CHAPTER

SOCIAL MEDIA DATA IN TRANSPORTATION

11

Sakib M. Khan[1], Linh B. Ngo[1], Eric A. Morris[1], Kakan Dey[3], and Yan Zhou[2]

[1]Clemson University, Clemson, SC, United States [2]Argonne National Laboratory, Lemont, IL, United States [3]West Virginia University, Morgantown, WV, United States

11.1 INTRODUCTION TO SOCIAL MEDIA

Social media is quickly becoming ubiquitous. According to one study, 89% of Americans use the Internet, and 72% use smart phones, with comparable figures across the developed world [1]. Moreover, these technologies are rapidly spreading in the developing world; for example, in a sample of developing nations including China and India, smart phone use increased from 21% of adults in 2013 to 37% in 2015, with more growth certain to come. Moreover, a majority of users of those technologies use them to take part in social media. According to Pew, 76% of Internet users worldwide participate in social media, with a rate that is even higher in much of the developing world. Thus multiple social media platforms have enabled interactions between millions and even billions of users in real time.

Social media can be defined much more generally than some specific formal social networking online sites, such as Twitter, Facebook, or LinkedIn. More generally, any website or application providing any type of social experience through interactions of users is considered to be a social network. For example, websites that are not formally known as social networking websites (e.g., Flickr or Youtube), but which allow social interactions, may also be considered to be social media. Similarly, many Internet-driven, cellphone-based chat applications allow social interaction also.

Over time, online social media platforms have become globally accepted as a means of sharing data by the public, as these platforms have few or no restrictions on distributing information in an affordable manner [2]. People share personal thoughts/sentiments, information about daily activities, and online images/videos via social media. Businesses share advertisements, and consumers share product reviews. Moreover, analysis of breaking news and political commentary is also posted regularly on social networking sites. In part because it is possible to retrieve valuable information from social media irrespective of location, social media data have been extensively utilized in different domains including political campaigns, organizing mass movements, disaster and crisis response, and relief coordination [3].

The role of social media in transportation is increasing. Federal, state, and local transportation agencies have been increasingly using social media to disseminate traffic and construction information. Looking ahead, social media might potentially revolutionize many aspects of transportation planning and engineering because it generates vast amounts of data in real or near-real time. Researchers have started mining social media and crowdsourced data for diverse transportation

Data Analytics for Intelligent Transportation Systems. DOI: http://dx.doi.org/10.1016/B978-0-12-809715-1.00011-0

engineering applications such as traffic forecasting for planned and unplanned events, traffic incident detection, and traffic condition assessment [4]. Multiple studies have demonstrated the viability of social media data for augmenting and even replacing the traditional transportation data collection platforms used by public agencies [4,5]. A recent Transit Cooperative Research Program report identified the five best uses of social media for transit operations: (1) real-time schedule updates, (2) service information, (3) customer feedback collection, (4) employee engagement, and (5) use of social media as an entertainment medium [6]. This study also summarized the major challenges standing in the way of the effective social media use, such as a lack of social data analytics expertise, the need to find appropriate online engagement protocols, cyber security issues, and issues of user privacy protection. All must be addressed to maximize the potential of social media data sources.

This chapter provides an overview of the recent applications of social media data in transportation. Six sections follow. Social media data characteristics are described in Section 11.2. Section 11.3 reviews the most recent social media data analysis tools and algorithms. Section 11.4 presents a brief overview of the emerging social media applications in transportation. Section 11.5 outlines future research challenges and potential solutions. Section 11.6 summarizes the chapter, and Section 11.7 offers concluding remarks on social media data for transportation applications.

11.2 SOCIAL MEDIA DATA CHARACTERISTICS

In order to exploit the data generated by a wide array of social media platforms, we need to understand the data characteristics. While there are many different social media platforms, only those which support certain types of social activities are appropriate for use in transportation applications. For example, LinkedIn, a highly successful environment which facilitates professional networking, will not be an appropriate source of traffic information. On the other hand, Twitter or Facebook are platforms that support spontaneous and heterogeneous contents that can be mined to support transportation applications.

The qualitative and quantitative characteristics of social media "Big Data" include volume, velocity, veracity, variety, and value. To illustrate each in turn we present a sample tweet that is a part of the real-time collection created by tracking the term "traffic accident" on the Twitter Streaming Application Programmers Interface (API).

Example 11.1

A tweet acquired via tracking the keyword "traffic accident" live on Twitter.

{"created_at":"Fri Jun 24 14:51:58 +0000 2016","id":746355191134359552,"id_str":"746355191134359552","text":"**Accident cleared on Crescent Cty Connection NB at US-90 #traffic #NOLA https:\/\/t.co\/9is45BqHqI**","source":"\u003ca href = \"http:\/\/www.sigalert.com\/Map.asp?region = New + Orleans\"rel = \"nofollow\"\u003eTTN NO traffic \u003c\/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":249850238,"id_str":"249850238","name":"TTN New Orleans","screen_name":"TotalTrafficNO","location":"New Orleans, LA","url":"http:\/\/www.totaltraffic.

com",“description”:“If you see traffic problems call (504) 620-1000”,“protected”:false,“verified”:false,“followers_count”:4122,“friends_count”:316,“listed_count”:115,“favourites_count”:39,“statuses_count”:40835,“created_at”:“Wed Feb 09 22:32:34 +0000 2011”,“utc_offset”:-18000,“time_zone”:“Central Time (US & Canada)”,“geo_enabled”:true,“lang”:“en”,“contributors_enabled”:false,“is_translator”:false,“profile_background_color”:“C0DEED”,“profile_background_image_url”:“http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png”,“profile_background_image_url_https”:“https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png”,“profile_background_tile”:false,“profile_link_color”:“0084B4”,“profile_sidebar_border_color”:“C0DEED”,“profile_sidebar_fill_color”:“DDEEF6”,“profile_text_color”:“333333”,“profile_use_background_image”:true,“profile_image_url”:“http:\/\/pbs.twimg.com\/profile_images\/439097660549500928\/nN84gOJS_normal.png”,“profile_image_url_https”:“https:\/\/pbs.twimg.com\/profile_images\/439097660549500928\/nN84gOJS_normal.png”,“default_profile”:true,“default_profile_image”:false,“following”:null,“follow_request_sent”:null,“notifications”:null},“geo”:{“type”:“Point”,“coordinates”:[29.95058,-90.08514]},“coordinates”:{“type”:“Point”,“coordinates”:[-90.08514,29.95058]},“place”:{“id”:“dd3b10083 1dd1763”,“url”:“https:\/\/api.twitter.com\/1.1\/geo\/id\/dd3b100831dd1763.json”,“place_type”:“city”,“name”:“New Orleans”,“full_name”:“New Orleans, LA”,“country_code”:“US”,“country”:“United States”,“bounding_box”:{“type”:“Polygon”,“coordinates”:[[[-90.137908,29.889574],[-90.137908,30.075628],[-89.884108,30.075628],[-89.884108,29.889574]]]},“attributes”:{}},“contributors”:null,“is_quote_status”:false,“retweet_count”:0,“favorite_count”:0,“entities”:{“hashtags”:[{“text”:“traffic”,“indices”:[56,64]},{“text”:“NOLA”,“indices”:[65,70]}],“urls”:[{“url”:“https:\/\/t.co\/9is45BqHqI”,“expanded_url”:“http:\/\/bit.ly\/17huQb8”,“display_url”:“bit.ly\/17huQb8”,“indices”:[71,94]}],“user_mentions”:[],“symbols”:[]},“favorited”:false,“retweeted”:false,“possibly_sensitive”:false,“filter_level”:“low”,“lang”:“en”,“timestamp_ms”:“1466779918832”}

11.2.1 VOLUME AND VELOCITY

Social data communications generated by individuals using mobile devices are typically short. A single tweet is limited to 140 characters, which equals approximately 200 bytes depending on implementation. Images and video contents embedded in Twitter and Facebook are often scaled down to ensure that they can easily be viewed on mobile devices over wireless connections. We also include message boards and forums as a form of social media: pure text postings on forums and message boards are often less than a page in length or several kilobytes in size. However, social media communication occurs at a very high rate and in various formats (e.g., text messages, images, and links to online resources). The volume of social media, which has often been taken for granted as “big” from the beginning of the Big Data deluge [7], comes from the fact that millions of participants actively produce and disseminate social content online. For example, in its IPO filing on 2013, Twitter reported that the number of monthly active users is more than 200 million, and they produced a staggering half a billion tweets per day [8]. This is equivalent to at least 100 gigabytes per day for the core tweet contents only. Moreover, for the tweets to be used in

transportation applications, they need to be packaged and transmitted under a JSON document format. As shown in Example 11.1, the highlighted 200-byte core content is miniscule comparing to the entire tweet size. For comparison, this rate of core tweet content is higher than the daily GPS sampling rate from developing cities such as Dhaka and Nairobi [9], and the actual streaming rate, due to the full size of the tweets' JSON format, would be several times higher.

11.2.2 VERACITY

Because social media data are created by individuals, they carry a degree of uncertainty about the accuracy and consistency of the information conveyed. This uncertainty can come from a multitude of reasons. It could be a difference in how certain words (e.g., street names) are spelled, especially with the 140-character cap imposed on a tweet. For example, highway I-85 could be spelled as I-85, I85, i85, or i-85. As shown in Example 11.1, the location "Crescent City Connection" is tweeted as "Crescent **Cty** Connection." As Twitter's tracker for the Streaming API only supports direct, case-sensitive matching, this aspect of veracity could lead to the inability to collect all relevant data. Another cause of potential uncertainty and inaccuracy is the unintentional dissemination of incorrect information due to the lack of direct knowledge from the person creating the tweet, or just simple misspellings. It is expected that, over time, incorrect information will be corrected by other social media participants. However, it should also be noted that if the individual who makes the initial incorrect tweets is an influential user, it will be more difficult to expect a correction from a community, which naturally trusts influential users, unless the influential user him/herself makes the correction.

11.2.3 VARIETY

Heterogeneity in social media results from two major causes: the variety of the content of the data itself and the variety of the formats of the data structures. While Twitter officially supports text-only messages, the back-end infrastructure allows for the embedding of pictures and videos that can be viewed through Twitter applications. From the perspective of data mining, this means that applications relying on Twitter data will need to have the capability to distinguish between tweets that contain only text and tweets that contain links to other resources, which could be more texts (external web pages), pictures, or videos. The second cause of heterogeneity is the diversity of the social media landscape itself, which leads to differences in the format of data structures required to disseminate the content. This can be observed by looking at the overall structure of a Twitter JSON object [10] where there are attributes which are either obsolete or yet to be active. Furthermore, different social media platforms take a wide variety of forms—including photo and video sharing, blogs, microblogs, social networking sites, social news, and wikis [11]—and these have different data structures. Thus any applications that need to integrate data from multiple sources will have to deal with this heterogeneity issue.

11.2.4 VALUE

The value of social media comes in two forms. First, by posting content to popular social media platforms, transportation agencies at the regional and local levels can and do participate in social media activities to disseminate news and real-time service alerts to the public, for example by

tweeting and blogging about traffic incidents, construction activities, or road closures in their jurisdictions. The second value of social media arises from its potential to furnish data in real time or near-real time to assist with traffic incident detection and management. While there has been much research examining the potential of this idea [4,5,12,13], an extensive investigation [14] shows the lack of accompanying geolocation information with tweets presents significant challenges in extracting relevant data. In Twitter, the geolocation feature is disabled by default to ensure privacy. Positive results are achieved only through analyzing special full-size data made available by Twitter, but access to these data are costly, in many cases the limited social media data available through public venues are unlikely to have enough value for transportation applications. Additional data and data infrastructure are needed to overcome this challenge.

11.3 SOCIAL MEDIA DATA ANALYSIS

Major social media platforms provide APIs that allow limited data access to conduct data mining research. Thus data from social media have been mined for diverse purposes, often with success [15]. Because of the huge number of users, businesses have been using various social media platforms extensively to gauge market sentiment, understand public opinion about products, forecast the success of potential ventures, etc. For example, the social media analytics service provider Samepoint identifies how users perceive different products and discuss them on the social platforms. Similarly, once academic institutions and research centers integrate the social media data analysis facilities, they can perform research by listening to the online community, discovering social perspectives, measuring public sentiment, and engaging in community conversation. Fig. 11.1 shows such a research facility at Clemson University, South Carolina, where social media data is captured and analyzed in the Social Media Listening Center. Salesforce Radian6 is used, to capture more than 150 million sources of social media (e.g., Facebook, Twitter, YouTube, LinkedIn, blogs, and other online communities) conversations, at this center. This center uses a graphical analytics platform to capture sentiment, trending topics, geolocation information and much more from social media contents, which have been used for further understanding online public opinion about college athletics, emergency management activities by law enforcement authorities, etc.

To be useful in the real-time management of transportation systems, transportation applications of social media data would focus on traffic management during diverse social activities such as sporting events or concerts, as well as traffic accidents, weather conditions, and more. However, there are major differences between static data sets and social media data; social media data are dynamic, huge, and noisy compared to traditional transportation data. While data mining techniques have been used in different domains for years on static data sets, the massive volume of dynamic social media data requires the development and refining of specialized tools to identify important and hidden users and group behaviors.

To analyze social media data, supervised and unsupervised machine learning algorithms have been used depending on the organizational structure of data sets (i.e., labeled, not labeled, or partially labeled). Rule-based classification and decision tree methods are the traditional supervised approaches applied to the subset of labeled data known as training data in order to discern patterns to be used to classify data sets. On the other hand, the clustering method makes use of an

FIGURE 11.1

Social Media Listening Center at Clemson University in Clemson, South Carolina [16].

unsupervised algorithm, typically using similarity across data sets without labels, to group data sets. However, to address the diverse/heterogeneous contents of social media platforms, different analytics tools have been developed such as group detection tools [17,18], group behavior analysis tools [19], and influence propagation identification tools [20,21].

Social media data can be considered as graph representations, and mathematical graph theory could be used to analyze social media data. In social media graphs, users are identified as vertexes/nodes. The relationship between nodes is represented by links in a graph. However, it is challenging to fit massive social media data into a graph structure, as it requires a tremendous amount of computer memory and processing power.

While data mining tools are primarily developed by computer scientists, due to the emergence of social media data mining applications in diverse disciplines there is substantial demand for innovative and efficient data mining solutions specific to domain perusing. As has been noted, social media data is in different formats such as text, image, and multimedia, where text is the most common data type. Thus, development of text mining algorithms is critical for different social media data analytics applications. There are substantial numbers of text mining algorithms that are specialized for text content analysis or linkage analysis between nodes, and several recent studies have developed algorithms combining both analysis features.

The keyword search method is very common in analyzing structured social media data such as data in tabular format or tree and graph format. However, it is complex to apply keyword searches on linked document data sets due to: (1) the complexity inherent in the formulation of accurate query semantics considering document content and linkage, (2) the difficulty in developing prioritization

strategies for identified subgraphs satisfying keyword searches, and (3) the complexity of achieving the requisite computation and time efficiency when applying keyword searches to large graphs. DBX-plorer and DISCOVER are the two most popular search algorithms used for XML and relational data sets which use complex database indexing methods [22,23]. In addition, link-by-link and index properties of a graph are used in keyword searches to identify subgraphs in schema-free graphs.

Social network nodes (i.e., user nodes) are often labeled by analyzing their textual contents using text classification algorithms. However, it is challenging to analyze social media text because they often contain nonstandard vocabulary, the labels are sparse (i.e., are distributed in large geometric places), and the contents themselves are noisy. Classification techniques can also use content and linkage characteristics between nodes. Chakraborti et al. [24] applied the first such combined method using web data sources which resemble social media data. In this study, a Bayesian method was used for data sets with known labels. When node labels were missing, a relaxation labeling method was introduced, which labeled nodes without labels using multiple trails until convergence levels were achieved.

Clustering algorithms can be used to partition large networks depending on common characteristics, such as connectivity between nodes. Partitioning of graphs is an NP-hard problem and is challenging for large networks such as social media graphs. Link-based clustering methods have been used to identify communities and clusters of information networks [25,26]. However, developing clustering algorithms is challenging for social media data which feature heterogeneous nodes with different types of contents.

Though keyword search, classification, and clustering algorithms have been introduced in social media data analysis and are used in different domains, most of these techniques are not quite scalable to massive social data, and are not efficient for dynamic data and heterogeneous data sets.

The growing number of sensor devices, such as various in-vehicle sensors or handheld mobile devices, is providing new sets of real-time data (e.g., vehicle location, speed, and direction) that could be integrated with social media data to further improve understanding of users' behavior and road conditions such as traffic congestion due to special events or traffic incidents. Also, cross-validation of findings from social media could be performed using sensor data, or vice versa. However, there are several challenges. First, these new sensor data provide personalized information, so protecting the privacy of users is critical while mining data. Second, though both social media and sensor data are dynamic in nature, sensor data are generated at a much faster rate and require more substantial storage and processing infrastructure. The Citisense application [27], Green GPS [28], and the INRIX vehicle tracking applications [29] are three platforms using sensor data. The Citisense application integrates multiple sources of sensor data such as cell phone location data, taxi cab location data, and fixed roadside sensor data to identify the most active (i.e., densely populated) locations in a city. These types of new analytics platforms provide valuable information to public agencies to assist in allocating resources appropriately to ensure better services.

While users of any social media platform have already established relationships (i.e., links) with other users, in sensor networks there are no explicit relations between data points, requiring the development of derived relationships by analyzing entities' behaviors, locations, and underlying interactions. Dynamic relationships between sensors can be modeled using stochastic properties of collected data and applying hidden Markov models [30]. Also, graph stream models could be used to study relationships between a group of changing participants/sensors by measuring the shortest paths between nodes, linkages between nodes, the topology of the networks, and spatiotemporal dynamics [31]. Synopsis algorithm development for these types of data sets must consider: (1) the

broad applicability of synopsis structures that could be used in different analytical scenarios, (2) compliance with one-pass constraints, (3) the efficiency, in terms of time and space, of large data sets, (4) the effectiveness of dynamic data sets, and (5) reservoir sampling to ensure sufficient sample size at any given time. Sketches, histograms, and wavelet decomposition are three commonly used techniques for dynamic sampling.

11.4 APPLICATION OF SOCIAL MEDIA DATA IN TRANSPORTATION

In the field of transportation, recent studies have explored the usability of social media data for transportation planning, traffic prediction, real-time traffic management during planned and unplanned events (e.g., sports events), and traffic information dissemination [4,13,32–38]. These studies have proposed various methods (including machine learning and forms of statistical analysis) to extract necessary transportation-related data (e.g., congestion status and incident information) from the user-shared contextual information available on social media platforms. Moreover, agencies have employed social media to disseminate and receive information to and from the public.

11.4.1 TRANSPORTATION PLANNING

For transportation planning projects, effective public participation should involve citizens in planning and decision-making processes. Public participation should promote a sense of community by bringing together people who have common goals [39]. In addition to traditional public involvement techniques (e.g., public meetings or surveys circulated online or on paper), online participation methods such as Facebook and Twitter posts are being employed as they may offer avenues for broader public involvement. Studies suggest that the observed tweet sentiments are representative of the opinions of the broader public [40,41].

In addition to promoting the flow of information from the public to transportation agencies, social media can improve information flow from agencies to the public. Majumdar [38] investigated the extent of social media usage by local governments for this purpose. A survey of regional councils of government (RCOGs) in Texas showed that almost half use multiple social media platforms which include Facebook, Twitter, YouTube, LinkedIn, and various blogs to create awareness of agency plans and information (such as maintenance activities that disrupt service) among the general public. It was found that agencies need to further develop social media policy to encourage public participation in long-term and/or short-term transportation planning.

11.4.2 TRAFFIC PREDICTION

Social media can also be used for long-term traffic prediction. A traffic prediction model was proposed by He et al.[12] where they used the traffic information from social media for a freeway network in San Francisco Bay area. They collected data using the Twitter streaming API with a geolocation filter. The filter contains a latitude/longitude bounding box of (−122.75, 36.934, −121.75, 38.369). The authors analyzed the correlation between traffic volume and tweet counts, and found a negative correlation between the web-based social media activity and the intensity of traffic activity

on the roads. Finally, the authors developed an optimization framework to extract relevant traffic indicators based on tweet semantics; results demonstrate that traffic prediction accuracy increases if the social media data are incorporated with the data derived from traffic sensors. Ni et al. [35] explored using social media data for the short-term freeway traffic flow prediction under special event conditions (i.e., sporting events). Both tweet rate features and semantic features were included in the prediction model. The authors found that the prediction results in models using both traditional traffic data (i.e., data collected from traffic detectors) and tweet features outperformed predictions employing only traditional traffic sensors. Incorporating tweet features in the model reduced the average root mean square error and mean absolute percentage error by about 24% and 7%, respectively.

11.4.3 TRAFFIC MANAGEMENT DURING PLANNED EVENTS

To assess travel demand related to public gatherings, Zhang et al. [4] conducted a study using Twitter hashtags to detect planned events. They studied tweets related to sporting events to predict the New York City subway passenger flow surrounding the events. Tweets were collected for baseball games during one season's 81 game days. Using the collected tweets, the method achieved a 98.3% precision rate for identifying the baseball games. To forecast traveler flow, the authors developed an ensemble model to leverage advantages from both Optimization and Prediction with Hybrid Loss Function (OPL) and Support Vector Regression (SVR). Findings from the analysis show that, compared to the individual performance of OPL and SVR, prediction accuracy and robustness further increases with the ensemble method employing both.

11.4.4 TRAFFIC MANAGEMENT DURING UNPLANNED EVENTS

Many studies have investigated social media data usage for traffic management during unplanned traffic events. For real-time incident information, Schulz and Ristoski [37] studied Twitter data using a semantic web technology (i.e., Linked Open Data/LOD Cloud) incorporating a machine learning approach. For tweet classification (they defined three classes including car crashes, shootings, and fires), the authors used features from LOD data and tweets. The resulting model achieved 89% accuracy for incident-related identification. Chen et al. [13] proposed a unified statistical framework for traffic congestion monitoring by combining a language model and hinge-loss Markov random fields. INRIX probe speed data sets and collected tweets gathered for Washington, DC, and Philadelphia were used, and their effectiveness was measured over a variety of spatial−temporal and other performance metrics. The study found that compared with traditional sensor data, social media data are not redundant. Sakaki et al. [42] considered social media as a social sensor and proposed a system to extract important event-related information (i.e., location and temporal information) from social media. Tweets were first classified into two groups—traffic-related tweets and nontraffic-related tweets. Eighty-seven percent precision was achieved when using tweets to identify heavy traffic conditions. In a different study, Gu et al. [5] used an adaptive data acquisition method to classify traffic-incident-related tweets. A dictionary of important keywords and their combinations was created after developing the adaptive data acquisition method to better identify traffic incidents.

To predict the impact of inclement weather on freeway operations, Lin et al. [43] applied linear regression models with the help of social media data. For the Buffalo−Niagara metropolitan area, the authors compiled weather data, Twitter data, and traffic information. The sensitivity and false alarm rate

were estimated against real-world weather data. The sensitivity value varied between 8% and 68.6% for four different data sets, and the false alarm rate varied between 1% and 18%. Following this, linear regression models were developed to predict the inclement weather impact on freeway speed. When the weather-related tweet variables were included, the linear regression models' accuracy was improved.

11.4.5 TRAFFIC INFORMATION DISSEMINATION

In addition to diagnosing traffic conditions, agencies have a need to disseminate important real-time information to the public. Wojtowicz and Wallace [36] studied agencies' use of social media for this task. The practices vary from one transportation agency to another. Several key factors can influence agencies' social media programs. These include the development of social media policy which includes the goals and objectives of the agency's use of social media data, proper message structure using standard language, and proper staffing and coordination with other agencies. Another study investigated how social media is used by transit agencies [44]. The authors conducted an online survey of 34 agencies, finding that agencies' most important goal is to use social media to communicate with current riders. This study suggested the use of captions and hashtags so that transit riders can easily identify the transit related information. This study also identified the barriers faced by the transit agencies, which include staff limitations and the time requirements to post updates, some riders' (including people with disabilities) lack of access to social media, and concerns that riders will criticize the agency via social media. Schweizter [45] found that Twitter posts reflect considerable dissatisfaction with transit service.

In a world that is increasingly connected with evolving technologies in communication techniques, sensors, and vehicles, transportation has become a more dynamic mobility system which integrates drivers, pedestrians, vehicles, infrastructures, etc. A mass wave of innovative concepts and booming technologies, such as ride sharing, vehicle sharing, autonomous vehicles and e-commerce, has hit the transportation system and shaped the way how people travel, shop, and commute. Beyond traditional traffic planning, operations and management, energy and emissions implications due to these mass changes have become more and more important. However, the vast range of energy implications due to uncertainty in vehicle miles traveled (VMT) require more research. Social media data could be utilized to analyze the trend of future travel and related systems impacts. Policy makers and business sectors could use social media as the front edge to promote adoption of more efficient travel modes, vehicle technologies, and sharing options.

11.5 FUTURE RESEARCH ISSUES/CHALLENGES FOR DATA ANALYTICS-ENABLED SOCIAL MEDIA DATA

11.5.1 SOCIAL MEDIA: A SUPPLEMENTAL TRANSPORTATION DATA SOURCE

In the future, the number and variety of sensor devices, such as different in-vehicle sensors or handheld mobile devices, will provide ever-increasing new sets of real-time data (e.g., on the locations, speeds, and directions of vehicles), which could be integrated with social media data to further improve the understanding of users' behaviors and real-time traffic conditions.

However, the need to protect social media users' sensitive data, while accessing less personal data, may compromise data mining outcomes. Thus, it is a major challenge to develop protection mechanisms that will help foster the widespread use of the social media data in concert with sensor data without compromising privacy.

11.5.2 POTENTIAL DATA INFRASTRUCTURE

Increasing penetration of social media data and innovations in text mining and analysis methods will leverage social media as a supplemental traffic data source. In previous work on social media analysis, the focus has been on the social media content itself. To the best of our knowledge, the issue of how to ingest and store social media data in transportation research has hardly been brought up for discussion. The challenge lies in the complex format of the tweets themselves. Tweets are collected in a raw JavaScript Object Notation (JSON) format containing hierarchical relationships among the attributes. A straightforward conversion to row-based tuples for a standard Structured Query Language (SQL) system is not possible. The only solution is to store tweets as special JSON-typed data blobs in either SQL or NoSQL systems.

Within the context of a research environment that emphasizes analytical study rather than data infrastructure, MongoDB is typically chosen as the go-to solution for storing tweets [46–48]. However, within the context of a production environment at transportation centers, additional components are necessary to ensure performance and resiliency, as is illustrated in Ref. [49]. Furthermore, restrictions imposed by Twitter will limit the number of connection points from which tweet streams can be collected. To support multiple usages of tweet contents, a mechanism must be devised to generate copies of tweets internally without impacting the analytical performance.

From a data storage and management perspective, the infrastructure for this work must be part of a comprehensive data infrastructure designed for connected transportation systems [50], as indicated by the components within the box with red dashes as shown in Fig. 11.2. While the original design calls for a sequential parsing and propagating of data across different components, this will create a delay in making data available for transportation applications based on service level agreement (SLA). To overcome this limitation, Rahman et al. presented a streaming data delivery approach [51]. With this approach, a message-oriented middleware component called Kafka [52] will be responsible for acquiring the initial Twitter stream and duplicating this stream in a parallel fashion. The resulting clones will be redirected toward different storage components, where different data-driven applications can be executed. This scheme is shown in Fig. 11.3.

The data infrastructure shown in Fig 11.3 is the backend to support a dynamic-filtering approach in working with online streaming data, including social media data. Its key novel features that are germane to the challenge of augmenting traffic incident information, yet are missing in the open domain tools, include:

1. Using an initial set of keywords provided by the user, data streams must be sampled in near-real time to dynamically suggest expansions of the original keyword set associated with emergent trends and terms. This will allow for dynamic data filtering that adapts to changes in topics and events driving primary sources such as social media data streams.
2. Developing a modular plug-in approach that facilitates keyword extraction from secondary data streams (such as content from news providers or RSS feeds), which allows finding relevant topics and keywords beyond the limitations of a primary data source.

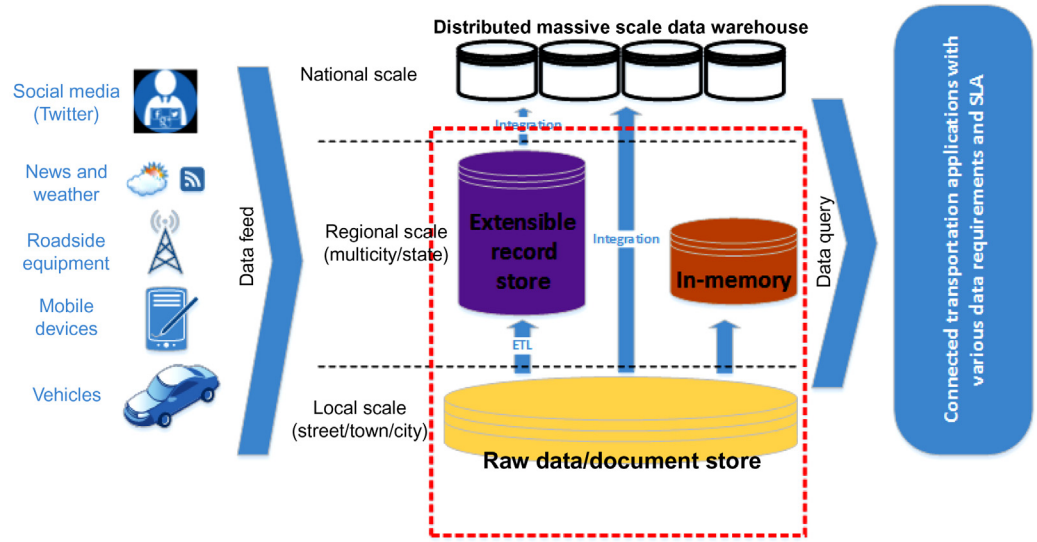

FIGURE 11.2

Data infrastructure for a connected transportation system.

3. Enabling fully automated as well as interactive keyword selection, to provide users with an easy-to-use data-filtering tool that permits "fine tuning" their queries and data collection.
4. Building a distributed stream management infrastructure to balance performance through dynamically provisioning hardware resources to the streaming and computing components. This may be done through either locally or remotely available resources.

Developing this tool, and the necessary supporting infrastructure, requires solving fundamental problems in Big Data. While topic detection and keyword optimization are well-understood research problems, combining the two produces unique algorithmic challenges. Solutions to these problems must be optimized to handle data from high-volume, high-velocity data streams. This further requires developing a computing infrastructure that is scalable and applicable across domains of inquiry.

While the management infrastructure depicted in Figure 11.3 is based upon the Lambda architecture for scalable real-time data systems [53], it may be required to deviate from the baseline model by combining the historical data in the batch layer into the final NoSQL database. This decision is based on two observations. First, the social media related research use cases focus on online streaming data rather than local repositories. Second, the proposed tool is targeted toward individual data-driven social researchers and research groups at institutions that may or may not have on-site administrative and technical support for large-scale cyberinfrastructure. Based on these observations, the developed infrastructure should have the following design choices:

1. The data infrastructure should not solely rely on existing static repositories. The data infrastructure should rely on the identification of traffic events that have just unfolded and assumes little to no prior information about these events. Instead, the stream-processing

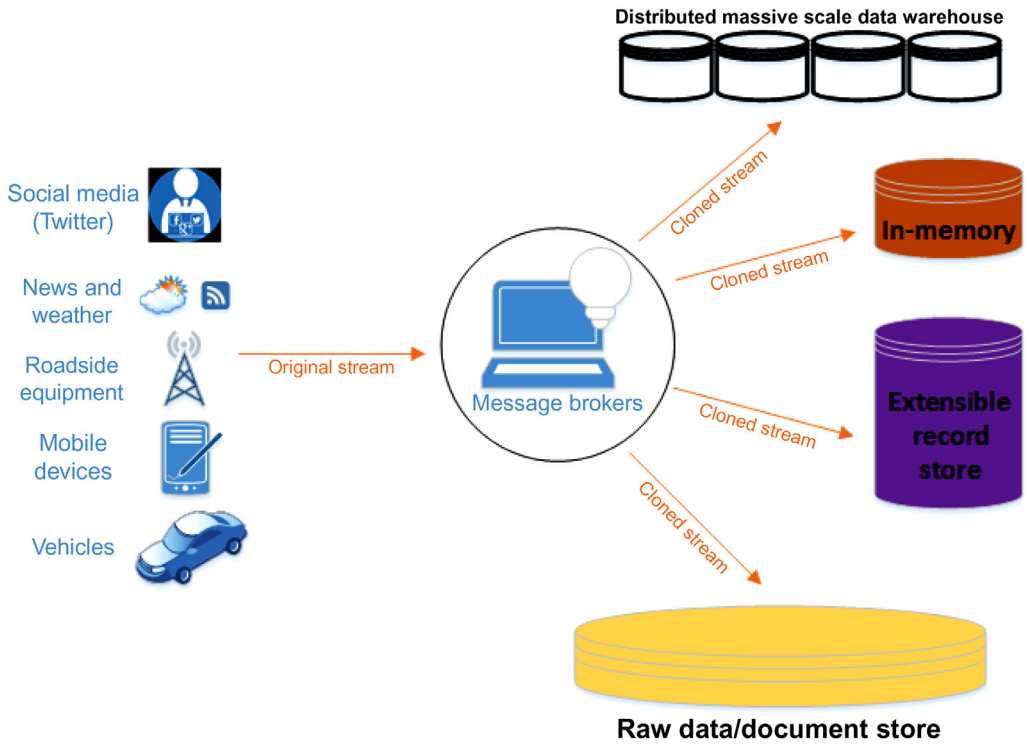

FIGURE 11.3

Streaming delivery approach for public agencies.

components will enable analyzing the global state of events through the integration of data streams from multiple sources over short bursts of time.

2. To enable use of the tool by a wide range of domain experts, many of whom might not be able to configure and maintain distributed computing resources, the computing infrastructure should be separated into interactive and distributed sections. The interactive infrastructure is targeted to fit on a single high-end computer, while the configuration and deployment of the distributed infrastructure is fully automated and can utilize either locally available or community resources.
3. The infrastructure should dynamically provision computing resources for the stream-processing component and the computation component. This allows the tool to support research areas that require access to different sets of information streams. Furthermore, as events unfold, new sources of information (e.g., an article linking to a new blog post) will become available to be streamed.
4. The tool should be designed and implemented with extensibility and modularity. That is, pluggable APIs should be designed and implemented for streaming data curation in order to support data integration and fusion from online sources. This allows the collaborating principal investigators leading the domain research areas, as well as any future community users, to participate in developing APIs for their data sources.

The envisioned architecture for the proposed dynamic distributed stream management infrastructure is illustrated in Fig. 11.4. By designing and implementing this infrastructure, the following issues can be addressed regarding the applicability of the infrastructure in the production environments of transportation departments:

1. *Evaluation and selection of the appropriate stream-processing engines (SPE)*: There exist a number of open source stream-processing systems [54–57]. However, there is no existing comprehensive evaluation and benchmarking of these tools in terms of performance, efficiency, scalability, and fault-tolerance, especially regarding the collection of diverse information streams. SPEs such as Samza, Spark Streaming, Storm, and S4 should be evaluated. A data streaming engine should be developed for benchmarking purposes, which can be used in research on synthetic Big Data generation.
2. *Integration of multiple SPEs*: While the performance evaluation study might provide a tentative ranking among the SPEs, it can be hypothesized that it will be necessary to combine multiple SPEs to take advantage of their possible strengths in terms of processing different information streams.
3. *Integration of infrastructure elasticity*: Previous work examines supporting elasticity within individual SPEs [58–60]. The infrastructure in Fig. 11.4 looks at elasticity at a different level of abstraction, which is to scale and balance resources for multiple instances of SPEs, as well as for the internal computing resources that analyze the integrated data coming from these different SPEs. Previous works on dynamic provisioning of Big Data infrastructure in a shared computing environment should be leveraged [61–62].

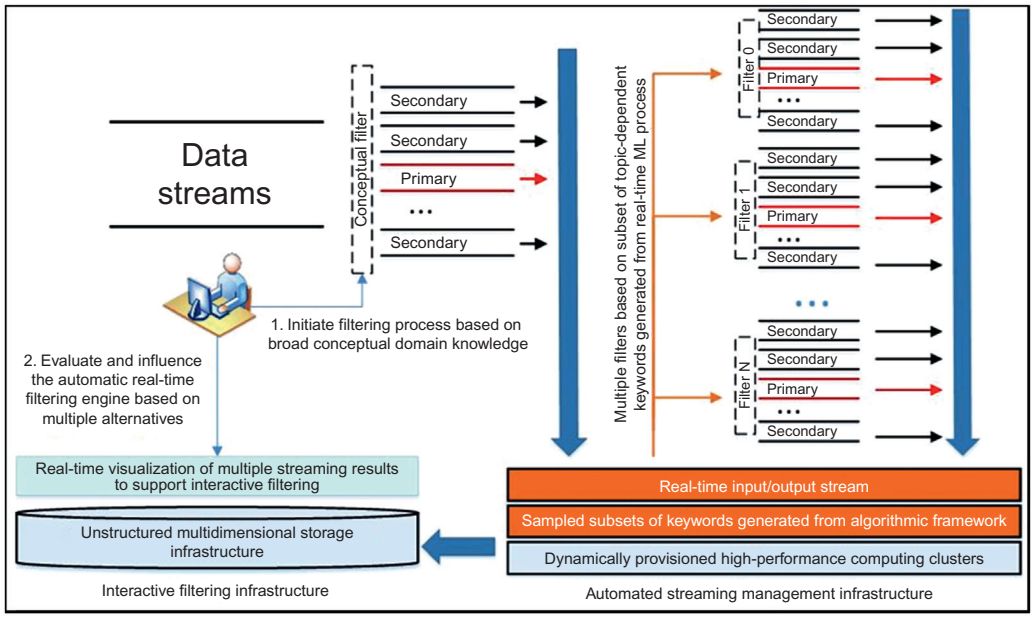

FIGURE 11.4

Dynamic distributed stream management infrastructure.

4. *Dynamic provision of computing resources from distributed locations*: Instead of limiting the elasticity to homogeneous resources (e.g., compute nodes within the same cluster or similar virtual machines on computing clouds), the infrastructure should be designed to support resource extension across geographically and administratively different locations.
5. *Data provenance*: To decrease the level of data duplication from different information streams and from the ML-based-filtering process, different approaches to storing the provenance of data through mechanisms such as data versioning and data index schemes should be investigated.

11.6 SUMMARY

Social media is allowing users to externalize their thoughts, ideas, and personal experiences, and share them with others. Emergent social media platforms, and rising engagement of citizens with social media, has created a unique opportunity for transportation agencies to collect traffic-related information from social media users with minimum resource investment. Transportation agencies are also using social media to disseminate information to users. This chapter provided a brief review of social media characteristics which determine the methods needed to extract the relevant information from the data generated across various social media platforms. Different data analytics algorithms will help to enable diverse methods of analyzing social media data to allow diverse applications of this data in transportation. Finally, future research issues and challenges, particularly in terms of data storage, processing, and accessibility, were outlined, which highlight that social media data, if incorporated with other streaming data, can be a potential reliable data source to augment roadway traffic-related data provided by public agencies. If the collected social media data (e.g., tweets) are not extensive, the public agencies can use a dynamic distributed stream management infrastructure to ingest and store the social media data with other traffic data.

11.7 CONCLUSIONS

Because of easy access to the Internet and high usage of devices such as tablets, smart phones, and laptop and desktop computers, different social media platforms with unique characteristics and user services have gained popularity among millions and even billions of users in last decade. These new data sources generate massive amounts of user-created heterogeneous contents and are becoming critical Big Data sources for understanding user behaviors in different domains such as business/product marketing and social trend analysis. There is a tremendous potential to apply these data to transportation planning, traffic incident detection, public outreach in transportation, and other transportation applications. Although substantial amounts of research have been conducted in last decades, the major challenge in developing efficient algorithms are: (1) development of analytics that ensure the privacy of data sources, (2) development of scalable algorithms that consider spatial and temporal dimensions, and (3) the creation of efficient computing platforms and algorithms for ever-increasing social media contents.

The future prospects for social media in terms of supplementing data needed for transportation planning, operations, and maintenance are promising. With an ever-increasing user base, data

available from social media will be more pertinent, frequent, and larger. It is thus important to prepare transportation planning and management for such useful and rich data sets. Planning and adopting the dynamic distributed social media data stream infrastructure discussed in the chapter will be key for such preparation.

11.8 EXERCISE PROBLEMS

1. What are the characteristics of social media data?
2. Using Twitter streaming API, collect tweets by tracking the following terms for 1 hour:
   a. "Traffic"
   b. "Accident"
   c. "Accident" and "I26"
   d. "Accident" or "I26"
3. Based on the tweets collected in Question 2, answer the following questions:
   a. What is the rate of arrivals for the tweets during (1) peak hours and (2) nonpeak hours?
   b. How many unique tweet handles are there?
   c. How many handles can be considered as "official sources" (e.g., public agencies like departments of transportation) or major media sources like news channels?
   d. Identify several events (traffic incidents), then calculate the number of retweets for all the handles that tweet about those events. Is there a difference between the official and nonofficial handles? Are there any cases where a nonofficial handle tweets about an event before an official handle does?
   e. Select a traffic incident from the collection and construct the propagation networks with the roots as the official handles. The rate of propagation is calculated as the number of additional Twitter handles who become aware of the incident either directly through the official handles or indirectly through retweets. Describe an approach for calculating the rate of propagation (how fast the information about the incident is being spread across the social network). Perform the calculation on various official handles and compare the results.
4. Do the collected tweets in Question 2 qualify as "Big Data"?
5. Identify the methods of social media data analysis. Which analysis method can be applied for the tweets collected for Question 2?

REFERENCES

[1] J. Poushter, Smartphone ownership and Internet usage continues to climb in emerging economies, <http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-Internet-usage-continues-to-climb-in-emerging-economies/>, 2016 (accessed 27.09.16).
[2] M. Adedoyin-Olowe, M. Mohamed, F. Stahl, A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617, 2013.
[3] P. Gundecha, H. Liu, Mining social media: a brief introduction, Tutor. Oper. Res. 1 (4) (2012) 1–17.

[4] Z. Zhang, M. Ni, Q. He, J. Gao, Mining transportation information from social media for planned and unplanned events, <https://www.buffalo.edu/content/www/transinfo/Research/socialmediaminingforevents/_jcr_content/par/download/file.res/MiningSocialMediaEvents_FinalReport.pdf>, 2016 (accessed 27.09.16).
[5] Y. Gu, Z.S. Qian, F. Chen, From Twitter to detector: real-time traffic incident detection using social media data, Transport. Res. C Emerg. Technol. 67 (2016) 321–342.
[6] S. Bregman, Uses of social media in public transportation, Transport. Res. Board 99 (2012).
[7] L. Manovich, Trending: the promises and the challenges of big social data, Debates in the digital humanities 2 (2011) 460–475.
[8] S. Kim, Twitter's IPO filing shows 215 million monthly active users, <http://abcnews.go.com/Business/twitter-ipo-filing-reveals-500-million-tweets-day/story?id = 20460493>, 2013 (accessed 27.09.16).
[9] K. Lantz, S. Khan, L.B. Ngo, M. Chowdhury, S. Donaher, A. Apon, Potentials of online media and location-based big data for urban transit networks in developing countries, Transport. Res. Record J. Transport. Res. Board 2537 (2015) 52–61.
[10] Twitter streaming API, <https://dev.twitter.com/overview/api> (accessed 27.08.16).
[11] C.C. Aggarwal, Social Network Data Analytics, Springer, New York, NY, 2011.
[12] J. He, W. Shen, P. Divakaruni, L. Wynter, R. Lawrence. Improving traffic prediction with tweet semantics, in: The International Joint Conference on Artificial Intelligence, 2013.
[13] P.T. Chen, F. Chen, Z. Qian, Road traffic congestion monitoring in social media with hinge-loss Markov random fields, in: 2014 IEEE International Conference on Data Mining, 2014, pp. 80–89.
[14] Z. Quan, Real-time incident detection using social media data, <http://www.dot7.state.pa.us/BPR_PDF_FILES/Documents/Research/Complete%20Projects/Operations/Real_time_Incident_Detection_Using_Social_Media_Data.pdf>, 2016 (accessed 20.09.16).
[15] P. Gloor, J. Krauss, S. Nann, K. Fischbach, D. Schoder, Web science 2.0: identifying trends through semantic social network analysis. Comput. Sci. Eng. 4 (2009) 215–222.
[16] Clemson University Social Media Listening Center, <http://www.clemson.edu/cbshs/departments/smlc/contact/index.html>, 2016 (accessed 01.11.16).
[17] E.-A. Baatarjav, S. Phithakkitnukoon, R. Dantu, Group recommendation system for Facebook. On the Move to Meaningful Internet Systems (2008), Springer Berlin Heidelberg, 211–219.
[18] D. Zhou, I. Councill, H. Zha, C. Giles. Discovering temporal communities from social network documents, in: Seventh IEEE International Conference on Data Mining, 2007, pp. 745–750.
[19] L. Tang, H. Liu, Toward collective behavior prediction via social dimension extraction, IEEE Intell. Syst. 25 (4) (2010) 19–25.
[20] N. Agarwal, H. Liu. Modeling and data mining in blogosphere, volume 1 of Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan and Claypool, 2009.
[21] P.K. Akshay Java, T. Oates, Modeling the spread of influence on the blogosphere. Technical Report UMBC TR-CS-06-03, University of Maryland Baltimore County, Baltimore, MD, 2006.
[22] S. Agrawal, S. Chaudhuri, G. Das. DBXplorer: a system for keyword based search over relational databases, in: ICDE Conference, 2002.
[23] V. Hristidis, Y. Papakonstantinou. Discover: keyword search in relational databases, in: VLDB Conference, 2002.
[24] S. Chakrabarti, B. Dom, P. Indyk. Enhanced hypertext categorization using hyperlinks, in: ACM SIGMOD Conference, 1998.
[25] D. Bortner, J. Han. Progressive clustering of networks using structure-connected order of traversal, in: ICDE Conference, 2010.
[26] N. Mishra, R. Schreiber, I. Stanton, R.E. Tarjan, Finding strongly-knit clusters in social networks, Internet Math. 5 (1–2) (2009) 155–174.
[27] Citisense, <https://citisense.com/> (accessed 09.30.16).
[28] R. Ganti, N. Pham, H. Ahmadi, S. Nangia, T. Abdelzaher, GreenGPS: A Participatory Sensing Fuel-Efficient Maps Application, Mobisys, San Francisco, CA, 2010.

[29] INRIX, <http://inrix.com/> (accessed 01.11.16).
[30] T. Choudhury, B. Clarkson, S. Basu, A. Pentland. Learning communities: connectivity and dynamics of interacting agents, in: International Joint Conference on Neural Networks, 2003.
[31] C.C. Aggarwal, H. Wang (Eds.), Managing and Mining Graph Data, Springer, New York, NY, 2010.
[32] M. Adedoyin-Olowe, M. Mohamed, F. Stahl, A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617, 2013.
[33] P. Gundecha, H. Liu, Mining social media: a brief introduction, Tutor. Oper. Res. 1 (4) (2012) 1–17.
[34] J. He, S. Wei, D. Phani, L. Wynter, R. Lawrence. Improving traffic prediction with Tweet semantics, in: The International Joint Conference on Artificial Intelligence, 2013.
[35] M. Ni, Q. He, J. Gao, Using social media to predict traffic flow under special event conditions, in: The 93rd Annual Meeting of Transportation Research Board, 2014.
[36] J. Wojtowicz, W.A. Wallace, The use of social media by transportation agencies for traffic management, in: Transportation Research Board 95th Annual Meeting, no. 16-6217, 2016.
[37] A. Schulz, P. Ristoski, The car that hit the burning house: understanding small scale incident related information in microblogs, in: Seventh International AAAI Conference on Weblogs and Social Media, 2013.
[38] S.R. Majumdar, The case of public involvement in transportation planning using social media, in: Transportation Research Board 95th Annual Meeting, no. 16-2604, 2016.
[39] American Planning Association, Planning and Urban Design Standards. Wiley Graphic Standards, John Wiley & Sons, Hoboken, NJ, 2006.
[40] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Election forecasts with Twitter: how 140 characters reflect the political landscape, Social Sci. Comp. Rev. (2010) 1–17.
[41] Tweetminster. Is word-of-mouth correlated to general election results? The results are in, <http://www.scribd.com/doc/31208748/Tweetminster-PredictsFindings>, 2010 (accessed 25.06.10).
[42] T. Sakaki, Y. Matsuo, T. Yanagihara, N. Chandrasiri, P. Naiwala, K. Nawa, Real-time event extraction for driving information from social sensors, in: 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012, pp. 221–226.
[43] L. Lin, M. Ni, Q. He, J. Gao, A.W. Sadek, Modeling the impacts of inclement weather on freeway traffic speed: exploratory study with social media data, Transport. Res. Record J. Transport. Res. Board (2482) (2015) 82–89.
[44] B. Susan, Uses of social media in public transportation, vol. 99. Transportation Research Board, 2012.
[45] L. Schweitzer, Planning and social media: a case study of public transit and stigma on Twitter, J. Am. Plan. Assoc. 80 (3) (2014) 218–238.
[46] K. Fu, C.L. Yen, L. Chang-Tien, TREADS: a safe route recommender using social media mining and text summarization, Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2014, 14.
[47] M. Liu, K. Fu, C.T. Lu, G. Chen, H. Wang, November. A search and summary application for traffic events detection based on twitter data, Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2014, 18.
[48] K. Fu, C.T. Lu, R. Nune, J.X. Tao, Steds: social media based transportation event detection with text summarization, in: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015, pp. 1952–1957.
[49] S. Barahmand, S. Ghandeharizadeh, J. Yap, A comparison of two physical data designs for interactive social networking actions, Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM, 2013.
[50] K. Lantz, S. Khan, L.B. Ngo, M. Chowdhury, S. Donaher, A. Apon, Potentials of online media and location-based big data for urban transit networks in developing countries, Transport. Res. Record J. Transport. Res. Board (2537) (2015) 52–61.

[51] M. Rahman, Y. Du, L. Ngo, K. Dey, A. Apon, M. Chowdhury, An innovative way to manage data for connected vehicle applications, in: The 95th Transportation Research Board Annual Meeting, 2016.
[52] J. Kreps, N. Narkhede, J. Rao, Kafka: a distributed messaging system for log processing, in: Proceedings of the NetDB, 2011, pp. 1–7.
[53] N. Marz, J. Warren, Big Data: principles and best practices of scalable realtime data systems, Manning Publications Co. Greenwich, CT, USA, 2015.
[54] R. Ranjan. Streaming big data processing in datacenter clouds, 78–83. <http://doi.ieeecomputersociety.org/10.1109/MCC.2014.22>. (accessed 31.07.16).
[55] S.G. Kamburugamuve, D.L. Fox, J. Qiu, Survey of distributed stream processing for large stream sources, Technical report. 2013, Available at <http://grids.ucs.indiana.edu/ptliupages/publications/survey_stream_processing.pdf>. (accessed 31.07.16).
[56] M. Gorawski, A. Gorawska, K. Pasterak, A survey of data stream processing tools, Information Sciences and Systems, Springer, New York, NY, 2014.
[57] J.W. Anderson, Kennedy K., Ngo L.B., Luckow A., Apon A.W. Synthetic data generation for the Internet of things, in: 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 171–176.
[58] V. Gulisano, R. Jimenez-Peris, M. Patino-Martinez, C. Soriente, P. Valduriez, Streamcloud: an elastic and scalable data streaming system, in: IEEE Transactions on Parallel and Distributed Systems, 23 (2012) 2351–2365.
[59] R. Tolosana-Calasanz, Bañares, J.Á., Pham, C. and Rana, O.F. Resource management for bursty streams on multi-tenancy cloud environments, Fut. Gener. Comput. Syst., 2015.
[60] T. Heinze, Z. Jerzak, G. Hackenbroich, C. Fetzer, Latency-aware elastic scaling for distributed data stream processing systems, in: Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, 2014, pp. 13–22.
[61] W.C. Moody, L.B. Ngo, E. Duffy, A. Apon, Jummp: job uninterrupted maneuverable mapreduce platform, in: 2013 IEEE International Conference on Cluster Computing (CLUSTER), 2013, pp. 1–8.
[62] L.B. Ngo, M.E. Payne, F. Villanustre, R. Taylor, A.W. Apon, Dynamic provisioning of data intensive computing middleware frameworks: a case study, in: The Workshop on the Science of Cyberinfrastructure: Research, Experience, Applications and Models (SCREAM-15), 2015.