Using Social Media to Predict Traffic Flow under Special Event Conditions

by

Ming Ni

07/26/2013

A thesis submitted to the
Faculty of the Graduate School of the
State University of New York at Buffalo
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Industrial and Systems Engineering

University at Buffalo, SUNY, USA
Acknowledgment

I would like to give my sincere gratitude to my advisor, Dr. Qing He, for your invaluable guidance and help, without which I would not have been able to accomplish this. I would like give many thanks to Dr. Jing Gao for your constant guidance and support.

Also a special thank you to my parents. Their constant support and encouragement is always my strength.
Abstract

Social media is a great resource of user-generated contents. Public attention, opinion and hot topics can be captured in the social media, which provides the ability to predict human-related events. Since social media can be retrieved in real time with no building cost and no maintenance cost, traffic operation authorities probably identify the social media data as another type of sensor for traffic demand. In this thesis, we aim to use social media information to assist traffic flow prediction under special event conditions. Specifically, a short-term traffic flow prediction model, incorporated with tweet features, is developed to forecast the incoming traffic flow prior to sport game events. Both tweet rate features and semantic features are included in the prediction model. We examine and compare the performance of four regression methods, respectively autoregressive model, neural networks model, support vector regression, and k-nearest neighbor, with and without social media features. To the end, we show the benefit gained by including social media information in the prediction model and its computational efficiency for potential practical applications.

Keywords: social media, Twitter, traffic flow prediction
## Contents

Acknowledgment ................................................................. iii

Abstract..................................................................................... iv

Chapter 1 Introduction and Motivation ........................................ 1

1.1 Thesis Scope ........................................................................... 1

1.2 Motivation............................................................................... 1

Chapter 2 Literature review .......................................................... 4

2.1 Social Media ........................................................................... 4

2.2 Traffic Prediction ..................................................................... 5

Chapter 3 Prediction Analysis ......................................................... 7

3.1 Data Description ...................................................................... 7

3.2 Data Pre-processing ................................................................. 12

3.2.1 Game and Non-game Day Comparison ............................... 12

3.2.2 Tweet Semantics Extraction ............................................. 17

3.3 Prediction Analysis and Modeling ............................................. 18

3.3.1 Prediction Features Selection ............................................. 18

3.3.2 Prediction Models ................................................................. 22

Chapter 4 Performance Results ....................................................... 24

Chapter 5 Conclusion and Future Work ........................................ 27

Bibliography ............................................................................... 29
List of Figures

Figure 1 The flow chart of the prediction analysis .............................................. 7
Figure 2 Twitter data collected location .................................................................. 9
Figure 3 Detectors layouts around Oracle Arena and O.co Coliseum ............... 12
Figure 4 Comparisons of tweet rates on game days and non-game days .............. 14
Figure 5 Hourly traffic flow for detectors ............................................................... 15
Figure 6 The de-trended prior-game traffic flow .................................................... 15
Figure 7 Correlation analysis between tweet features and game traffic
incremental ............................................................................................................. 19
Figure 8 The MAPE and RSME for each traffic detector ...................................... 25
List of Tables

Table 1 Characteristics of collected tweets.......................................................... 9
Table 2 The home games of the Athletics and Warrior from February to May.. 11
Table 3 Keywords to select game-related tweets.............................................. 13
Table 4 Features for Tweet Rates..................................................................... 13
Table 5 ANOVA test results for factor “game day”......................................... 16
Table 6 Selected tweet features for each detector with different $\lambda$.......... 21
Table 7 Average of MAPE and RMSE for SVR............................................... 26
Chapter 1
Introduction and Motivation

This chapter is divided into two parts. The first part is devoted to define the scope of the thesis in order to give overall view of the work. The second part discusses the motivation and initiatives to give us the power to do the research for the last six months.

1.1 Thesis Scope

The thesis is initiated by the potentially existing relationship between special event caused traffic and social media user-generated contents.

The scope of traffic in this thesis is constrained in the freeway, according to the data accessibility and the big influence of traffic. We focus on fusing social media data with event traffic data, in order to predict the dynamics of prior-event traffic. The possible linkages of social media data and the special event-caused traffic will be analyzed and recognized.

The scope of the social media data is restrained to the Tweets, which created, shared and modified by the Twitter users. Twitter Inc. provides relatively convenient and open free application interface for people to collect and convert to social media database comparing with other social media services. Also the Twitter, one of the extremely popular services in United States, has a vast number of users, 200 million active monthly users in December 2012. The above two reasons guarantee the quality and quantity of the potential tweet data.

1.2 Motivation

This thesis aims to use social media data to predict traffic flow under special event conditions. The motivation for this research on the relationship between social media and traffic flow is to provide effective models and indicators for traffic prediction in
the age that has the largely increased users in social media and vast user-generated contents. By understanding the potential linkage with social media and freeway traffic flow, future researchers can probably identify the social media data as another type sensor for traffic demand, which is updated in real time with no building cost and no need for maintenances.

Social media has become an indicator of modern people and lifestyle in the Internet virtual community. Vast of user-generated contents strengthen linkage and interaction between each individuals within the community circle, also provide large amount of information related to various area. Examples include Facebook, Twitter, YouTube, Google Plus and Wikipedia.

Through the booming of smartphone and cellular networks in the last 5 years, it is easier to use and reach social media than ever before. More and more users now are able to post and view contents of the virtual circle simply using their handset devices. The trend of easy accessing social media will continuously grow with the development and commercialization of wearable computer devices, like Google Glass and other smart watches.

On the other hand, Traffic congestion is one of the interesting and long-lasting problems in the world. For centuries, people worked to invent better tools and vehicles to help them to travel and move goods. For recent decades, the intrigued traffic network has been established around the world, but coming with both day-to-day recurrent traffic congestion and event-caused non-recurrent traffic congestion, which significantly affect the quality of life and impact the U.S. economy. Among different traffic mitigation counter-measures, traffic flow prediction alerts traffic congestion beforehand and facilitates pro-active traffic management and control. Therefore, traffic flow prediction is one of the key methods and the foundation for alleviating traffic congestion.
Motivated by the potentially great value of the social media and the difficulty of traffic flow prediction, in this thesis, we intend to build relationship between traffic data and social media data. Specially, the following question have been answered: Will social media help to predict prior-event traffic?

Furthermore, the purpose of research is to demonstrate one way to uncover the potential great value of big data generated by the social media. In this case, it can be an effective indicator of freeway traffic performance. This possibility may transfer to other branches of transportation, such as transit scheduling and operations, as well post-disaster recovery. In this way, our work can assist future researchers in their initiatives of social media data mining.
Chapter 2
Literature review

The main focus of this thesis is to investigate the potential relationship between freeway traffic flow and social media user-generated contents. This chapter investigates the existing social media aided researches, characteristics of the possible prediction made by the social media, and the prior short-term traffic prediction models under special conditions. The review generally comes from two perspectives, namely social media and traffic prediction.

2.1 Social Media

Although the social media includes a variety of web services, the previous studies generally fall into two parts: social relations and user-generated contents (Yu and Kak 2012). Social relations reflect the virtually social network and connections between people on the web. The user-generated contents are created and published by the people, which intend to strengthen linkage and interaction between each individual within the community circle. In this thesis, we focus on the previous studies with the user-generated contents.

Positive or negative emotions were able be extracted from Twitter data about certain company brands and products (Jansen et al. 2009), which can be an auxiliary survey data applying to the company marketing strategy. Public opinion of election in US was studied by some researches (Tumasjan et al. 2010, Balasubramanyan, Routledge, and Smith 2010), effectively using the social data as a public survey to predict the election results with low cost comparing to the traditional telephone survey. As a different idea, a model was built by the rate of Twitter messages, related to movie, to predict movie box-office revenue and confirmed that the Twitter feeds can be effective indicators of real-world performance (Asur and Huberman 2010). An evaluation system about stock market was proposed to collect public mood and
sentiment form Twitter service as an economic indicator (Bollen, Mao, and Zeng 2011). Other research also used social media to predict stock market indicator by analyzing Twitter posts. It is concluded that the correlation between stock market and social media indeed existed by examining emotional Twitter posts and market index such as Dow Jones, NASDAQ and S&P 500 (Zhang, Fuehres, and Gloor 2011).

One research (Yu and Kak 2012) identified three characteristics of event when social media data can be treated as a good predictor. The three characteristics are, respectively, human related event, masses of people involved, and the event should be easy to be talked in public.

### 2.2 Traffic Prediction

Since people intended to publish on social media corresponding to the non-recurrent traffic conditions, such as traffic special events, we focus on the literatures related to the traffic prediction under atypical conditions.

Neural Network was used to build the traffic volume prediction model which was based on the time-series data (Yasdi 1999). In a similar manner, a supervised statistical learning technique called Online Support Vector machine for Regression, or OL-SVR, was applied for the prediction of short-term freeway traffic flow under both typical and atypical conditions (Castro-Neto et al. 2009). Guo et al. (Guo, Krishnan, and Polak 2012) pre-processed traffic data using Singular Spectrum Analysis, and utilized k-nearest neighbor method to predict traffic. Since pre-processing step reduced the noisy sensor inputs, this model can be used under normal and special conditions. An Online boosting nonparametric regression (OBNR) model also was used to perform traffic prediction, which consists of two major parts, respectively, the base part and the boosting part. The base part was constructed under normal conditions, while the gradient boosting part undertook special conditions into account (Wu et al. 2012).
From those previous studies for traffic prediction, four models frequently used as experimental or reference methods, respectively, autoregressive model, neural networks, support vector regression, k-nearest neighbor, to build the models under both typical and atypical conditions. In the next chapter, we will utilize those four good and proved methods to construct the traffic prediction models.

Because there are very few references about the linkage of traffic and social media, the most challenge for this thesis is to find the right use case to further investigate if social media can assist traffic analytics. From the previous papers about social media in this chapter, we can see social media may do prediction or examination about events, which the ideas of this research come from. However, in spite of the constraint of idea, there are several possibly feasible linkages existing. And the only way to investigate their feasibility is to build and check the models. For instance, examining traffic incidents by social media is one of the possible ideas. We used the social media data related to the incidents temporally and spatially to build the model. But lack of the adequate number of related tweets on the freeway became an obstacle to continue the study. In sum, it is essential but time consuming to locate the right study topics for this thesis.
Chapter 3
Prediction Analysis

The framework of prediction analysis is presented as Figure 1. Starting with data preprocessing, both Twitter and traffic features are introduced and extracted by the following four components, traffic data de-trending, game traffic impacted detector identification, game tweet extraction and aggregation, and tweet semantics extraction.

![Figure 1 The flow chart of the prediction analysis](image)

3.1 Data Description

In this thesis, traffic data were obtained from the Caltrans Performance Measurement System (PeMS)*. PeMS is a system designed to maintain California freeway traffic data and compute annual congestion for facilities with surveillance systems in place, typically loop detectors spaced approximately 0.5 mile apart on each freeway lane (Choe, Skabardonis, and Varaiya 2013). The analysis uses 1-hour aggregated volume data, collected in four months from February 1, 2013 to May 30, 2013.

Some of the detectors may miss or report invalid data in the practice. In order to compensate the missing or incorrect data samples, the diagnostics algorithm and imputation regression models developed in (Chen et al. 2003) were applied to detect

*http://pems.dot.ca.gov/
the bad detectors and fill the missing value. This method generated total 12,242,699 entries of hourly traffic flow records in the database.

We used the following linear model to compensate the missing data from neighbor loop detectors:

\[
q_i(t) = \alpha_0(i, j) + \alpha_i(i, j)q_j(t) + \text{noise}
\]

For each pair of neighboring detectors \((i, j)\), \(q\) here stands for the traffic flow at timestamp \(t\). And the parameters \(\alpha_0(i, j)\) and \(\alpha_i(i, j)\) are estimated using maximum-likelihood estimation (MLE) by the valid data \(t \in (1, n)\) in the four months.

\[
\alpha_0(i, j), \alpha_i(i, j) = \arg \max_{\alpha_0, \alpha_i} \left\{ \frac{1}{n} \sum_{i=1}^{n} [q_i(t) - \alpha_0 - \alpha_i q_j(t)]^2 \right\}
\]

Twitter data was collected by the same spatial and temporal window accordingly, through the Twitter Streaming API † with geo-location filter. We use cURL ‡, transferring data computer software, to grab the data from the Twitter Streaming API. cURL works day and night to collect the real-time tweets came from Bay area. The original file is JavaScript Object Notation (JSON) format. We used Python § to convert the row files to the plain text format and combine the files together as the tweet database in the following analysis.

Table 1 shows the characteristics of collected tweets in the database. In fact, there are more characteristics for a tweet. But in this thesis, the below nine is concerned and extracted rather than other features.

† https://dev.twitter.com/
‡ http://curl.haxx.se/
§ http://www.python.org/
Table 1 Characteristics of collected tweets

<table>
<thead>
<tr>
<th>Parameters of a Collected Tweet</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Who sent the tweet</td>
</tr>
<tr>
<td>screen_name</td>
<td>The user ID of Twitter</td>
</tr>
<tr>
<td>timestamp</td>
<td>When the tweet posted</td>
</tr>
<tr>
<td>texts</td>
<td>The content of the post</td>
</tr>
<tr>
<td>Latitude and longitude</td>
<td>The GPS location</td>
</tr>
<tr>
<td>Location</td>
<td>The name of the location</td>
</tr>
<tr>
<td>Mentions (@someone)</td>
<td>Whether this post @ other user</td>
</tr>
<tr>
<td>Hashtags (#topic)</td>
<td>How many topics of the post</td>
</tr>
<tr>
<td>URL</td>
<td>Whether this post has link</td>
</tr>
</tbody>
</table>

Right after tweets collected, the filtration of spam and commercial tweets is implemented by the list of Twitter users. This results in a total number of 5,444,527 valid tweets.

Figure 2 shows the bounding box to grab the tweets and to consist of the main network of freeways with the bay area. Also it shows that part of collected tweets display on the map, one dot represents one tweet.

(a) The bounding box to collect tweets  (b) Collected tweets display on the map

Figure 2 Twitter data collected location
As a basic introduction of tweets, each user of Twitter can be subscribed by other users known as followers. The message posts or status updates, called as tweets, is able to contain 140 characters or less, which typically is one kind of user-generated contents. Tweets can act as vehicles of text, image, URL, location service, video and so on. Published Tweets are shown in the user’s profile page as well as followers’ home pages. The "@" sign followed by a username is used for mentioning or replying to other users. Hashtag of tweets are the words or phrases prefixed with a "#" sign, which is used to group posts together or indicated to a topic.

Note that there are some limitations of the Twitter Streaming API with geo-location filter. Firstly, the lots of users do not have location enabled smart phone or location in their profile so that their tweet will not be collected. Secondly, the location matching of Twitter is not entirely accurate. In addition, the Twitter data collection system sometime experienced unexpectedly short-term outage during the four months. Although having those handicaps, thankfully to the hotspot – special events, the characteristics of the possible prediction made by the social media data, we assume, still can be fulfilled by the collected large part of tweets. The characteristics are, respectively, human related event, masses of people involved, and the events should be easy to be talked in public.

The sport events are considered as good venues to perform both traffic and social media analysis, since it can be observed by the tweet posts, namely quantity and semantics. In addition, the traffic volume is dramatically influenced by the local sport events due to its popularity.

In the study, we specially consider game traffic impact on I880 near Oracle Arena and O.co Coliseum in Oakland, California. Oracle Arena and O.co Coliseum was home to the Oakland Athletics (Athletics) and Golden State Warriors (Warriors) in 2013 game seasons. Oakland Athletics are a Major League Baseball team and Golden State Warriors are an American professional basketball team. There were 51 home games
of the Athletics and Warrior from February to May in 2013, the details is in Table 2. Oracle Arena and O.co Coliseum is located right besides Interstate Highway 880 (I-880). Six traffic detectors, located before the exits of I-880 to the entrance of the Oracle Arena, were chosen to analyze incoming prior game traffic, shown as Figure 3.

Table 2 The home games of the Athletics and Warrior from February to May

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Team</th>
<th>Timestamp</th>
<th>Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/2/2013 19:30</td>
<td>G.S. Warriors</td>
<td>3/30/2013 13:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>2/12/2013 19:30</td>
<td>G.S. Warriors</td>
<td>4/1/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>2/20/2013 19:30</td>
<td>G.S. Warriors</td>
<td>4/2/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>3/6/2013 19:30</td>
<td>G.S. Warriors</td>
<td>4/12/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>3/13/2013 19:30</td>
<td>G.S. Warriors</td>
<td>4/16/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>4/7/2013 17:00</td>
<td>G.S. Warriors</td>
<td>4/30/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>4/15/2013 19:30</td>
<td>G.S. Warriors</td>
<td>5/14/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>5/2/2013 19:30</td>
<td>G.S. Warriors</td>
<td>5/18/2013 18:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>5/10/2013 19:30</td>
<td>G.S. Warriors</td>
<td>5/19/2013 13:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td>5/16/2013 19:30</td>
<td>G.S. Warriors</td>
<td>5/28/2013 19:05</td>
<td>O. Athletics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5/31/2013 19:05</td>
<td>O. Athletics</td>
</tr>
</tbody>
</table>
3.2 Data Pre-processing

In this section we study the tweet and traffic data sets in order to obtain the potential features in the prediction modeling.

3.2.1 Game and Non-game Day Comparison

In order to see the differentiation between game day and non-game day for tweet data as well as traffic flow, the raw data have been pre-processed as the usable features. And validation of the data will be analyzed in this part.

Twitter

The sport games, like Athletics’ and Warriors’, were considerably interested topics in the Twitter, especially among the Twitter users in the Bay area in order to comment and support their home teams. In order to identify the tweets which are relevant to
each game and each team, the appropriate keywords, shown as Table 3, were used to select the game-related tweets with the help of disambiguation checks.

Table 3 Keywords to select game-related tweets

<table>
<thead>
<tr>
<th>Gold State Warriors</th>
<th>Oakland Athletics</th>
</tr>
</thead>
<tbody>
<tr>
<td>· letsgowarriors</td>
<td>· greencollar</td>
</tr>
<tr>
<td>· warrior</td>
<td>· letsgooakland</td>
</tr>
<tr>
<td>· dubnation</td>
<td>· athletics</td>
</tr>
<tr>
<td>· letsgodubs</td>
<td>· astalk</td>
</tr>
<tr>
<td>· warriorsground</td>
<td></td>
</tr>
<tr>
<td>· warriorsgame</td>
<td></td>
</tr>
</tbody>
</table>

The game-relevant tweets were aggregated into hourly time series in order to be compatible with Traffic data. For each hour in the 4 months, we extracted 5 features of tweet rates, shown as Table 4.

Table 4 Features for Tweet Rates

<table>
<thead>
<tr>
<th>Features for Tweet rates</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_of_tws</td>
<td>Number of tweets related to game in one hour</td>
</tr>
<tr>
<td>no_of_users</td>
<td>Number of independent users sent those tweets</td>
</tr>
<tr>
<td>no_of_hashtags</td>
<td>Number of hashtags in all those tweets</td>
</tr>
<tr>
<td>no_of_tws_mention</td>
<td>Number of tweets mentioned other twitter user</td>
</tr>
<tr>
<td>no_of_tws_urls</td>
<td>Number of tweets which has URLs</td>
</tr>
</tbody>
</table>

Note the non-game days stand for the days which don’t carry on either home game or away game, while game days denotes only for the days carrying on home game at Oracle Arena or O.co Coliseum.

Figure 4 shows the number of tweets per hour for Golden State Warriors at their game days and non-game days. One can see that there are much more tweets on the game day than that on the non-game day. Therefore, we can use the tweet rate series on the same day to indicate the total public attention for this game.
First of all, in order to eliminate the effects of day-to-day traffic fluctuations, we detrended the traffic flow by subtracting the average hour-of-day and day-of-week traffic volumes. For every record of traffic flow, the below equation was used to reduce the periodic variance so that we can only concentrate on the increment of the traffic flow.

\[
\Delta y_{wt} = y_{wt} - \overline{y}_{wt}
\]

Where \( y \) stands for the traffic flow; \( w \) indicates it is weekday or weekend; \( t \) denotes the hour of day; \( \overline{y}_{wt} \) is the average time-of-day and day-of-week traffic flow.
The original and de-trended time-of-day patterns of traffic flow are demonstrated in Figure 5 for both weekday and weekend.

(a) Hourly traffic flow for weekdays
(b) Hourly traffic flow for weekends
(c) De-trended hourly traffic flow for weekdays
(d) De-trended hourly traffic flow for weekends

Figure 5 Hourly traffic flow for detectors

Figure 6 The de-trended prior-game traffic flow
The de-trended prior game traffic is plotted in Figure 6, which takes the last 2 hours before the game event starting into account. The majority of de-trended game traffic series is greater than zero, which indicates more traffic demand on the game day.

In order to test the impact of game on traffic data, a single-factor analysis of variance (ANOVA) was used to differentiate between two populations of data sets through factor “game day”. One dataset contains the de-trended traffic flows within 4 hours before the game starting time. The other includes the binary game indicator during the same time on every non-game day.

Table 5 ANOVA test results for factor “game day”

<table>
<thead>
<tr>
<th>Detector</th>
<th>P-value</th>
<th>Significant or not</th>
</tr>
</thead>
<tbody>
<tr>
<td>400498</td>
<td>0.0064</td>
<td>YES</td>
</tr>
<tr>
<td>400955</td>
<td>1.24e-13</td>
<td>YES</td>
</tr>
<tr>
<td>400190</td>
<td>6.3e-07</td>
<td>YES</td>
</tr>
<tr>
<td>400956</td>
<td>0.029</td>
<td>NO</td>
</tr>
<tr>
<td>400360</td>
<td>4.52e-13</td>
<td>YES</td>
</tr>
<tr>
<td>400333</td>
<td>7.45e-13</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 5 presents the results from ANOVA analysis. As one can see, sport games generate significantly impact on the traffic volume of most surrounding detectors on game days. Note that the traffic flow at detector 400956 does not demonstrate significant differences caused by game events, since the detector is located at downstream of the exit towards Oracle Arena. In addition, the distance between detector 400333 and Oracle Arena is much larger comparing with other detectors. For the purpose of simplification, these 2 traffic detectors are treated as only the neighboring detectors rather than the analytical targets to be predicted in the following analysis.
3.2.2 Tweet Semantics Extraction

One of important differences between tweets and other types of data is that tweets consist of rich sentiments. Utilizing this information of tweet data, sentiment extraction is necessarily introduced to the analysis.

Sentiment analysis is an important part in machine learning. In here, sentiment results are treated as potential tweet features. In this thesis, the sentiment analysis is performed based on an R (R Development Core Team 2009) package, called “tm.plugin.sentiment”, which provides the functions for natural language sentiment processing. The words of text in each Tweet are labeled as positive, negative and neutral references. Before conducting the sentiment analysis, the text part of tweets was initially preprocessed by the following steps.

1. Make each letter lowercase,
2. Remove punctuation and stopwords.
3. Replace the abbreviations from the formal English words.

Our study implemented a unified letter transformation approach (Liu et al. 2011) to normalize the Tweet post, to replace the non-standard tokens to the Standard English words. This method tends to alleviate the biased effect of abnormal text on the sentiment analysis.

In order to quantify the sentiments of tweets, a sentiment scores system was implemented and based on the Lydia/TextMap system (Godbole, Srinivasaiah, and Skiena 2007) (Agarwal et al. 2011). There are 5 measures listed as follows.

- polarity \( (p - n / p + n) \): difference of positive and negative sentiment references / total number of sentiment references
- subjectivity \( (p + n / N) \): total number of sentiment references / total number of references
Prediction Analysis

- pos_refs_per_ref \( (p / N) \) : total number of positive sentiment references / total number of references
- neg_refs_per_ref \( (n / N) \) : total number of negative sentiment references / total number of references
- senti_diffs_per_ref \( (p - n / N) \) : difference of positive and negative sentiment references / total number of references

Above 5 measures were calculated for every tweet post.

The polarity is the feature to measure the bias (positive or negative) of the text, while the subjectivity is the feature to measure the rate of bias reference in the text. The pos_refs_per_ref and negRefs_per_ref are the rate at which positive and negative references occur in the text. senti_diffs_per_ref is a combination of polarity and subjectivity: it determines the bias of the text in relation to the size of the text.

3.3 Prediction Analysis and Modeling

After the traffic and tweet data have been processed, there are many usable features of those two, respectively the traffic flowing of neighboring detectors, and 5 tweet rate features plus 5 tweet sentiment score features. This section, two kinds of features are combined together to build the short-term traffic prediction model under the special event conditions.

3.3.1 Prediction Features Selection

Because there are many potential features of traffic and tweet at hand, it is very necessary to select the right significant features to build a parsimonious prediction model.

It varies that the correlations between tweets features and traffic flow for each detectors, as Figure 7 shows. Therefore, the features selection is an essential step and very significantly affects the performance of the prediction modals.
Figure 7 Correlation analysis between tweet features and game traffic incremental

Where the “W” stands for games of Golden State Warriors. “A” stands for games of Oakland Athletics. “AW” denotes the game time happened to have two teams together at the Oracle Arena. “ALL” is for correlation for the all of games together.

Those traffic and tweet features were combined into an optimal matrix for each traffic detector. Here, we used the least squares optimization with L1-norm regularization (Schmidt) to solve this problem.
\[
\min \sum_{k=1}^{K} \sum_{t=1}^{T} \left( \Delta y_{k}^{g-t} - \alpha - \sum_{m=1}^{M} \sum_{l=1}^{L} \beta_{ml} \Delta y_{km}^{t-t-l} - \sum_{p=1}^{P} \chi_{kp} T_{g-t-r} V_{p} \right)^{2} + \lambda \sum_{p=1}^{P} |V_{p}|
\]

Where \(K\) is the number of game; \(T\) stands for the time steps before game start to be predicted and \(T_{g}\) indicates the game starting time; \(M\) is total number of neighboring detectors; \(L\) denotes the number of time lags of the neighboring detectors; \(P\) is the number of tweet features; \(r\) represents the number of time lags of tweets relative to traffic data; \(\beta_{ml}\) is the coefficients for traffic variables; \(V_{p}\) is the coefficients for tweet variables; \(\alpha\) indicates the intercept of the model. \(\lambda > 0\) is the regularization parameter that balances the L1 regularization term and least square term. One may replace \(|V_{p}|\) by two additional positive variables \(V_{p}^{+}\) and \(V_{p}^{-}\), as following:

\[
V_{p} = V_{p}^{+} - V_{p}^{-}
\]

\[
V_{p}^{+} \geq 0, \; V_{p}^{-} \geq 0
\]

\[
|V_{p}| = V_{p}^{+} + V_{p}^{-}
\]

By applying different values of \(\lambda\), we solved the above optimization problem by CPLEX. Appropriate traffic and tweet features were selected with non-zero coefficients. For the traffic features, the traffic flow variables of neighboring detectors were always chosen by the optimization model. For the tweet features, every target detector had different combination of tweet rate features and tweet sentiment features, shown as Table 6. In the practice, we selected tweet features when \(\lambda = 6000\), but also considered the effect of different lambda values.
## Table 6: Selected tweet features for each detector with different $\lambda$

<table>
<thead>
<tr>
<th>$\lambda = 2000$</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$\lambda = 4000$</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$\lambda = 6000$</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
<th>Traffic Flow</th>
<th>Selected Tweet Features</th>
</tr>
</thead>
</table>

21
In the Table 6, the independent variables are the volumes of traffic flow (tf) in last 1 hour prior to the games. Because there are 2 sport teams, in here, “w” stands for Golden State Warriors and “a” stands for Oakland Athletics. “sum4” means that the tweet feature is aggregated in last 4 hours before the predicting time.

### 3.3.2 Prediction Models

There were many existing prediction approaches for short-term traffic flow, although most of them focus on traffic features only. To compare with the performance under different methods, we examine the following four popular methods to predict short-term event traffic (Smith and Demetsky 1994, Smola and Schölkopf 2004, R. K. Oswald, William T. Scherer, and Brian Lee Smith 2001):

- Autoregressive Model (AR)
- Neural Networks (NN)
- Support Vector Regression (SVR)
- K-Nearest Neighbor (KNN)

In order to test the performance of social media features in short-term traffic prediction, two models is constructed for every target detector under each method. The model 1 is typically based on the traffic features only, while the model 2 depends...
on both traffic features and tweet features. By comparing the results between Model 1 and Model 2, we are able to tell whether the social media features improve the overall performance of the accuracy of traffic prediction.

- Model 1: Target traffic flow~ (Neighboring Traffic flow)
- Model 2: Target traffic flow~ (Neighboring Traffic flow) + (Tweet Features)

The implementation of those methods is as following. AR model basically implements the linear regression model. The architecture of NN model consists of the input layer, a single hidden layer with 6 neurons and 1 output neuron. In terms of SVR, we used typical linear kernel and epsilon-regression. KNN method takes the straight average of the 10 nearest points in the training set in Euclidean distance to obtain the prediction results.
Chapter 4
Performance Results

The prediction performance is evaluated by two measures, namely Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), defined as follows

\[
MAPE = \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{|\Delta y_k^{T_g-t} - \Delta \hat{y}_k^{T_g-t}|}{y_k}
\]

\[
RMSE = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{T} (\Delta y_k^{T_g-t} - \Delta \hat{y}_k^{T_g-t})^2}
\]

Where \( K \) stands for the total number of games; \( \hat{y} \) denotes the estimated value of traffic flow \( y \); \( T \) represents total number of time steps to be predicted before the game starts, and \( T_g \) indicates the game starting time;

Every prediction model is implemented with above four methods. The training data occupies the 70% of entire dataset, while the remaining 30% was treated as test data. We generated 100 instances of training and test data with 100 different random seeds. Every model runs for 100 times. And the performance is evaluated by the average MAPE and RMSE of 100 experiments.

Figure 8 shows the evaluation results of two models in four regression methods for four target detectors. As one can see, although the MAPEs doesn’t show consistent results across four regression methods and four detectors, the RMSEs of model 2 with tweet features always outperform those of model 1 with only traffic features. Therefore, it is very beneficial to incorporate tweet features in the prediction models. In other words, the tweet features indeed improve the results of short-term traffic prediction. More importantly, such performance improvement exists regardless of forecasting methods.
Performance Results

Figure 8 The MAPE and RMSE for each traffic detector.
As indicated from Figure 8, SVR shows overall best results among four methods for game traffic prediction. The average of MAPE and RMSE of SVR are presented in Table 7.

Table 7 Average of MAPE and RMSE for SVR

<table>
<thead>
<tr>
<th></th>
<th>Model 1 Traffic only</th>
<th>Model 2 Traffic Tweet</th>
<th>Error reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>385.5252</td>
<td>311.0429</td>
<td>23.946%</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.050437</td>
<td>0.047146</td>
<td>6.9804%</td>
</tr>
</tbody>
</table>

The percentage of error reduction in Table 7 indicates the promising value of social media in event traffic prediction. Because it is very difficult to predict the short-term traffic flow under abnormal conditions, any improvement is significant for traffic management and operations, especially when using no-cost and real-time social media data.

The computational efficiency of short-term traffic prediction is critical for practical applications, such as Advanced Traveler Information Systems (ATIS). The traffic flow data usually reports every 30 seconds from detectors, which can be used directly in real-time. The social media data comes from the Twitter Stream application interface in real time but it requires additional data processing to get the tweet rate features and tweet sentiment features. Considering the hourly aggregation and keywords filtration to obtain sport game related tweet, it is feasible to prepare the right social media data in order to get the desirable tweet features in an hour. Estimation of processing time for an hour tweet feature extraction is twenty minutes. As a result, the proposed prediction models can be directly applied for real-time practical applications.
Chapter 5
Conclusion and Future Work

In this thesis, social media data is incorporated with traffic flow data for with prior-event traffic prediction. The linkages of tweet data and the sport game traffic is analyzed and recognized to build a short-term traffic flow prediction models.

We identify the traffic flow of neighboring detectors as traffic features; the tweet rates and tweet sentiments as social media features. The traffic features are treated as the typical components in the traffic prediction models, while social media features are treated as the experimental components selectively in the models.

There are many potential tweet features at hand. Least squares optimization with L1-norm regularization to select the right significant features to build a parsimonious prediction model. For target detectors, there are different set of the traffic features and tweet features be selected.

In order to construct effective prediction models for short term traffic under special event condition, four popular forecasting methods that frequently used in previous studies, respectively, autoregressive model, neural networks, support vector regression, and k-nearest neighbor, are implemented to build models.

We then find that the prediction results with both traffic and tweet features outperformed those with only traditional traffic features across different detectors, no matter what kind forecasting models in those four forecasting methods. With support vector regression, average MAPE and RMSE was reduced by 6.98% and 23.95% after including tweet features in the model.

We also argue the efficiency of computation time for social media data processing in practical real-time applications. It is feasible to do traffic and tweet feature extraction in the practice for hourly aggregation. The time cost is estimated.
Another purpose of research is to demonstrate one way to uncover the potential great value of big data generated by the social media. In this case, it can be considered as an effective indicator of public attention and opinions. This possibility may transfer to other branches of transportation, such as travel behavior analysis, transportation related choices, transit scheduling and operations, and so on, which can be treated as other complicated and challenging aspects of social media collaboration with traffic that needs to be addressed.
Bibliography


