Extended Implementation Methodology for Virtual Sensors: Web-based Real Time Transportation Data Collection and Analysis for Incident Management

Abdullah Kurkcu, M.Eng. (Corresponding Author)  
Graduate Research Assistant, UrbanMITS Laboratory, Department of Civil & Urban Engineering,  
Polytechnic School of Engineering,  
Center for Urban Science + Progress (CUSP), New York University (NYU)  
One MetroTech Center, 19th Floor, Brooklyn, NY 11201, U.S.A.  
Tel: 1-(646)-997-0538  
Email: ak4728@nyu.edu

Ender Faruk Morgul, M.Sc.  
Graduate Research Assistant, UrbanMITS Laboratory, Department of Civil & Urban Engineering,  
Polytechnic School of Engineering,  
Center for Urban Science + Progress (CUSP), New York University (NYU)  
One MetroTech Center, 19th Floor, Brooklyn, NY 11201, U.S.A.  
Tel: 1-(646)-997-0531  
Email: efm279@nyu.edu

Kaan Ozbay, Ph.D.  
Professor, UrbanMITS Laboratory, Department of Civil & Urban Engineering,  
Polytechnic School of Engineering,  
Center for Urban Science + Progress (CUSP), New York University (NYU)  
One MetroTech Center, 19th Floor, Brooklyn, NY 11201, U.S.A.  
Tel: +1 (646) 997 0552  
Email: kaan.ozbay@nyu.edu

Word count: 5,170 + 9 Figures x 250 words (each) = 7,420 words.

TRR Paper number: 15-3374  
Submission Date: August 1, 2014

Paper submitted for Publication in the  
Transportation Research Record, Journal of Transportation Research Board after being presented  
Transportation Research Board’s 94th Annual Meeting, Washington, D.C., 2015
ABSTRACT

Open data sources and the use of social media data are increasingly gaining attention as important information providers in transportation and incident management. In this paper, we present practical evidence for the emerging potential of on-line and open data sources. We combine and extend our prior research on virtual sensors (1) by integrating real-time incident information and social media network engagement. The fundamental contribution of this paper is to develop an extended virtual sensor (EVS) framework to provide an automated travel time data collection methodology as incidents occur. In addition, it has also been shown that social media data can be potentially useful for more effective real-time incident response. The proposed framework can be easily modified and used to evaluate travel time impacts of incidents on roadways, clearance times, and make use of social media data in terms of obtaining time critical incident related information.
INTRODUCTION

The Internet has evolved dramatically in a very short period of time. A decade ago, on-line information management depended mostly on human beings for today’s simple tasks such as data input or sharing data with other people. Over time, the internet has become the network of devices leveraging smart technologies reducing the need for human intervention. The advances in network technologies have also had significant impacts on Intelligent Transportation Systems (ITS). One such enhancement is the emergence and widespread use of mobile sensor technologies. As of today, Global Positioning System (GPS) devices are among the most commonly used form of sensors that provide high quality real-time traffic data for monitoring transportation systems. These ubiquitous smart devices such as mobile phones also act as sensors that can capture and share various form of relevant information in a very efficient way. As more data are becoming available for researchers, the opportunity to gain insights into transportation related problems using those new forms of information has also emerged.

Widespread deployment of smart devices provides not only a cost effective way to monitor transportation systems, but also creates an ideal environment for real time interaction with travelers through the use of these technologies. Many transportation agencies publish on-line traffic information for travelers including real-time alerts for incidents, work zones, and road closures, as well as disruption or scheduled changes in transit service. This information allows individuals to adjust their travel schedules and/or route choices accordingly. Therefore, accurate and timely reporting of travel time measurements are key factors in the success of traffic information services.

In a previous paper (1), we introduced a “virtual sensor” methodology which utilized open traffic data sources from web-based map providers. The presented framework in that paper has a strong potential to be an alternative to fixed traffic sensors, most of which have high installation and maintenance costs. Moreover, the comprehensive network coverage of virtual sensors is a major advantage over traditional data collection techniques. The accuracy of real-time data gathered from on-line sources was statistically validated using reference data from loop detectors and electronic toll tag readers. In addition, evaluation of travel time reliability for a 90-mile section of New Jersey Turnpike was given as an empirical application to the developed virtual sensor methodology. Therefore, the framework can be effectively used to collect vast amounts of high quality real-time data in any section of a roadway to supplement existing data sources (1).

In this paper, the virtual sensor methodology is extended by integrating the capability of automatically acquiring real-time incident information and Twitter (2) feeds related to the captured incidents. Figure 1 exhibits the new framework that is mostly aimed at the representation of travel time variations during and after incidents in minor roads where physical sensor infrastructure is limited or non-existent. We introduce an efficient web based software implementation framework to capture travel times within a short period of time after incidents occurred. We compared Twitter feeds with the actual real-time incident reports from an official incident reporting website, 511NY (3).
LITERATURE REVIEW

Traffic incidents have been identified as a major contributor to increased congestion, generating one quarter of the congestion on US roadways (4). Incidents can impact the traffic flow and operations in regional levels and cause major delays. Furthermore, both large-scale and small-scale incidents can be associated with secondary incidents (5). In order to ameliorate the consequences of incidents and response in a timely and effective manner, traffic incident management (TIM) strategies benefit from real-time data sources, information dissemination among agencies, traffic surveillance information and advanced communication systems (6). Traffic incidents such as accidents or disabled vehicles cause uncertainty in average travel times. TIM systems need to estimate such incidents’ magnitude, delay and queue length (7). The precautions must be taken to decrease the duration of experienced delay and limit the possibility of secondary accidents (8). Moreover, TIM needs to provide guidance information to motorists to reduce the travel time experienced by most travelers.

On the other hand, user generated content on social networking web sites can provide new layers of information about traffic behavior (9). Although, physical traffic sensors and mobile devices with GPS capabilities may detect traffic delays, they usually fail to explain the underlying reasons for the disturbance. It has been shown that social media streams, such as Twitter, not only provide valuable information for such events but also highlight the essential causes of current traffic conditions (10). When combined with other real-time traffic data sources, this information from social media networks can provide significant insights to understand the causes of non-recurrent traffic congestion.

While social media platforms are often used for sharing information about incidents, studies using social media usage for incident management paid more attention to large-scale incidents, such as natural disasters (11). The fundamental reason that makes large-scale events more attractive for researchers is the abundance of social media data. For example, there usually exist more social media messages with a wide spatial and temporal coverage when an earthquake occurs, compared to a traffic accident which only shows its impacts in a smaller area. Sakaki, Okazaki and Matsuo (12) explored the real-time interaction of users with large-scale events within Twitter and suggested an algorithm to collect user posts and to locate a target event. They classified tweets into positive and a negative classes. After classifying tweets, they applied Kalman filtering and particle filtering methods in order to estimate the event location and notify people of an earthquake event in real-time. Vieweg et al. (13) analyzed microblog (i.e. on-line services that users broadcast messages to other subscribers, e.g. Twitter) posts during two large-scale incidents to understand if microblogging services could be used to improve situational awareness. The authors concluded that tweets including geo-location information could easily be identified and automatically extracted for a variety of useful purposes during large-impact rare events.

There are quite a few studies that focus on detecting small-scale incidents using social media data (11). For example, Abel et al. (14) introduced a framework and a web-based system, Twitcident, which connects to emergency broadcasting services and automatically filters out the relevant incident information that has been published through social media messages and microblogs. The system analyzes tweets semantically and provides an automatic filtering of relevant information from social web streams. Semantic enrichment was stated as a key factor that improves the accuracy of the information gathered from Twitter posts.
There are some efforts in the literature to combine real-time traffic data sources with the feedback coming from social networking sites. Lécué et al. proposed a system called STAR-CITY, which has been designed for the analysis of current traffic conditions, semantic traffic analytics and aggregating heterogeneous data for Dublin City, Ireland. Their system presented how the severity of road traffic congestion can be conveniently diagnosed, analyzed, and predicted using historical and real-time data sources. Similarly, Daly, Lecue and Bicer introduced a methodology which merges open data sources and social media data in order to understand the traffic behavior in real-time and to be able to explain the causes of traffic congestion. Schulz, Ristoski and Paulheim proposed a methodology for real-time identification of small-scale incidents using microblogs to enhance the situational awareness by gathering detailed information about incidents. They compared the detected incidents with real-time incident reporting systems in Germany and found a hundred percent accuracy with approximately ten tweets per accident. Most of the aforementioned research did not compare the event reporting time with real-time public incident information sources and did not explore the actual impact of the incidents on traffic.

Transportation researchers have also explored the potential to harvest the extensive information provided by social media networks for travel demand modeling, emergency planning and analysis. In particular, Collins, Hasan and Ukkusuri demonstrated the use of social media data to evaluate transit rider satisfaction. They utilized sentiment analysis of transit riders’ social media messages. The analysis showed that riders are more inclined to post negative comments than positive comments about a transit situation. Cebelak showed the potential ways to estimate origin-destination (OD) locations with higher temporal resolution and a lower cost compared to traditional methods. The results of that study showed an advantage for reducing the OD estimation errors.

Available real-time transportation data sources are widely dispersed and there are no standards for the provided data format. An integrated real-time traffic data collection methodology that combines publicly open data sources with social media data in a single database can be extremely useful for timely decision-making both for public agencies and travelers. While combined real-time traffic data can be analyzed to evaluate the reliability of travel times during regular operations, congestion patterns due to incidents can be analyzed for future improvements.

The major contributions of this current study are threefold. First, we outline a methodology for capturing incident information in real-time from open traffic data sources from web-based map providers and social media data. Second, the location information of an incident is automatically geo-located on the virtual sensor framework proposed in and real-time information is obtained from Bing Maps traffic application programming interface (API) services. The travel time collection process continues until the normal travel conditions are recovered, for a post-event evaluation of the impacts on network. Third, Twitter feeds for the event location are investigated using a machine learning algorithm. Overall, the developed system enables an effective integrated system for incident management in a cost-efficient way.

The rest of this paper is structured as follows. Section 3 describes the methodology and developed framework used to capture travel times during incidents. Section 4 presents the data used and defines the machine learning algorithm used in the paper. In Section 5, we introduce a real world application of the framework for incident detection and comparison with Twitter data. The last section offers conclusions and indicates possible future extensions.
METHODOLOGY

The proposed travel time and incident data collection process only requires a web browser and internet connection. Our methodology relies on an analysis framework that consists of three major components. Firstly, we improved our “virtual sensor” methodology which uses open data sources provided by web-based mapping services to collect travel times. Secondly, we developed a program to collect real-time incident data from the 511NY data feeds provided by New York State Department of Transportation. These two components are developed using the PHP programming language (version 5.4.16) and integrated to start collecting travel times when incidents occur on the predefined road segments. Finally, we developed an application in Python programming language to filter Twitter stream for incident related keywords and location variables using Tweepy (27), a user-written library for accessing the Twitter API.

Twitter API

In its current version (.v1.1), the Twitter API does not allow users to apply two filters to the stream at the same time. In particular, twitter stream cannot be simultaneously filtered by location and keywords to find incident related tweets with location information. For a case when only the geo-location filter is applied, it is not feasible to identify incident related tweets due to the vast amount of irrelevant data. To address this issue we performed semantic analysis on the location filtered tweets. Figure 2 summarizes the machine learning algorithm used in this paper.

The Naïve Bayes (NB) (28) method is utilized for text classification to identify incident related tweets because of its simplicity and high efficiency. In the NB classifier each tweet is considered as a collection of words. The probability of a class value \( c \) given a test document \( d \), a tweet in this case, is computed as:

\[
P_{NB}(c|d) = \frac{(P(c) \sum_{i=1}^{m} P(f_i|c) n_i(d))}{P(d)}
\]

Where \( n_i(d) \) represents the count of feature \( f \) found in tweet \( d \). There are a total of \( m \) features. Parameters \( P(c) \) and \( P(f_i|c) \) are obtained through maximum likelihood estimates.

The machine learning algorithm explained in detail is trained and used to classify incident related tweets which contain geo-location information. The tweets are collected by applying a keyword filter to the Twitter feed for the training set. The list of keywords used in this study are given in below.

\[
['crash', 'accident', 'delay', 'congestion', 'traffic', 'collision', 'emergency', 'incident']
\]

These tweets are further filtered by tweets only in English language and then preprocessed. In the preprocessing phase, first tweets are converted to lower case and duplicate statuses are removed. Second, mentions\(^2\) and replies in the posts are excluded since they do not provide any additional

---

\(^1\) All the codes are available upon request.

\(^2\) A mention is a Twitter update that contains a username after “@” sign in the body of the tweet. Replies are also considered as mentions.
information. Third, the most commonly used words such as “the”, “be”, “along” are removed to improve the classifier’s performance. Finally, punctuations and web addresses (URLs) are screened out from the posts. After preprocessing, the tweets are manually classified as positive or negative. The approach used in the study mainly focus on the relation between the tweet and the incident rather than understanding the content. If the user post is an incident related tweet, it is classified as positive indicating that an incident occurred. Each tweet in the training set is divided into words and each word is joined to the feature vector. The feature vector is used to build a model which the classifier learns from the training data and decides if the tweet is incident related or not. For example, for a tweet “I just destroyed my car”, the feature vector is [‘destroyed’, ‘car’] and since it is an incident related tweet, its label is positive.

**Incidents API**

The real-time incident information is publicly available in XML format and the feed is updated real-time on 511NY developer resource area. For automated log-ins to the 511NY server, we used PHP extension cURL to send POST requests. To pull the incident data in XML format for Incident API, the username, password and data type parameters are passed within http POST request. The response from the server is converted into tabular format and stored in MySQL tables. Each event in the response contains 39 different attributes with descriptions. Since the incident data feed is updated every 30 seconds, Incident API sends http POST requests every 30 seconds.

**Virtual Sensor API**

Several web mapping service providers deliver real time route information, travel times, traffic conditions and service disruptions. Google (29) was the first mapping service to aggregate data from many sources including historical traffic statistics and real time traffic conditions to estimate the travel time between any two locations. After Google, many other web mapping services started to provide real time traffic information and launch APIs to allow developers to share content and data between each other such as Microsoft Bing Maps (30) and MapQuest (31). The Bing Maps API is utilized for extracting real-time traffic information in the EVS methodology.

The first improvement to virtual sensor methodology was to switch the programming language from JavaScript to PHP. For the roadway segments to be monitored, users need to edit a csv file containing OD information for start and end points of the segments. Each added route in the csv file should have longitude, latitude information and a unique route identifier (route_id). Routing requests for each OD pair are sent every five minutes automatically through the PHP interface. The response from the server is stored in a MySQL database in tables that are specific to each road. Furthermore, we developed a system that checks the incidents and automatically starts collecting travel time information as soon as an incident occurs on the monitored roads. For the study, the Route 1, Route 18, Route 27 of New Jersey and Garden State Parkway (GSP) were investigated. The location information for the test roads were entered manually for 5 to 10 mile length segments in both directions to create an OD lookup table.

In the EVS methodology, the algorithm first checks the facility name. If the facility name is matched, then it determines the direction of the incident. Depending on the direction of the incident, the OD pairs are created and stored in another MySQL table. The OD pairs are usually located within a 5 mile radius of the incident. The virtual sensor API reads the OD information
from the look-up table, and then it starts collecting travel times. However, to avoid overloading the database, the time period of travel time collection for a single incident is limited to one week. Figure 3 exhibits all the steps in the incident detection and travel time collection processes.

DATA

Twitter Dataset

For illustrating the performance of twitter analysis, the data was collected from Twitter’s real-time streaming API between July 6th, 2014 and July 16th, 2014. Twitter limits the streaming feed at any time to 1% of the total tweet volume. However, when the keyword filter is applied to the global stream, it is possible to harvest more incident related tweets as long as the resulted dataset size is less than 1% of total tweet volume. This limitation may be overcome by utilizing Firehose access level provided by Twitter. This feature allows researchers to access 100% of all public tweets. It may be possible to fetch more incident related tweets using the full access to the stream. Figure 4 shows the normalized geo-location filtered tweets with respect to the maximum number of tweets collected during July 8th, 2014 (333,251). In total, 5,044,502 unique user posts were obtained and 44 per cent of the harvested tweets had geo-location information.

Training Dataset

2 million tweets were collected to build a training set using the Twitter streaming API from July 1st, 2014 to July 4th, 2014. The randomly selected set of 5 thousand tweets were manually labeled as positive and negative. Positive examples were the ones that match with the reports of actual incidents, tweets about incidents from users, news agencies as well as transportation agencies. The rest of the tweets were labeled as negative. The examples of positive and negative tweets can be seen below.

| (+)  | Getting in a car accident is a great way to start the day!!! I hate my life!!! |
| (-)  | Nate I accidentally favorited that tweet |
| (+)  | Seen two car accidents today - Drive safe guys! |
| (+)  | Yup if you guys see that huge car accident in pittsford some ***** swerved into my lane and hit me head on URL |
| (-)  | That was an accident Lexi |
| (+)  | Avoid Morris in front of Kean, 4 car accident #hope #everyoneisok |
| (+)  | Just witnessed a car accident, Im ***** shaking |

Test dataset

A new set of tweets was used to measure the training set’s performance by checking its ability to classify tweets on unseen data (i.e. data not in the training set). The test data were collected during
the period from July 5th, 2014 to July 6th, 2014 and it consists of 355 incident related tweets and 355 not incident related tweets.

The accuracy of the filtered tweets, whether they refer to actual incidents, is investigated using different sizes of training sets and the results are depicted in Figure 5. Using a small training set of 100 tweets resulted in a poor accuracy of 40% in notifying a real incident using the test data set. When the size of training set was increased to 1000, the accuracy improved to a level that is more than 80%. However, increasing the training set size further did not make a significant enhancement in accuracy. The final training set contains 959 positive and 959 negative tweets.

Once the text identification model is trained, it is applied to geo-location filtered Twitter data set to detect incidents and incident locations. The filtering process can also be used to check the slices of the entire tweet sample, -i.e. more than 2000 tweets per hour, and to classify these tweets if they have useful information for incident detection. Even though the filtering was not applied real time to the stream in this study, it is feasible to be used in practice.

One of the major goals of this study was to show the capability of Twitter data as a supplement to existing web based data feeds. The findings support that Twitter data have great potential to be used in developing a tool with real-time incident location detection functionality. An interesting extension to the current study might be the assessment of the coverage area of the event reporting tweets. For example, urbanized or highly populated areas could be better candidates for high social media user response to incidents, whereas sample size collected from local roads or minor arterials could be smaller.

**Incident Dataset**

The incident data used this study consists of 2 weeks of records from all incidents that occurred between July 6th, 2014 and July 16th 2014. There are 3997 incidents in the data set and each record is retrieved from 511NY’s real-time XML data feed. Below is an example of an incident record.
As it can be seen from Figure 6, delays, constructions and accidents constitute the majority of incidents occurred within 2 weeks (i.e. 49% of all incidents). Each event that is less than 2% is grouped under “other”.

**CASE STUDY I: INCIDENT LOCATION DETECTION**

The EVS framework serves as a fully automated real-time data collection tool to avoid human errors in simple tasks such as manually selecting the site for data collection or inputting data to relevant tables. As such, correctly determining the location of an incident is a key initial step. In this first case study we demonstrate the incident location detection using two different data sources: 511NY website and Twitter feeds. The first data source is the public 511NY data feed that publish real-time incident information. The geo-location information is usually included for an incident therefore locating the incident on a map is an easy step as long as the latitude-longitude information is provided correctly. The EVS system uses Bing Maps and during the case study data collection for 4 weeks and for the study period 100% of the geographical coordinate information was accurately matched on the map. For every road of interest, milepost geo-location look-up tables are included into the database. Below is an example of the look-up table for the GSP.

---

| Event Id: 2014072913470901104 |
| State: opened |
| Class: special event |
| Type: Special event |
| Report Organization ID: NYC DOT |
| Facility Name: 4th Avenue |
| Direction: northbound |
| Article code: between |
| From Location Point: Lafayette Street |
| To Location Point: E. 15th Street |
| Create Time: 7/29/2014 1:52:57 PM |
| Last Update: 7/29/2014 1:52:57 PM |
| Event Description: NYC DOT: special event on 4th Avenue northbound between Lafayette Street (New York) and E. 15th Street (New York) Summer Streets, Saturday August 2nd, 2014, 07:00 AM thru 01:00 PM, all lanes closed |
| City: New York |
| County: New York |
| State: NY |
| Estimated Duration in Seconds: 342423 |
| Latitude: 40.7307792109442 |
| Longitude: -73.9904952049255 |
| Lane Status: closed |
| Lane Description: all lanes |
| Start Date: 8/2/2014 7:00:00 AM |
| End Date: 8/2/2014 1:00:00 PM |
| Closure Type: 0 |
Once the exact location of the incident is spotted on the map and the direction of the incident is verified, the framework selects a route ID that is a sufficiently long section of the roadway, and monitors the change in travel time during and after the incident. In this particular case, the incident occurred at coordinates 40.648693, -74.2873188. The EVS algorithm automatically selected the route ID 25, a 6 mile section of the road in the northbound direction, and the virtual sensor location for the OD pair was automatically determined. The accident was reported in 511NY system on 11 July 2014 at 07:12 p.m. and virtual sensors started collecting travel time information at 07:13 p.m.. Data collection continued for a week until July 18th with the main goal of capturing normal operational characteristics of this section. Figure 7 shows the location of the accident and virtual sensors on the GSP. As mentioned earlier, the primary focus of virtual sensors presented in (1) is on the minor roads and local arterials where physical sensor infrastructure is limited or does not exist. For those parts of the network, virtual sensors immediately starts collecting travel time information following an incident and stores the collected data for further analysis. Although data collection can be performed without time and location limitations using the virtual sensors framework, such an application could require large data storage space and more computational process power as the coverage network gets larger. The provided framework on the other hand, only gets activated for a local road when there is an incident, therefore reducing expensive computation time. Moreover, collected data can easily be exported to simpler data formats (i.e. MS Excel) and be classified by incident type and actual impacts on the network can be analyzed later.

Although 511NY is a well-developed system that offers verified incident information, the report times for some situations can be delayed from the actual occurrence time of that incident, especially for minor roads. In particular, for some cases real-time data collection could only start after the incident is partly cleared from the road, which results in a missed time period that the effect of incident is usually at its peak. Therefore as a second data source, Twitter feeds were analyzed. As shown in the literature, people respond to small-sized events on social networks and for some cases they can post information minutes before the 511NY data feed. For example, below tweets were posted by Twitter users who witnessed an accident on Thomas Mathis Bridge in New Jersey.

```
“Accident on the bridge on my way home :(”
4:15 PM - 13 Jul 2014
“Car accident on the bridge is ridiculous”:
5:15 PM - 13 Jul 2014
```
The information about the same accident appeared in 511NY data feed system at 04:39 p.m., 14 minutes after the information about the same incident started to appear on social media networks. The time difference could be critical for later uses of the collected data, especially when there are no additional data sources as in the case of some local roads. Twitter data can be worth investigating for an improved accuracy of the accident start times. However, Twitter data is extremely noisy. Thus, it can be useful only after a number of filtering phases as explained in the methodology section. The screened out tweets should also be used with extreme care assuming the risk of a false alarm. If selected keywords such as “accident”, “crash” or “delay” have been frequently posted for certain locations and by a set of different users these tweets can be used for incident location detection in the EVS framework.

In some cases, 511NY reports delays including the incident start and end locations. For example, a delay was reported at 3:28 p.m. on July 14th, 2014 for a 15-mile section on Route 1. The report was updated twice until 5:21 p.m. and 511NY system did not state the reason for the northbound delay. In such cases, the filtered tweets can be used to understand the reasons of the incidents, the effects on traffic and to locate the critical sections of the highway. Figure 8 exhibits the virtual sensor’s OD, incident start location and the corresponding Twitter posts. In this case, a picture of the incident explaining the reason of the delay was attached, thus, enabling transportation officials to get further information about the incident that might be relevant.

CASE STUDY II: INCIDENT DURATION

An important application area of the EVS is the incident duration measurement based on the collected data in real-time. Actual impacts of an incident on travel times can be of great interest to both decision-makers and travelers. Therefore in this second case study we present examples from incidents that were detected in the proposed EVS framework and corresponding data measurements showing the effect of incidents on travel times.

Figure 9a shows the fluctuations in travel times when a disabled bus reported on the Route 18. The incident was reported on June 3rd, 2014 at 9:20 a.m. In the current framework, data collection starts within a short time after the incident occurrence, however for illustration purposes travel time data right before the incident was also shown. Regularly for that segment of Route 1, average travel times vary between 10 to 20 minutes for 7 miles, depending on the time-of-day. During the disabled bus incident, average travel time jumped to 32 minutes for the same section. The red line in Figure 9a shows the moment that the alert was posted in 511NY system. Traffic conditions recovered to normal state within 164 minutes according to the travel time measurements. Proposed EVS framework continued to collect data for one week starting from the incident time, to compare the magnitude of the impact with regular traffic conditions observed along this segment of Route 1.

The second example is an accident and recurrent traffic conditions on Route 1. The accident was reported at 07:03 a.m. on 26 June 2014. In this case, 511NY system also had a description about the incident along with a delay estimation. This information was verified with the data collected in virtual sensors, as shown in Figure 9b. The segment where the accident was detected, usually had an average travel time of 10 minutes, whereas for the time period when the accident occurred, average travel time was recorded as 34 minutes. The clearance time of this incident was observed as 136 minutes. During the one week data collection following the incident, all congestion points
were marked. The red line shows the incident and the green lines are the recurrent traffic congestion periods which appear in cycles almost all weekdays. During these cycles average travel times were also higher than the overall average travel time for this section.

Twitter data was also inspected for the two examples given in this section, however there were no tweets associated with the incidents. The reason for weaker social media user engagement in responding traffic events could be correlated with the lower traffic demand on minor roads.

CONCLUSION AND FUTURE DIRECTIONS

Alternative data collection procedures have gained increased popularity among researchers in various fields, especially with the advances in mobile and wireless devices in the last two decades. This paper presents an advanced data collection methodology for transportation analysis. Using on-line open data sources for real-time traffic information, an integrated data base system is described. The primary objective of this study is to acquire and use data for sections in a transportation network where physical sensor infrastructure is limited at best.

As opposed to the deployment of costly fixed road-based technologies, the proposed EVS methodology focuses on utilizing free web based data sources as an efficient way of real-time data generators. This nature of EVS can make it attractive to public agencies that would like to avoid high sensor installation and maintenance expenditures. One of the major functionalities of the EVS is the incident information management that is based on two major independent data sources: real-time public data feeds and social media network, -i.e. Twitter, data. Incidents reported from both sources are automatically analyzed, located on the map and real-time average travel time information is also stored. The traffic data collection methodology, mainly follows from Morgul et al. (1)’s virtual sensor methodology. As shown in illustrative case studies, presented methodology can deliver useful information for post-evaluation of traffic incidents including actual impacts on roadways, clearance times, and social media users’ engagement in event management.

This study also highlights some interesting findings about real-time incident reporting using social media data. Despite being largely unexplored for transportation topics, several studies in the literature showed the effectiveness of social media networks in large- and small-scale events. In testing the EVS framework, a number of cases were observed where the incident reporting were considerably faster in Twitter data compared with public data feed, 511NY. However, proper filtering of this social media is necessary to avoid false reports and alarms.

Future research should focus on further integration of social media data in the EVS framework. Semantic enrichment methods can improve analyzing the content of the user posts and for certain incidents such as major accidents, timely information can be transferred to control centers for effective allocation and mobilization of resources and teams. One useful outcome of EVS framework can be pre-estimated clearance times by different criteria such as road type, time-of-day or incident type. This estimation can be extracted based on the historical observations stored in the integrated data base and disseminated to travelers as an immediate prediction of clearance times when an incident occurs.
ACKNOWLEDGMENTS AND DISCLAIMER

The contents of this paper only reflect views of the authors who are responsible for the facts and accuracy of the data and results presented herein. The contents of the paper do not necessarily reflect the official views or policies of sponsoring agencies. The authors appreciate the New York University (NYU) Polytechnic School of Engineering and Center for Urban Science and Progress (CUSP) for partially funding this study.
REFERENCES


LIST OF FIGURES

FIGURE 1  Proposed Web Based Data Acquisition Methodology
FIGURE 2  Architecture of the Tweet Filtering System
FIGURE 3  Extended Virtual Sensors (EVS) Algorithm
FIGURE 4  Normalized Tweets by Day
FIGURE 5  Twitter Training Set Size: Sensitivity to Accuracy
FIGURE 6  The Distribution of Incidents
FIGURE 7  EVS Origin and Destination Locations
FIGURE 8  Incident and Tweet Location
FIGURE 9a  Travel Time Collection: Disabled Bus
FIGURE 9b  Travel Time Collection: Accident with Injuries
FIGURE 1 Proposed Web Based Data Acquisition Methodology
FIGURE 2 Architecture of the Tweet Filtering System
FIGURE 3 Extended Virtual Sensor (EVS) Algorithm
FIGURE 4  Normalized Tweets by Day
FIGURE 5  Twitter Training Set Size: Sensitivity to Accuracy
FIGURE 6  The Distribution of Incidents
FIGURE 7 EVS Origin and Destination Locations

NJ Turnpike Auth.: Accident on Garden State Parkway northbound North of Exit 136 - Stiles St (Cranford Twp) right shoulder blocked
Create Time 7/11/2014 7:12:00 PM
FIGURE 8 Incident and Tweet Location
FIGURE 9  a) Travel Time Collection: Disabled Bus
FIGURE 9  b) Travel Time Collection: Accident with Injuries