An Exploratory Study on the Correlation between Twitter Concentration and Traffic Surge

Zhenhua Zhang
Department of Civil, Structural and Environmental Engineering
The State University of New York, Buffalo, NY 14260
Email: zhenhuaz@buffalo.edu

Ming Ni
Department of Industrial and Systems Engineering
The State University of New York, Buffalo, NY 14260
Email: mingni@buffalo.edu

Qing He
Department of Civil, Structural and Environmental Engineering
Department of Industrial and Systems Engineering
The State University of New York, Buffalo, NY 14260
Email: qinghe@buffalo.edu

Jing Gao
Department of Computer Science and Engineering
The State University of New York, Buffalo, NY 14260
Email: jing@buffalo.edu

Jizhan Gou
Serco Inc., Reston VA 20190
Email: Jizhan.Gou@serco-na.com

Xiaoling Li
Northern Region Operations,
Virginia Department of Transportation, Fairfax, VA 22030
Email: Ling.li@vdot.virginia.gov

Submitted to TRB 95th Annual Meeting for both Presentation and Publication
January 2016, Washington D.C.
To appear in Transportation Research Record, 2016

Word Count: 7460
Abstract and Manuscript Text: 5960
Number of Tables and Figures: 10 (= 2500 words)

1 Corresponding Author
Abstract

Social media receives increasing attentions as a crowdsourced information source in traffic operations and management. The tweets, which are blogged and shared by the broad masses of people, may be associated with some major social activities. These tweets are called “Twitter concentrations” in this paper. The public activities behind Twitter concentrations potentially pose more pressure on traffic network and cause traffic surge within a specified time and location. However, it still remains unknown how closely the Twitter concentration and traffic surge are correlated with each other. Our study fuses a set of tweets and traffic data collected during the whole year of 2014 in North Virginia Region, and mainly investigates the correlation between Twitter concentration and traffic surge in July. The results show the promise and effectiveness of our proposed methods and even provide insights in the causality of the non-recurrent traffic surge.

Keywords: traffic surge; tweet analysis; Twitter concentration; social media

1 Introduction

Road traffic surge aggravates the jammed condition and worsens the level of service of road links. The consequences of traffic surge may vary, including traffic delay, fuel wasting, drivers’ frustration, etc. Some of the traffic surges may be accounted by the recurrent features of traffic patterns such as time-of-day characteristics and weekday-weekend differences. This kind of traffic increase is predictable in most cases, and people living by usually get accustomed to it. Other kinds of traffic surge, which are more unpredictable and hazardous, may correlate with non-recurrent traffic patterns such as road accident, bad weather, malfunction of traffic signals, festival parades, etc. Figure 1 illustrates the differences between recurrent and non-recurrent traffic surge. The non-recurrent traffic patterns and corresponding traffic surge problems, which are caused by major social activities, are the main interests of this paper.

For decades, the traffic surge problem, which potentially causes and even worsens traffic congestion, has been given much attention. The state-of-art studies attempt to unveil the correlation between traffic increase and other variables, and several explanations have been put forward, which are described as follows. Bando et al. (1) found that there exists the congestion that is induced by a small perturbation without any specific origin such as a traffic accident or a traffic signal. Arnott et al. (2) argued that cars cruising for parking add to traffic congestion. Duranton et al. (3) showed no evidence that the provision of public transportation affects vehicle kilometers traveled (VKT). Further, they proved that increased provision of roads or public transit is unlikely to relieve congestion. Anderson (4) concludes that the cessation of transit service may increase average highway delay by up to 47%. Other studies even showed that higher congestion through restraining capacity for additional travel appears to be associated with the decrease in regional employment growth rates (5). State-of-art studies investigate under
which conditions and activities the traffic operations are influenced, and the traffic congestion
deteriorates. However, the same as many other traffic problems that are closely related to human
activities, the answers to correlation studies of the traffic congestion problems may be quite
diverse.

Figure 1 Recurrent traffic fluctuations of hourly traffic occupancy in one sample detector

Created in March 2006, Twitter founds a perfect stage of “We Media” and this makes possible
the wide-range information retrieval about public activities from the massive majority of people.
Over the past decades, social media has been validated useful to broadcast major events such as
natural disasters (6; 7), bird flu (8), politic events (9), etc. The problems are whether it can detect
the traffic surge and whether there exists any possible correlation between the tweets and traffic
surge. Usually, the Twitter contents that are related to specified events will occur
disproportionally frequently over certain time and space, and it is possible to make a direct
connection between tweets and events like ceremony opening (10), celebrity death or festival
parades (11). However, the non-recurrent traffic surge may be quite different from that of other
events. This is because a variety of public activities possibly incurs more on-road traffic and
cause non-recurrent traffic congestions. For example, on June 9th, 2013, an anomalous 10-mile
traffic jam was detected on a major Southern California freeway. At the same time, the keywords
“Obama” and “Impeach” occur much more frequently in the current tweets than the former ones
(12). The contents that most people discuss and post on Twitter may imply a major social trend
and bring more on-road traffic over specific time and space. In our study, we define **Twitter
concentrations** as the tweets that involve a variety of traffic-related activities whose contents are
widely created, consumed, distributed or shared. The goal of this paper is to explore the
correlation between Twitter concentrations and traffic surge.
There are mainly three challenges to be addressed: The first challenge lies in how to quantify the traffic surge that may result from major public events. The time-of-day traffic data collected by loop detectors inheritably contain detection errors, and the traffic occupancy may fluctuate over time-of-day. Thus, it may not be easy to reasonably interpret the traffic data from the observations of one or two days and also not easy to determine the traffic surge according to data records previously collected. Also, traditional studies focused mainly on several intersections (13) or corridors (14; 15). In comparison, archiving, interpreting and summarizing the high-resolution traffic occupancy data in a large road network is quite challenging, considering the huge data size collected by large-scale fixed detectors. In our study, a clustering method is employed, and a detector-based probabilistic model to detect traffic surge is proposed.

The second challenge lies in the inheritable complexity and unstructured nature of tweet data: language ambiguity (16) and how to exact the traffic-related Twitter concentrations from the large collections of tweets is worth studying. We extract the Twitter concentrations by one or more keywords that make the tweets discriminatively different from that of others. The extracted tweets are further classified to label whether they are traffic-related or not. The prevailing methods of classifying tweets can be categorized into supervised and unsupervised techniques including Naïve Bayes classifier (17), online clustering (18), support vector machine (7), hierarchical divisive clustering (19), discrete wavelet analysis (7), continuous wavelet analysis (20), decision trees (21) etc. The performance of these methods is partially decided by the applications and data sources. In this paper, we employed widely-accepted unsupervised and supervised learning techniques to classify the traffic-related tweets.

The third challenge is that public events that are reflected on the Twitter concentrations may exert different levels of influences on its surrounding traffic. Some of the activities may pose influence on more than one detectors, and some may be in effect within more than one hour. To properly interpret the traffic surge by the data collected from different locations and time periods, we aggregate the data by two different statistics measures: mean and 75th percentile. These values can be compared with the traffic-related Twitter concentrations to explore their correlations with traffic surge.

Our main contributions can be summarized as follows. First, a probability index is proposed to quantify the level of detector-based traffic surge in a large-scale road network that is detailed in Sections 3 and 4. Second, an effective detection method is proposed to extract, filter and classify the traffic-related Twitter concentrations from a total collection of tweet posts, and the method is introduced in Sections 5 and 6. Third, we develop a methodology to evaluate the correlation between a specified tweet post and its surrounding traffic. The details of the correlation studies are in Section 7. The paper ends with some useful conclusions and thoughtful ideas in Section 8.

2 Data and Incentives

The study area, shown in Figure 2, is located in the vast road network of Northern Virginia (NOVA). The network that is more than 50 square kilometers consists of roads connected by
more than 1200 signalized intersections. For each intersection, an average of 12 loop detectors is fixed on the approaches of the intersections. With these traffic detectors, the access to real-time traffic information in our study area is becoming routine as under growing pressure for improving traffic management (22). The traffic occupancy data, which are usually employed as an index of a traffic jam, are collected at an interval of 15 minutes in July 2014. We only study the data collection in the daytime from 4:00 a.m. to 21:00 p.m.

Within NOVA area, we also collected the tweets through Twitter Streaming API with geo-location filter. More than 584,000 tweets were collected throughout 2014. We mainly investigate the tweets in July. To ensure that the tweets are collected from the general public, Twitter users from traffic authorities such as “I95VA”, “MKA_NVA” have been removed from our study. Also, some of the tweets from media or press such as “nbcwashington” etc. were also excluded after our empirical examination. Our incentives are intrigued by some preliminary examinations of the tweets. For instance, in February 2014, there was a keyword surge of “capitalweather”. People tweeted the delay caused by the “Biggest snow storm since Snowmageddon”. Public service info feed “Metrobus Info” lively broadcasted the congestion at North Capitol & New York Ave in DC area. This was widely retweeted by other users. One can say that the public events reflected by Twitter concentrations potentially exert pressure on the road network and aggravate the traffic congestion.

3 Traffic clustering algorithm

Besides the large-scale sensor data, the traffic data inheritably possess time-of-day features such as AM peak and PM peak. Thus, the traffic surge should be justified by a time-of-day clustering method. It is worth mentioning that we do clustering on the data separately on weekdays and weekends because the traffic conditions may be quite different. The following algorithm works almost the same for weekdays and weekends, and we do not intentionally distinguish that.
Backed by this setup, we start from detector level. The traffic occupancy data of detectors are collected every 15 minutes, and we take the median of the traffic data collections in different hour period as the traffic signatures of the detectors. That is:

\[ O^d = (O^d_1, O^d_2, \ldots, O^d_j, \ldots, O^d_N) \]

where \( O^d_j \) is the occupancy median in the \( j \)th hour period in the detector \( d \). There are in total \( N \) hour periods in one traffic signature \( O^d \). As our study hour period is from 4:00 a.m. to 21:00 p.m., \( N \) is set to be 18. The median value possibly eliminates the fluctuations of traffic data in different days and is less likely to be influenced by outliers than mean. Previous study argues that given the combination of direction, connectivity and locality of a road segment, one can distinctively determine the corresponding traffic signature (time-of-day features of speed) of a road segment with high probability (23). Enlightened by this idea, we also assume the following:

• Assumption 1: there exist unchanged traffic signatures in a given detector. The time-of-day traffic occupancy over a certain hour period fall into a reasonable range, and those that are obviously higher from the feasible range are traffic surge.

The traffic signatures in more than 15000 traffic detectors constitute the raw database. To find the feasible occupancy range of each hour period, we employed the K-means algorithm with a principled way of finding the number of clusters and the cluster centers. This algorithm can partition the traffic signatures into finite groups of similar patterns and output the centers of clusters as well as the cluster IDs detectors belong to. The algorithm is shown as follows:

---

**Algorithm 1: Traffic signature clustering**

**Input:** The maximum number of clusters \( k \) and the traffic signature matrix \( O^d \) (in this paper, this matrix contains 15000 rows and 18 columns. Each row is the traffic signatures of detectors).

**Output:** Centers of clusters \((C^1, \ldots, C^i, \ldots, C^k)\);

The cluster IDs detectors belong to.

Assign the initial number of clusters \( k=2 \), initialize \( AIC= +\infty \)

Repeat

Implement K-means clustering algorithm with \( k \) clusters:

Pick randomly the cluster centers \((C^1, \ldots, C^i, \ldots, C^k)\);

Repeat

Cluster each traffic signature \( O^d \) to the nearest cluster center \( C^i \) with \( \min (d(O^d, C^i)) \);

Replace \( C^i \) by \( mean(O^d) \);

Until none of the detectors switches clusters

Calculate the ratio of AIC change: \( diff(AIC)/AIC \)
Until $\text{diff}(AIC)/AIC \leq \epsilon$ or $k=\kappa$

In the algorithm, $\epsilon$ is a threshold value set to be 3% in this paper. The Akaike information criterion (AIC) (24) is employed to measure the relative quality of the clustering results.

$$AIC = \sum_{l}^{k} \sum_{d \in \text{dom}(i)} d(O^{dl}, C^l) + K \cdot N$$

where $O^{dl}$ denotes the traffic signature of the $d$th detector that belongs to $l$th cluster. $C^l$ is the center of the $l$th cluster. $d(O^{dl}, C^l)$ is the Euclidean distance between a traffic signature $O^{dl}$ and its cluster center $C^l$. $\text{dom}(i)$ is the domain (collection) of all detector ID whose traffic signature belongs to $l$th cluster. $k$ is the current number of clusters. $N$ is the count of elements in a traffic signature which equals to 18 in our study. Theoretically, the smaller the $AIC$ is, the better the clustering result should be. For computational efficiency, the algorithm stops when the increase of cluster number brings no more than 3% additional benefits. The AIC results are shown in Figure 3. When $k=15$, $\text{diff}(AIC)/AIC$ goes lower than 3%.

There recommend two important criteria in selecting the number of clusters: First, the cardinality of small size clusters may decrease with the increasing of the number of clusters. The cardinality should not be too small because corresponding results from clusters of large cardinality produce more reliable cluster centers in the later study. Second, we may also use BIC or other statistics to measure the quality of clustering. Same as that of AIC, the results of other statistics do not also indicate an unconstraint large number of clusters.

The clustered centers are shown in Figure 4. From the shape of our clustering results, it is not surprising that different cluster centers vary not only in shapes but also in scales. It reveals a
clear time-of-day feature for each detector. This method can find the outliers in the traffic
occupancy due to several of its advantages:

• The method fully considers the time-of-day features of traffic patterns inherited in the traffic
data.
• The method is totally based on the field data which is in large scale. The aggregation of large-
scale data may eliminate the possible noises from the results.
• The method clusters the traffic occupancy only in July, which can diminish the effects of
traffic operations in different months.

Figure 4 12 different clustered centers of traffic signatures

4 Traffic surge definition in detector level
For each cluster, the traffic occupancy over a specified hour period should be distributed around their cluster center. An outlier is far away from the cluster center, and its level of deviation from the center can be justified by calculating its probability. We empirically check the distributions of traffic occupancy in all hour periods in different clusters, and two of them are shown in Figure 5. After reviewing the empirical distributions, we can conclude they reasonably approximate them with a normal distribution.

![Figure 5 Distribution of traffic occupancy in two clusters at 7:00 p.m.](image)

Thus, we can justify the severity of traffic increase based on the normal distribution. We define the traffic surge probability as probability $\Phi \left( Z = \frac{O_{d_l} - C_j^i}{\sigma_j^i} \right)$ during the $j$th hour period in detector $d$ and cluster $i$. Here $C_j^i$ is the cluster center in Cluster $i$ during hour $j$; $\sigma_j^i$ is the standard deviation of traffic occupancy in cluster $i$. The closer the value is to 1, the greater the traffic surge should be.

5 Twitter concentration extraction and filtering

We can automatically extract Twitter concentrations that have the following features:

- They are related to real events that many people witness and are willing to share their observations or experience via Twitter.
- They potentially have connections with on-road traffic-related activities and may involve some kind of traffic movement.

First, we extract the Twitter concentrations from the set of tweets from January 2014 to December 2014. In most of the cases, keywords contained in the tweet can differentiate it from other tweets. Our algorithm first splits the tweet texts into separate word characters that form a
large word database. In the database, we first search for the keywords that frequently appear in each day. Then, for all the frequent words, we select the words that frequently appear in one day but not so frequently or even vanish in other days.

**Algorithm 2: Twitter concentration extraction**

**Input:** Tweets collections throughout 2014  
**Output:** Keywords of each day and tweets that contain the keywords

For each day  
Decompose the tweets into vectors of words on that day  
Count frequency $f_w(d)$ of each word $w$ on day $d$;  
Pick the words that satisfy $f_w(d) \geq \epsilon \sum w f_w(d)$  
For all selected $w(d)$  
Select the words $w(d)$ as the keywords $k(d)$ that satisfy  
$\text{count}[w(d) \cap \text{dom}(w(d))] \leq \tau$

For all $k(d)$  
Extract the tweets that contain $k(d)$ in July

We set the frequency threshold for the keywords by a ratio parameter $\epsilon$. $\text{count}[w(d) \cap \text{dom}(w(d))]$ counts the frequency of $w(d)$ in the domain of all $w(d)$, i.e. $\text{dom}(w(d))$ and whose frequency is no larger than $\tau$ is selected. For the selection of $\epsilon$, if we increase the value of $\epsilon$, it may miss some important keywords due to the large datasets. If we lower $\epsilon$, it may incur more computations. Our experiences show that there are not so much difference between 10% and values lower than 10%. The value of $\tau$ should not be too large because longer periods of events will diminish the enthusiasm of the people and these events may not be a reflection of Twitter concentration. One may increase $\tau$ if their data covers more than one year because there may exist yearly events. In our study, we set $\epsilon$ and $\tau$ to be 10% and 3 respectively. By comparing the frequent words in different days, the stop-words such as “is”, “and”, “us”, etc. can be eliminated and the remaining frequent words are the keywords that may indicate a kind of social activity. Table 1 shows some keywords of the day in July and some possible related social events.

Table 1 6 keywords and related social events in July 2014

<table>
<thead>
<tr>
<th>Date</th>
<th>Keyword</th>
<th>Social events</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/1/2014</td>
<td>waffles</td>
<td>Waffle House restaurant just tweeted the most American waffle breakfast possible</td>
</tr>
<tr>
<td>7/3/2014</td>
<td>louis</td>
<td>Louis Zamperini, an American war survivor in World War II, died</td>
</tr>
<tr>
<td>7/4/2014</td>
<td>freedom</td>
<td>Celebration activity such as Freedom Fest 2014 fireworks</td>
</tr>
<tr>
<td>7/18/2014</td>
<td>fotosdeprinceroyce</td>
<td>Prince Royce concerts</td>
</tr>
<tr>
<td>7/16/2014</td>
<td>wjmc2014</td>
<td>Washington Journalism and Media Conference</td>
</tr>
<tr>
<td>7/26/2014</td>
<td>silver</td>
<td>A new metro line: silverline, is opened</td>
</tr>
</tbody>
</table>
Table 2 Transportation lexicon

<table>
<thead>
<tr>
<th>Accidents</th>
<th>Carpooling</th>
<th>Drive</th>
<th>Junc</th>
<th>Passenger</th>
<th>Seatbelts</th>
<th>Trains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival</td>
<td>Carriage</td>
<td>Driver</td>
<td>Junction</td>
<td>Passengers</td>
<td>Shuttle</td>
<td>Transit</td>
</tr>
<tr>
<td>Arrivals</td>
<td>Cars</td>
<td>Drivers</td>
<td>Junctions</td>
<td>Passing</td>
<td>Sidewalks</td>
<td>Transport</td>
</tr>
<tr>
<td>Arrive</td>
<td>Collision</td>
<td>Drives</td>
<td>Kilometer</td>
<td>Pedestrian</td>
<td>Speed</td>
<td>Transports</td>
</tr>
<tr>
<td>Arriving</td>
<td>Combustion</td>
<td>Driving</td>
<td>Kilometers</td>
<td>Pedestrians</td>
<td>Speeding</td>
<td>Trasport</td>
</tr>
<tr>
<td>Auto</td>
<td>Commute</td>
<td>Drop</td>
<td>Lane</td>
<td>Periods</td>
<td>Speedlimits</td>
<td>Travel</td>
</tr>
<tr>
<td>Automobile</td>
<td>Commuter</td>
<td>Eastbound</td>
<td>Licence</td>
<td>Petrol</td>
<td>Speeds</td>
<td>Travelcar</td>
</tr>
<tr>
<td>Automobiles</td>
<td>Commuters</td>
<td>Eastern</td>
<td>Line</td>
<td>Pickup</td>
<td>Standstill</td>
<td>Travelcard</td>
</tr>
<tr>
<td>Automatic</td>
<td>Commuting</td>
<td>Exhaust</td>
<td>Lines</td>
<td>Priced</td>
<td>Steer</td>
<td>Travelcards</td>
</tr>
<tr>
<td>Baggage</td>
<td>Congested</td>
<td>Exit</td>
<td>Link</td>
<td>Queues</td>
<td>Steering</td>
<td>Traveline</td>
</tr>
<tr>
<td>Bicycle</td>
<td>Congestion</td>
<td>Flows</td>
<td>Metering</td>
<td>Rd</td>
<td>Stops</td>
<td>Travelled</td>
</tr>
<tr>
<td>Bicycled</td>
<td>Connect</td>
<td>Freeflow</td>
<td>Motor</td>
<td>Ride</td>
<td>Taxi</td>
<td>Traveller</td>
</tr>
<tr>
<td>Bicycles</td>
<td>Connection</td>
<td>Freeway</td>
<td>Motorbike</td>
<td>Rider</td>
<td>Taxicabs</td>
<td>Travellers</td>
</tr>
<tr>
<td>Bicyclists</td>
<td>Connections</td>
<td>Freight</td>
<td>Motorbikes</td>
<td>Riders</td>
<td>Taxiing</td>
<td>Travellers</td>
</tr>
<tr>
<td>Bike</td>
<td>Crossing</td>
<td>Heading</td>
<td>Motorcycle</td>
<td>Riding</td>
<td>Taxis</td>
<td>Travelling</td>
</tr>
<tr>
<td>Bikeability</td>
<td>Crossings</td>
<td>Heathrow</td>
<td>Motorcycles</td>
<td>Road</td>
<td>Taxiway</td>
<td>Travelling</td>
</tr>
<tr>
<td>Biker</td>
<td>Crossroad</td>
<td>Highspeedrail</td>
<td>Motorcyclist</td>
<td>Roadmap</td>
<td>Taxiways</td>
<td>Travels</td>
</tr>
<tr>
<td>Bikes</td>
<td>Cyclist</td>
<td>Highway</td>
<td>Motorcyclists</td>
<td>Roads</td>
<td>Terminal</td>
<td>Trip</td>
</tr>
<tr>
<td>Bikesafe</td>
<td>Cyclists</td>
<td>Highways</td>
<td>Motoring</td>
<td>Roadsafety</td>
<td>Tolled</td>
<td>Trips</td>
</tr>
<tr>
<td>Bikesharing</td>
<td>Delay</td>
<td>Incident</td>
<td>Motorist</td>
<td>Roadside</td>
<td>Tolling</td>
<td>Truck</td>
</tr>
<tr>
<td>Biking</td>
<td>Delays</td>
<td>Incidents</td>
<td>Motorists</td>
<td>Roadways</td>
<td>Tolls</td>
<td>Trucks</td>
</tr>
<tr>
<td>Biofuel</td>
<td>Departing</td>
<td>Int0change</td>
<td>Motorization</td>
<td>Roadworks</td>
<td>Tour</td>
<td>Trunk</td>
</tr>
<tr>
<td>Biofuels</td>
<td>Departure</td>
<td>Interchange</td>
<td>Motorization</td>
<td>Roundabout</td>
<td>Tourists</td>
<td>Tunnels</td>
</tr>
<tr>
<td>Brake</td>
<td>Departureboard</td>
<td>Interchanges</td>
<td>Motorized</td>
<td>Route</td>
<td>Tow</td>
<td>Turn</td>
</tr>
<tr>
<td>Brakes</td>
<td>Departures</td>
<td>Intercity</td>
<td>Motorized</td>
<td>Routes</td>
<td>Towing</td>
<td>Turning</td>
</tr>
<tr>
<td>Braking</td>
<td>Destination</td>
<td>Intermodal</td>
<td>Motorpoint</td>
<td>Routing</td>
<td>Townbound</td>
<td>Uncongested</td>
</tr>
<tr>
<td>Breaks</td>
<td>Destinations</td>
<td>Intersection</td>
<td>Motorway</td>
<td>Runway</td>
<td>Track</td>
<td>Van</td>
</tr>
<tr>
<td>Bus</td>
<td>Direction</td>
<td>Intersections</td>
<td>Motorways</td>
<td>Runways</td>
<td>Tracking</td>
<td>Vans</td>
</tr>
<tr>
<td>Buses</td>
<td>Directional</td>
<td>Interurban</td>
<td>Navigating</td>
<td>Rush</td>
<td>Tracks</td>
<td>Vehicles</td>
</tr>
<tr>
<td>Busstop</td>
<td>Distance</td>
<td>Journey</td>
<td>Navigation</td>
<td>Safety</td>
<td>Traffic</td>
<td>Vehicular</td>
</tr>
<tr>
<td>Car</td>
<td>Districts</td>
<td>Journeyplanner</td>
<td>Parkers</td>
<td>Scooter</td>
<td>Trail</td>
<td>Wait</td>
</tr>
<tr>
<td>Cargo</td>
<td>Dock</td>
<td>Journeys</td>
<td>Parking</td>
<td>Scooters</td>
<td>Trailers</td>
<td></td>
</tr>
</tbody>
</table>
is shown in Table 2 which is referred to (26). We made some revisions by excluding some words that are related to air, water, railway traffic, etc. The tweets that contain at least one term in the lexicon are reserved otherwise discarded. We finally extracted 1179 candidate traffic-related tweets.

6 Twitter concentration classification and labeling

Twitter concentration classification is a supervised learning method that calculates the correlation between the tweets and traffic. We employed the logistic regression model, which is first introduced in 1958 (27), as our learning model to train and test the tweets. We first train the model and use the model to label the candidate tweets obtained in Section 5 is traffic-related. The model is as follows:

\[
F(X) = \frac{1}{1 + e^{\beta \cdot X + \beta_0}}
\]

where \(X = (X_1, X_2, ...X_i ...X_m)^T\). \(X_j\) represents the vector of \(i\)th feature and there are \(m\) features in total. \(\beta = (\beta_1, \beta_2, ...\beta_i ...\beta_m)\) is the vector of coefficients of each feature. \(\beta_0 = (\beta_0 ... \beta_0)\) is the vector of intercepts. \(F(X) = (F(X_1), F(X_2), ...F(X_i) ...F(X_m))\) is the vector of probability values of the dependent variable.

The classification method proceeds in the following steps:

First, we pick randomly 2000 tweets that contain one or more words in the Transportation Lexicon in Table 2 from the tweet collections of the whole year. We manually label them to judge whether they are traffic-related. The labeled results are taken as the ground truths as well as the dependent variables \(F(X)\).

Second, each tweet is further decomposed into separate word characters that are called “tokens” in our study. The tokens can be English character, number or even Latinized letters and are taken as the candidates of independent variables. There are more than 6000 tokens in total.

Third, we conduct a stop-word filtering on the candidate feature words. The stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (28). It can rule out the tokens that have no apparent linguistic meanings or significant event indications including articles, conjunctions, prepositions, pronouns, etc. The stop-word list we used referred to (29).

Fourth, we include those tokens that may correlate with the labels. The correlation benchmark we choose is phi coefficient (30), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as \(\phi\)) between two variables \(x\) and \(y\) is calculated as:

\[
\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})}}
\]
When \( x = 1 \), \( n_{11} \) and \( n_{10} \) are the counts separately for \( y = 1 \) and \( y = 2 \); when \( x = 2 \), \( n_{01} \) and \( n_{00} \) are the counts separately for \( y = 1 \) and \( y = 2 \). Those tokens whose correlation coefficient \( \phi \) are higher than 0.05 are selected. These tokens totaling 71 are taken as our covariate features \( X \).

Fifth, we estimate the coefficients of the variables in the regression model by maximum likelihood estimation (MLE). This likelihood estimation can be realized by an iterative process such as Newton’s method (31) and the estimation of both coefficient values and significance are detailed in (32). To increase the accuracy of the predicted model, we implement 5-fold cross validation (33), which is a popular model validation method. Cross-validation can give insight on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data set. We finally obtained an overall estimation by averaging 5 test results. The accuracy of the model is 0.76.

Finally, the prediction model obtained in the previous step is employed to test the candidate traffic-related tweets obtained in Section 5. In our study, we take \( F(X_i) \) as the traffic accident probability of \( i \)th tweet data. The results show that of all 1179 tweets from the first classifier, 164 tweets may correlate with the traffic with \( F(X_i) > 0.5 \).

7 Correlations between Twitter concentrations and traffic operations

For each tweet, we mainly study the traffic related information within certain spatial and temporal ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. It is worth noting that:

- Public activities related to Twitter concentrations may happen either before or after when the tweet is blogged. So does the traffic surge.
- As the geographic impact of public activities may vary, the traffic surge may exist in one or even more intersections nearby.

Thus, influenced by public activities related to the Twitter concentrations, there are mainly two different traffic surges: traffic surge in part of the detectors or over shorter time periods; traffic surge in most detectors and over long time periods. The first kind of traffic surge can be justified by the 75th percentile value of traffic surge probability:

\[
q_{traffic} = Q3\left\{ \phi \left( Z = \frac{O_{dj} - C_{j}}{\sigma_j} \right), d \in dom(d) \cap j \in dom(j) \right\}
\]

Where \( j \) is the hour period, \( d \) is the detector ID and \( i \) is the cluster ID. \( dom(d) \) is the domain of all the detectors within the geo-scale of the tweets and \( dom(j) \) is the domain of all time periods within the time-scale of the tweets. \( Q3() \) is the operator of 75th percentile. As this kind of traffic
surge is dramatic in only part of the detectors while relatively mild in other detectors, value of $q_{\text{traffic}}$ corresponding to a tweet should be relatively high to justify a traffic surge.

The second kind of traffic surge can be justified by averaged traffic surge probability:

$$p_{\text{traffic}} = \frac{1}{NUM} \sum_{j \in \text{dom}(j)} \sum_{d \in \text{dom}(d)} \phi \left( Z = \frac{O_{jt}^{d} - C_{j}^{d}}{\sigma_{j}} \right)$$

Where $NUM$ is the total number of traffic occupancy data related to a tweet.

For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by a threshold value of either $p_{\text{traffic}}$ or $q_{\text{traffic}}$. Here is our assumption on the detection of traffic surge from Twitter concentrations:

- **Assumption 2**: For a traffic-related Twitter concentration, its correlation to traffic surge can be justified by either $q_{\text{traffic}} \geq q_{\text{traffic}}^{0}$ or $p_{\text{traffic}} \geq p_{\text{traffic}}^{0}$, where $q_{\text{traffic}}^{0}$ and $p_{\text{traffic}}^{0}$ are two parameters.

Given Assumption 2, the public events of Twitter concentrations can impact the surrounding traffic, and this impact can be justified by these two traffic surge probabilities. Two important findings are worth mentioning:

- Empirical results show that the impact of different threshold values of $q_{\text{traffic}}$ on the result is low and even negligible.
- If we set the threshold of $q_{\text{traffic}}$ to be 0.8, the percentage values of traffic-justified Twitter concentration events may change with $p_{\text{traffic}}$ as shown in Figure 6. Given $q_{\text{traffic}}^{0} = 0.8$ and $p_{\text{traffic}}^{0} = 0.5$, 127 out of 164 Twitter concentrations (77.4%) can be justified by traffic surge.

![Figure 6 Percentage of traffic-justified Twitter concentrations under different threshold $p_{\text{traffic}}^{0}$](image)

Figure 6 Percentage of traffic-justified Twitter concentrations under different threshold $p_{\text{traffic}}^{0}$
Different threshold values $p_{\text{traffic}}^0$ may influence the final results. It is obvious that a higher $p_{\text{traffic}}^0$, indicating a more serious traffic surge condition, may correspond to a public event that arouses more Twitter concentrations. It will be of great use in future study to further explore and quantify the severity levels of traffic surge using Twitter concentration.

Table 3 presents two Twitter concentrations and its corresponding public events. Figure 7 illustrates the time-of-day fluctuations of traffic surge probabilities of these two Twitter concentrations. One can see that the overall traffic surge probabilities in Figure 7(a) and 7(b) are above 0.5 for Twitter concentration (1). As a comparison for (2), traffic surge probabilities in one detector are high (see Figure 7(d)), but low in the other (see Figure 7(c)). This figure characterizes the influence levels of different Twitter concentrations in different geographic scales. In Table 3, keywords “silverline” and “4thofjuly2014” can justify the correlation between tweets and major public events. The Twitter concentrations indicate the occurrence of traffic-related activities that result from the public events. The results prove the potentials of Twitter concentrations in detecting the traffic surge. One can see that without knowing the type of public events in advance, detecting the traffic-related Twitter concentrations assists in interpreting the causality of traffic surge and provides insights for better decision-making in urban traffic management.

Table 3. The keywords, Twitter concentrations and public events corresponding to Figure 7

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Twitter concentration</th>
<th>Public events</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) silverline</td>
<td>waiting at Wiehle to ride silverline</td>
<td>Silverline metro opened on July 26</td>
</tr>
<tr>
<td>(2) 4thofjuly2014</td>
<td>waiting for my friend to get here so we can roll out easternshore, 4thofjuly2014</td>
<td>Celebration activities on Independence Day</td>
</tr>
</tbody>
</table>
8 Conclusions and discussions

Our paper investigates the correlation between Twitter concentrations and the traffic surge on July 2014. The results prove the potentials of using tweets to detect the traffic surge within a given scale of space and time. First, the traffic occupancy over a certain period may follow a normal distribution, and this feature is fully exploited to derive the probability that quantifies the traffic surge. Second, the correlation between Twitter concentrations and traffic surge indicate that the major social activities that are related to traffic could possibly deteriorate the nearby traffic congestions. Our experiments show that 77.4% of traffic-related Tweeter concentrations can be justified by local traffic surge.

Following these findings, one may further the study by tackling the limitations of the current approach. In our study, the tweets are collected through Twitter Streaming API with geo-location.
filter and cannot possibly cover all the traffic surge of the whole region. This may be due to the limited volume of geo-tagged tweets. Also, the classification method employed in our paper may be limited by the size of the training datasets, and the precision of the results may increase by incorporating more tweets.

The potential applications of our study are also promising: First, the traffic surge detection algorithm is built on the “big data” analysis of previous data collections. It can precisely unveil the traffic patterns in a large road networks and even identify the anomaly traffic conditions.

Second, this study can help traffic operators understand the cause of traffic surge and improve short-term prediction of traffic congestion (especially non-recurrent congestion) on roadways in the future. Third, the Twitter concentration can broadcast the traffic-related events in a much more timely and quickly manner than traditional broadcasting media. Monitoring the social media data may deliver useful traffic event information, including traffic accident, traffic jam, road construction, etc. It will also be interesting to analyze the spatial-temporal correlations between traffic patterns and Twitter concentrations in future research.

Acknowledgement

This research is supported by Transportation Informatics (TransInfo) University Transportation Center at University at Buffalo, The State University of New York.
Reference


