \{ y_{t-k-v+1}, \ldots, y_{t-k} \} \). Then, for each Wiki-page, the average performance of ARW(\( m \)) is measured on the validation set, \( V \). The top \( N \) pages that result in the highest accuracy are then selected to perform the final forecasting. We represent this set of top-N Wiki-pages as \( W^{Top}_N \).

Then, the final forecast is performed as follows:

\[
\hat{y}_t^W = \frac{1}{N} \sum_{W_i \in W^{Top}_N} \hat{y}_t^{W_i} \tag{7}
\]

where \( \hat{y}_t^{W_i} \) is determined based on Eq. 6. This algorithm, which we name it as Wiki-based Regression(\( m,N \)), is illustrated in Algorithm 1.

5.2 Regression with Google Search Trends

Let us assume that we have \( Q \) time series of GST and we aim to forecast the tourism demand at time \( t \), \( y_t \), based on the data available up to time \( t-k \). Then, we will have:

\[
\hat{y}_t^{G_i} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j \hat{y}_{t-k-j}^i\tag{8}
\]

The general model of tourism demand forecasting using GST is defined in Eq. 2. Using a linear regression setting for the forecasting function \( F(\cdot) \), the forecasting model based on the usage trend of GST is as follows:

\[
\hat{y}_t^G = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j \hat{y}_{t-k-j}^G\tag{9}
\]

The accuracy of tourism demand forecasts. A well-known approach for tourism demand forecasting is to use autoregressive models, AR(m). Here, we deploy this approach as the building block of other regression methods. The general model of AR(m) is as follows:

\[
y_t^{AR} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j}
\]

where, \( k \) is the forecasting lead time, \( \alpha_j \)'s and \( \beta \) are regression coefficients, and \( m \) is the order of the model. In the following subsections, we extend this model to incorporate WUT and GST time series in the forecasting process.

5.1 Regression with Wikipedia Usage Trends

The general model of tourism demand forecasting using WUT is defined in Eq. 2. Using a linear regression setting for the forecasting function \( F(\cdot) \), the forecasting model based on the usage trend of Wiki-page \( W_i \) is as follows:

\[
y_t^{W_i} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j W_{t-k-j}^{i}\tag{6}
\]

where \( \beta, \alpha_j \)'s, and \( \gamma_j \)'s are regression coefficients. In the rest of the paper, we denote this model by ARW(\( m \)). Note that ARW(\( m \)) is an autoregressive exogenous (ARX) model where the external variable is a WUT time series.

Experiments show that each of the Wiki-pages results in a slightly different estimation and the accuracy of estimation based on each individual Wiki-page changes over time. To overcome this variation and in order to improve the accuracy of the forecasts, we use an ensemble method to aggregate the estimations that have been achieved based on different Wiki-pages. Let us assume that we have \( P \) Wikipedia pages and we aim to forecast the tourism demand at time \( t \), \( y_t \), based on the data that is available up to time \( t-k \). For this purpose, we name the latest \( v \) months of the available data as the validation set, \( V = \{ y_{t-k-v+1}, \ldots, y_{t-k} \} \). Then, for each Wiki-page, the average performance of ARW(\( m \)) is measured on the validation set, \( V \). The top \( N \) pages that result in the highest accuracy are then selected to perform the final forecasting. We represent this set of top-N Wiki-pages as \( W^{Top}_N \).

Then, the final forecast is performed as follows:

\[
\hat{y}_t^W = \frac{1}{N} \sum_{W_i \in W^{Top}_N} \hat{y}_t^{W_i} \tag{7}
\]

where \( \hat{y}_t^{W_i} \) is determined based on Eq. 6. This algorithm, which we name it as Wiki-based Regression(\( m,N \)), is illustrated in Algorithm 1.

5.2 Regression with Google Search Trends

Let us assume that we have \( Q \) time series of GST and we aim to forecast the tourism demand at time \( t \), \( y_t \), based on the data available up to time \( t-k \). Then, we will have:

\[
\hat{y}_t^{G_i} = \beta + \sum_{j=0}^{m-1} \alpha_j y_{t-k-j} + \sum_{j=0}^{m-1} \gamma_j \hat{y}_{t-k-j}^G\tag{8}
\]
one of the important issues in tourism industry is to know when people start to plan their trips and how they use online resources for that purpose. In this section, we propose a data-driven framework based on WUTs to answer this question.

We categorize tourists into two groups: Wikipedia readers and non-Wikipedia readers. A Wikipedia reader tourist is a person that uses Wikipedia as an online resource before planning their trip. The importance of each category of Wiki-pages for domestic and international tourists is illustrated in Fig. 5. In this experiment, similar to Algorithm 1, we used the method of ARW$(m)$ to determine the top-250 Wiki-pages, $W^T_{250}$. Then, we measured the ratio of pages that belong to each category in $W^T_{250}$. This figure depicts that domestic and international tourists have different behaviors. As an example, while for domestic tourists, the Hotel category is highlighted with lead time of 1, for international tourists it is highlighted with lead time 6. We can also observe from this figure that Wiki-pages in Flight, Hotel, Events, and Transportation categories are more important than the other three ones.

### 6 Wikipedia Usage Behavior
are constant coefficients that determine the ratio of Wikipedia reader tourists, and we have:

\[ w_t^i = \sum_{j=0}^M K_{W_t}^D b_j^i y_t^{D+j} + \sum_{j=0}^M K_{W_t}^I c_j^i y_t^{I+j} + \nu_t + \epsilon_t \] (11)

where, \( w_t^i \) is the number of visits to the Wiki-page \( W_t \), and \( y_t^{D+j} \) and \( y_t^{I+j} \) are the number of domestic and international tourists at time \( t+j \), respectively. Furthermore, \( K_{W_t}^D \) and \( K_{W_t}^I \) represent the ratio of domestic and international Wikipedia reader tourists, respectively. Also, \( b_j^i \)'s and \( c_j^i \)'s are constant coefficients that determine the ratio of Wikipedia reader tourists that have read \( W_t \), \( j \) months before the trip time. Note that since coefficients \( b_j^i \) and \( c_j^i \) represent the reading distributions, these coefficients cannot be negative and each set should be summed up to 1. In Eq. 11, \( \nu_t \) and \( \epsilon_t \) are two error components, illustrating the uncertainties that we have in the model. Uncertainty (or variation) in the assumptions and data (i.e. noise in \( K_{W_t}^D \), \( w_t^i \), etc.) is modeled by \( \epsilon_t \) and Wiki-page visits that have a non-tourist reason are shown by \( \nu_t \). The intuition behind this model is that the number of Wikipedia page-visits for a specific area is a linear function of the number of tourists that will visit that specific area (i.e. in this paper Hawaii) in the future, plus some noise.

Yearly trip-planning surveys conducted for Hawaii (HTA Report 2002-2013) show that trip planning activities follow a normal-like distribution. Since we assumed that Wikipedia-reader tourists read Wiki-pages during their trip planning, we expect that the reading distributions follow a normal-like distribution as well. Therefore, in order to apply this prior knowledge to the model and to dictate the estimations model to follow a normal-like distribution we need to add the following set of constraints:

\[ b_0 \leq b_1 \leq \cdots \leq b_{\beta-1} \leq b_\beta \geq b_{\beta+1} \geq \cdots \geq b_M \] (12)

\[ c_0 \leq c_1 \leq \cdots \leq c_{\zeta-1} \leq c_\zeta \geq c_{\zeta+1} \geq \cdots \geq c_M \]

where \( \beta \) and \( \zeta \) are indexes that the maximum value of distributions occur for \( b \) and \( c \), respectively.

In order to estimate the Wikipedia reading distributions of domestic and international tourists, we first drop the error factors, \( \nu_t \) and \( \epsilon_t \), from Eq. 11 to calculate the estimated number of visits to the Wiki-page \( W_t \), i.e. \( \bar{w}_t^i \). Then, we solve the following optimization problem:

\[ b_j^*, c_j^*, K_{W_t}^{D*}, K_{W_t}^{I*} = \arg \min \frac{1}{T} \sum_{t=1}^T (\bar{w}_t^i - w_t^i)^2 \]

\[ s.t. \quad 0 < K_I \leq 1, \quad 0 < K_D \leq 1 \]

\[ 0 \leq b_j^i \leq 1, \quad 0 \leq c_j^i \leq 1, \quad j = 0, 1, \cdots, M \]

\[ \sum_{j=0}^M b_j^i = \sum_{j=0}^M c_j^i = 1 \]

\[ b_0 \leq b_1 \leq \cdots \leq b_{\beta-1} \leq b_\beta \geq b_{\beta+1} \geq \cdots \geq b_M \]

\[ c_0 \leq c_1 \leq \cdots \leq c_{\zeta-1} \leq c_\zeta \geq c_{\zeta+1} \geq \cdots \geq c_M \]

In the optimization problem of Eq. 13, \( M \), \( \beta \), and \( \zeta \) are model parameters which are estimated through cross validation.

Note that solving the problem of Eq. 13 results in two reading distributions: \( b_j^i \)'s for domestic tourists and \( c_j^i \)'s for international ones. It is easy to show that by taking the weighted average of these two distributions, one can calculate the average reading distribution of all tourists. In other words, if \( N \) is the total number of tourists and \( N_D \) is the number of domestic tourists we will have:

\[ a_j^i = \frac{N_D}{N} b_j^i + \frac{N-N_D}{N} c_j^i, \quad j = 0, 1, \cdots, M \] (14)

The average of reading distributions for category of pages (as defined in Section 4) are illustrated in Fig. 6. In order to

Figure 4: Forecasting RMS errors for different regression algorithms with (a) one month lead-time, (b) 3 months lead-time, (c) 6 months lead-time, and (d) 9 months lead-time.

Figure 5: Word clouds illustrating the importance of each class of Wikipedia pages in forecasting the tourism demand.
calculate the average distribution of category $A$ we use the following equation:

$$\bar{a}_j^A = \frac{\sum_{W_i \in A} K_{W_i} a_i^j}{\sum_{W_i \in A} K_{W_i}}$$

(15)

where $K_{W_i}$ is the average ratio of Wikipedia reader tourists for page $W_i$. Similar to Eq. 14, $K_{W_i}$ is calculated using $K_D$ and $K_I$. Fig. 6(a) shows the average reading distributions for all the tourists. The average reading distributions of domestic and international tourists for category $A$, i.e. $b_j^A$ and $c_j^A$, can also be calculated in a similar way and are illustrated in Fig. 6 (b) and (c). Figure 6 also shows the average ratio of Wikipedia reader tourists for each category which is calculated as follows:

$$K = \frac{\sum_{W_i \in A} K_{W_i}}{|A|}$$

(16)

where $|A|$ represents the size of category $A$.

Figure 6(a) depicts that on average, most of the Wikipedia reading activities have occurred about 4 to 8 months prior to the trip. This is inline with the surveys reported in (HTA Report 2002-2013) as the mean decision date for most of the activities are between 4 to 8 months before the actual arrival date. Figure 6(a) also indicates that the most popular categories are Events, General information, and Flights. A comparison of Fig. 6(b) and (c) shows that international tourists start planning their trip about 4 to 6 months before the domestic ones. Furthermore, we can observe that on average, the ratio of Wikipedia reader tourists is higher among international tourists than the domestic ones.

7 Conclusion

In this paper we showed that Wikipedia usage trends can be effectively used in tourism planning. Experimental results indicated that WUT time series improve the accuracy of tourism demand forecasts. We also used WUT and tourism demand time series to estimate the behavior of tourists in trip planning. Results are consistent with the survey results gathered from domestic and international tourists of Hawaii. As future work, we aim to use other sources (e.g. TripAdvisor) to rank the Wikipedia pages and consider the relationship graph of Wiki-pages to improve forecasting and behavior estimation performances.

References