Using Web Mining to Support Low Cost Historical Vehicle Traffic Analytics

Charanjeet Kaur, Diwakar Krishnamurthy, Behrouz H. Far
Department of Electrical and Computer Engineering, University of Calgary, AB, Canada
{ckaur, dkrishna, far}@ucalgary.ca

Abstract—Analyzing historical vehicle traffic data has many applications including urban planning and intelligent in-vehicle route prediction. A common practice to acquire this data is through roadside sensors. This approach is expensive because of infrastructure and planning costs and cannot be easily applied to new routes. In this paper, a low-cost Web mining approach is proposed to address these limitations. Our system gathers information about vehicle commute times, accidents, and weather reports from heterogeneous Web sources. Information from these sources can be combined to support road traffic analytics. We illustrate the utility of our system through a clustering analysis that investigates the traffic patterns of the busiest highway in Calgary along with factors having the most impact on commute time. The analysis shows that most of the accidents are localized around a small section of the highway near the city center and that the commute time in this segment is significantly more than that in other segments. Bad weather increases the typical evening rush hour commute time by 60% for days with moderate accidents and by a factor of 100% for days with large number of accidents. Overall, commute times can vary by a factor of 4 depending on accidents and weather.

Keywords-road traffic; clustering; data analysis; Web mining; traffic management

I. INTRODUCTION

In recent years, there has been an increased interest in road traffic analysis and predictions. Traffic data can be collected in various ways. A common practice to acquire this data is through roadside sensors. This approach is expensive because of infrastructure and planning costs and cannot be easily applied to new routes.

In this paper, we propose Web mining as an alternative approach. This approach leverages the capabilities of existing Web sources and does not require any expensive infrastructure to capture current traffic data. Commonly available APIs are used to capture this already available information from the Web which can then be stored for a longer period of time. This approach is flexible and can be tailored to any route of interest and it can also incorporate new data sources and factors which affect the commute time.

Typically, the data to be collected for vehicle traffic analysis include the following:

- Number of cars, speed, flow, occupancy
- Time of the day, weekday, weekend, day of the month, season, year
- Weather, temperature, humidity
- Road type
- Events scheduled, e.g., hockey games
- Events unscheduled, e.g., fire
- Accidents, detours, lane closures
- Police archives of incidents
- Parking: location, occupancy
- Zones: schools, elderly houses, event locations, historically accident prone areas
- Glare (direction of sun)
- Tweets: incidents, locations
- Points of interest (POIs)

Many of the above information can be collected from multiple Web sources such as Google Maps [1], Twitter [2] and the weather websites. Data mining techniques can then be applied on such information to infer how the traffic pattern on a given road is related to factors such as time of the day, day of the week, accidents, and weather events. In contrast to the traditional approach of relying on specialized roadside instrumentation, this approach is more flexible in that it can be adapted with little effort and cost to analyze any road for which such Web data is available.

This paper describes our efforts to develop such a Web mining driven traffic analytics system. As a proof of concept, our system continuously collects and maintains historical commute time estimates for 18 heavy traffic roads in Calgary, Alberta, Canada. The system also extracts reports of accidents on these roads by mining Twitter posts, i.e., Tweets, related to these accidents and overlays this information with the commute time data. Finally, detailed information about weather conditions such as temperature, snowfall, and snow accumulation on the ground are mined from the Web and associated with the other pieces of data.

We illustrate the utility of our system through a study that investigates the traffic patterns of the busiest highway in Calgary along with factors having the most impact on commute time. Specifically, we use clustering [3] to discern commute time trends over a period of 153 days spanning September 2013 to February 2014. Our study shows that there was at least one accident in all of the workdays in the dataset with most accidents happening between 5 PM to 6 PM. It also shows that most of the accidents are localized around a small section of the highway near the city center and that the commute time in this segment is significantly more than the commute times in other segments. On a day when the weather is good, i.e., no snow falling or on the ground, typical evening rush hour
 commute times increase by a factor of 35% for days with large number of accidents. Whereas bad weather increases the typical evening rush hour commute time by 60% for days with moderate accidents and by a factor of 100% for days with large number of accidents. Overall, commute times can vary by a factor of 4 depending on accidents and weather.

This paper makes several new contributions. Firstly, we are not aware of other studies which have exploited Web mining for traffic analysis. Our methodology can serve as an example for others attempting similar studies. Secondly, we provide new insights on commute time patterns of Calgary’s busiest highway and the factors that influence them. This could for example be useful for those that manage the highway, researchers who want to validate traffic simulators for the highway, and those that are interested in predicting commute times based on historical trends. Finally, we plan to share datasets generated by our system with others for use in their studies.

The paper is organized as follows. Section 2 discusses related work. Section 3 describes our data collection, storage and clustering analysis methodologies. Results of our commute time analysis are presented in Section 4. In Section 5, limitations of this work are discussed and Section 6 concludes the paper.

II. RELATED WORK

There has been an abundance of research work to understand traffic characteristics. Several researchers have exploited data mining techniques such as clustering and regression to characterize as well as predict traffic conditions \[4\][5][6][7][8][9][10]. Most of the existing studies rely on specialized roadside instrumentation such as sensors and cameras to collect traffic information. For example, Jain et al. exploit image processing techniques to automatically infer traffic congestion based on live camera feeds \[8\]. Zhang et al. performed visualized spatial–temporal traffic data analysis to evaluate the traffic situation in the road network level \[9\]. Quek et al. developed a fuzzy neural network model for short term traffic flow prediction \[10\].

Our work is different from all these studies in two important ways. First, it does not rely on video or data feeds from sensors. It is open and low-cost as data is gathered from various Web sources and there is no need for special infrastructure to be in place to collect the data. Second, these studies mainly focus on real-time traffic problems while our system focuses on historical traffic trends. Last, our technique is general and flexible. It is not limited to predefined routes and can be used on any route anywhere for which the Web data is available. Thus it complements the sensor approach by offering a solution for municipalities and businesses to conduct a fast and cost-effective analysis of the routes of their interest.

III. METHODOLOGY

In this section we present the data collection and analysis mechanisms used in this work. The first subsection describes the methodologies used to gather heterogeneous data from multiple Web sources and the overlaying of data from one source to the others. The second subsection describes the methodology behind the clustering analysis on the collected traffic data.

A. Data Collection

Our data collection system design is motivated by our previously stated objective of understanding some factors that influence commute time patterns in major urban roadways. In particular, we are currently collecting data on 18 heavily congested highways and arterial roads in Calgary. Due to space constraints, we focus our analysis in Section IV on the busiest of these roads called Deerfoot Trail. Deerfoot Trail is a multi-lane highway that spans about 50 KM within Calgary and features 21 interchanges \[11\]. It has 3 to 4 lanes in each direction and has a speed limit of 100 KMs/hr. The roadway is the province of Alberta’s busiest highway with traffic volumes ranging between 27,000 and 158,000 vehicles per day \[11\]. Although we focus our analysis on one specific roadway in Calgary, our data collection and analysis methodology can be replicated in a straightforward way for other roads.

Our source of commute time information is the Google Maps website. Given a source and a destination address, Google Maps provides a basic estimate of commute time based on route distance and posted speed limits of segments constituting the route. Furthermore, the service also reports a more advanced measure called the “in current traffic” (ICT) estimate, which represents a commute time estimate that takes into account current traffic conditions on the route. This estimate is obtained by continuously tracking in real-time GPS locations of participating mobile phone users travelling on the target route \[12\].

A number of challenges need to be addressed to exploit the commute time data provided by Google Maps. The Google Maps API \[1\] is an application programming interface which supports a programmatic approach to retrieve the information displayed by the Google Maps website. However, the free version of the API does not provide the advanced ICT commute time estimate. As a result, one needs to query the Web service through a browser, save the resulting Web page, and parse the saved page to extract this measure. Furthermore, the Google Maps website does not support queries that request historical ICT estimates for a given route. Such a feature is crucial for understanding how factors such as time of the day, day of the week, and month of the year impact commute times.

We developed custom C# scripts to continuously collect route information and ICT commute time estimates for specified routes. Both scripts take as input the source and destination GPS coordinates of a target route. To allow finer grained analysis, they also accept as inputs GPS coordinates of sub-segments in the route. Finally, the second script additionally takes as input the time instants at which the ICT commute time estimates need to be collected.

The first script queries the Google Maps API and is invoked once for every route of interest and the sub-segments within those routes. The API returns compact JavaScript Object Notation (JSON) files containing information for the route and its sub-segments. Each file contains several options for traversing from the source to the destination along with the basic commute time estimates for the options. The JSON
format makes it easier to programatically query route information during the traffic analysis phase. However, as mentioned previously, these files do not contain the ICT commute time estimates for the routes.

The second script uses the HTTP protocol to capture the ICT commute time estimates for the route and its sub-segments for various times of the day. It queries the Google Maps website at each of the time of the day instants specified as input. Each query returns a set of Web pages containing the Google Maps response for the route as well as its sub-segments. The script uses a custom parser that we developed to extract the ICT commute time estimates pertaining to the route and its sub-segments from these files. The extracted values are then stored along with other information pertaining to the route and its sub-segments contained in the previously obtained JSON files to facilitate subsequent traffic analyses.

We chose our inputs carefully to limit the number of queries issued through the Google Maps API and the number of queries submitted to the Google Maps website. The Google Maps API imposes a limit of 2,500 queries per day. Moreover, we wanted to limit the number of queries sent to the Google Maps website for obtaining the ICT commute time estimates to avoid our scripts from adversely influencing the Quality of Experience (QoE) of human users of the website. As a result, we limited our time of the day queries to 19 different time instants. Queries were issued at 30-minute intervals during rush traffic periods and 2-hour intervals during other periods. We also limited the number of sub-segments monitored in each road. For example, for Deerfoot Trail we specified 6 nearly equidistant sub-segments with each sub-segment roughly encompassing 3 consecutive interchanges. The start and end coordinates of such sub-segments were obtained manually. We defer automating this task to future work.

We augment the commute time information gathered with traffic accidents information mined from the Twitter social network [2]. After an initial analysis of all traffic related Tweets in Calgary, we selected two sources namely, Canadian Traffic Network Calgary (@CTNCalgary) and 660 News Traffic (@660NewsTraffic). Key reasons for choosing these services were the comprehensiveness of updates and the consistent formatting of the updates, which permitted easy parsing of the location of accidents.

Tweets from these sources were programatically gathered using the Twitter Search API [2]. Twitter Search API is an application programming interface which queries the indices of recent tweets and returns the collection of most recent tweets posted by a specified user. These tweets are retrieved in a JSON file. The API restricts the number of queries per day and limits the number of Tweets returned per query to 100. Consequently, queries are issued once every hour and each of these queries retrieved the 100 most recent Tweets. Tweets from both the sources (@CTNCalgary and @660NewsTraffic) are compared to avoid duplicity as same accidents are tweeted by both. For several representative days, we manually compared the Tweet stream from the API calls to the corresponding Twitter feeds displayed on a browser and made sure that no information was lost due to the API’s restrictions.

Text parsing using regular expressions is performed on the collected Tweets to look for accident reports on the target routes and to eliminate duplicate information. Furthermore, accidents are also assigned to the sub-segments of all target routes being monitored. We manually validated that the Twitter script captured all accident reported for several representative days.

Finally, we augment the commute time and accident information with historical weather data collected automatically from the climate.weather.gc.ca website [13]. Informal insights from Calgary commuters indicated that significant snowfall on a particular day and significant snow and ice on the ground due to previous precipitation activity are better predictors of traffic woes than the temperature. For example, it is not uncommon to witness smooth flow of traffic even when the temperature is -25 Celsius provided there is no precipitation and the roads are clear. As a result, we focus on these two metrics in this paper. Results presented in Section IV vindicate the choice of these metrics since they correlate well with the number of accidents and commute times.

B. Clustering Analysis

We now describe the methodology used to analyze the behavior of Deerfoot Trail. We consider 153 days of data spanning the period September 2013 to February 2014. Since it covers both fall and winter, the data captures diverse scenarios with respect to weather conditions. As shown in Figure 2, we first classified the data into “Good weather days”, i.e., days with no falling snow or snow on the ground, and “Bad weather days”, i.e., days with one or both of falling snow and snow on the ground. For each of these categories, we analyze days with accidents separately from days without accidents. For days with accidents, we apply k-means clustering [3] to ascertain a small number of unique commute time patterns.

For k-means clustering, each day in our dataset was represented by a 19-element feature vector. The elements of this vector are the commute times collected by our system at various times of the day for the entire 50 KM stretch of Deerfoot Trail. We used the Weka toolset [14] with the default k-means settings. The centroid of a cluster reported by Weka is a 19-element vector whose elements represent the average commute times at the various times of the day for all days included within that cluster.

We followed the approach presented by Menasce et al. to determine the appropriate number of clusters for a given dataset [15]. This approach uses $\beta_{var}$ the ratio of the variance of mean intra-cluster distances to the variance of mean inter-cluster distances to decide on the appropriate k-value, i.e., number of clusters. For good quality clustering, intra-cluster distance, i.e., the distance of feature vectors within a cluster from their centroid, must be low while inter-cluster distance, i.e., the distance of the centroid of one cluster to the centroid of another cluster, must be high. Therefore, a lower $\beta_{var}$ value indicates better clustering results. Clustering exercises with progressively higher k-values are carried out till the $\beta_{var}$ value shows no appreciable decrease or starts to increase. The k-value which caused the least $\beta_{var}$ is then chosen.
As shown in Figure 1, by applying this technique we identified 3 clusters for good weather days with accidents (C2, C3, and C4) and 2 clusters for bad weather days with accidents (C5 and C6). On closer analysis, bad weather days without accidents and good weather days without accidents contained only weekends and holidays and exhibited similar commute time patterns. Therefore, they are represented by a single cluster (C1) as the traffic pattern in these kinds of days is very similar.

IV. RESULTS & DISCUSSION

Due to space constraints, we only discuss the north-to-south traffic on Deerfoot Trail. We first describe some general trends from the data before discussing characteristics of the individual clusters.

From Figure 1, about 65% of the days in our observation period had bad weather. This is due to the very short fall and very long winter season in Calgary. Surprisingly, the likelihood of observing at least one accident on Deerfoot Trail is about 85% for both bad weather days and good weather days. However, closer inspection revealed that the average number of accidents per day is 1.8 times higher on bad weather days than on good weather days.

Analysis of the sub-segment information provided several interesting observations on the accidents. More than 34% of the total accidents reported in the Tweets were concentrated on a 4 KM stretch covering the 16th Avenue N, Memorial Drive, and 17th Avenue S interchanges. This sub-segment recorded higher number of accidents than any other sub-segment. Based on the speed limit of the highway, it should only take about 3 minutes to travel this sub-segment. However, the ICT commute time data indicates that commute times in this sub-segment went as high as 66 minutes at 3 PM on 23rd December 2013 when it was snowing and there was 18 cm of snow on the ground.

We next focus our attention on the clusters shown in Figure 1 and consider first the cluster C1 encompassing days with no accidents. Figure 2 shows the commute time trends for this upper, and the day that diverges least from the centroid, i.e., lower.

Figure 2 shows that weather does not seem to impact commute times significantly on weekends and holidays when there are no accidents. There is very little variability in commute times among different days as well as across different times of the day. The maximum commute times for days in this cluster varies from 42 minutes, observed on a day with bad weather, to 33 minutes, on a good weather day.

We next focus on good weather days with accidents. As shown in Figure 1, clustering identified three distinct clusters namely, weekends and holidays (C2), workdays with moderate number of accidents (C3), and workdays with large number of accidents (C4). Holidays only had an average of 1.8 accidents per day. Comparing the centroids of Figures 2 and 3, these accidents increase the commute time by up to 35% when compared to holidays without accidents. From Figure 3, one
can observe that commute times are longer in the evening pointing to more accidents in the evening.

Figure 4 shows traffic patterns for cluster C3, i.e., workdays with moderate number of accidents in good weather conditions. From Figure 1, this cluster represents the most likely pattern for good weather days and contains more than 50% of all good weather days. The number of accidents per day ranges from 3 to 6 with an average of 4.5. Figure 4 shows that the morning rush hours are from about 7:30 AM to 8:30 AM while the evening rush hours are from 3:30 PM to 5:30 PM. As with holidays, closer analysis showed that more accidents happen during evenings, which might explain why the evening rush hour period is longer. Figure 5 shows the traffic pattern for cluster C4, i.e., good weather days with large number of accidents. The number of accidents per day in this cluster range from 7 to 13 with an average of 8.2. Comparing the centroids of Figures 4 and 5, the magnitude of rush hour commute times and the durations of the rush hours increase significantly as the number of accidents increase. From Figure 5, the maximum commute time observed for this cluster was 1 hour and 46 minutes and evening rush hours are worse than morning rush hours.

Closer analysis of sub-segment commute times showed that for both C3 and C4 evening rush hour commute times in the stretch leaving the city center were much higher than the morning rush hour commute times in the stretch leading into the city center. This suggests that the evening traffic volume out of the city center is higher than the morning traffic volume into the city center. This evening traffic represents residents living south of the city center getting back to their homes. These results confirm the longstanding intuition among Calgary residents that there is a lack of good alternatives to Deerfoot Trail for residents in the south.

We now focus our attention on bad weather days with accidents. About 66% of bad weather days were classified into a single cluster C5 as shown in Figure 1. Figure 6 shows the patterns for this cluster. The number of accidents per day range from 1 to 7 for these days with an average of 3.9. Comparing the centroid of Figure 6 with that of Figure 4 which represents the good weather day cluster with comparable average number of accidents, the maximum evening rush hour commute increases by 15 minutes due to bad weather.

Finally, Figure 7 shows the cluster C6 that contains bad weather, accident days with unusual patterns where commute time peaks happen outside of the usual peak hours. About 20% of the bad weather days fall into this cluster. The number of accidents per day varies from 8 to 16 with an average of 9.6. The upper curve of this cluster represents December 2, 2013 when Calgary was hit by an extreme snow blizzard. Commute
times were as high as 2 hours and 4 minutes with 16 accidents reported just on the north to south segment of Deerfoot Trail alone.

From Figure 1, clusters C3 and C5 contain about 62% of the days in our dataset. A noteworthy feature of these clusters is that the days in these clusters have well-defined patterns and are very close to their respective centroids. This suggests that clustering based commute time prediction models can be very effective. We intend to explore such models as future work.

V. LIMITATIONS

Our methodology relies on the ICT commute time estimates from the Google Maps website. While our results seem to conform to real traffic trends in Calgary, a more rigorous evaluation is required to validate the accuracy of this metric. Our system also only focuses on accidents and weather as factors that influence commute time. As outlined in Section I, there are other factors such as sun position, special events, and lane closures that can impact commute times. While some of these factors, e.g., lane closures, can be identified by enhancing our Twitter scripts’ text parsing capabilities, others require us to identify other Web sources, e.g., hockey game schedules from nhl.com. However, there may be factors such as traffic volumes that are currently not reported by any Web service.

As mentioned previously, the identification of sub-segments is currently done manually. Automating this would increase the flexibility of our system towards handling more routes. Our system’s monitoring capabilities are also limited by query restrictions of Google Maps API and Twitter Search API.

Our analysis ignores fine-grained characterization of weather conditions based on factors such as temperature, visibility, and precipitation amounts. This limitation can be overcome easily since the climate.weather.gc.ca website [13] provides many of these metrics. We also did not focus on evaluating the effectiveness of other clustering techniques or on alternative configurations of k-means, e.g., sensitivity to various distance measures. Finally, our analysis could also benefit from a larger dataset.

VI. CONCLUSION AND FUTURE WORK

This paper proposed a low-cost system for supporting historical analysis of vehicle commute times on roadways. The system does not require dedicated roadside sensors and associated infrastructure. Instead, it collects commute time estimates for any given route and its sub-segments from the Google Maps website over a long period of time. Furthermore, it overlays traffic accident information from the Twitter social network and weather information onto the collected commute times. We show that such a system is able to support analyses that characterize commute time patterns and their dependency on factors such as weather, accidents, and time of the day.

Future work will focus on first addressing the limitations outlined in Section V. Furthermore, the scope of the analysis could be expanded to cover multiple roads to study how traffic on a particular road affects traffic on other roads connected to it.

ACKNOWLEDGMENT

This work has been partially supported by a grant from Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES