ARTICLE IN PRESS

Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Review

Sensing and detecting traffic events using geosocial media data: A review

Shishuo Xu[a,b], Songnian Li[a,*], Richard Wen[a]

[a] Department of Civil Engineering, Ryerson University, 350 Victoria St., Toronto, ON M5B 2K3, Canada
[b] School of Environment Science and Spatial Informatics, China University of Mining and Technology, No. 1 Daxue Road, Xuzhou, Jiangsu 221116, China

ARTICLE INFO

Keywords:
Traffic event
Event detection
Geosocial media
Twitter data stream

ABSTRACT

Social media platforms, or social networks, have allowed millions of users to post online content about topics related to our daily lives. Traffic is one of the many topics for which users generate content. People tend to post traffic related messages through the ever-expanding geosocial media platforms. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable traffic related information, which can be mined to extract traffic events to enable users and organizations to acquire actionable knowledge. A great number of literature has reported on the methods developed for detecting traffic information from social media data, especially geosocial media data when geo-tagged. However, a systematic review to synthesize the state-of-the-art developments is missing. This paper presents a systematic review of a wide variety of techniques applied in detecting traffic events from geosocial media data, arranged based on their adoption in each stage of an event detection framework developed from the literature review. The paper also highlights some challenges and potential solutions. The aim of the paper is to provide a structured view on current state-of-art of the geosocial media based traffic event detection techniques, which can help researchers carry out further research in this area.

1. Introduction

Traffic events, including traffic jams, roadworks, road closures, traffic accidents, and bad weather conditions, pose great challenges to both drivers and traffic management agencies (Gutiérrez, Figuerias, Oliveira, Costa, & Jardim-Goncalves, 2015). Identifying the time, location and type of traffic event in a real-time manner is important for drivers and traffic managers to generate proactive operation strategies to improve traffic conditions (Fu, Lu, Nune, & Tao, 2015; Fu, Nune, & Tao, 2015; Gu, Qian, & Chen, 2016).

Traditional methods applied to detect traffic events mainly focus on measuring traffic speed, traffic density and traffic flow using a wide variety of physical sensors (e.g., imaging sensor, inductive loop, magnetic sensor, acoustic detector, and passive infrared), which are usually installed at fixed locations along roads. They are implicitly embedded with the assumption that significant changes in flow characteristics immediately follow the traffic events (Gu et al., 2016). Guralnik and Srivastava (1999) and Ihler, Hutchins, and Smyth (2006) detected traffic events using the data collected from loop detectors by adopting time series algorithms. This study indicated that the proposed approach performed significantly better than a non-probabilistic threshold-based technique. The accuracy of loop detector data was further investigated by Coifman and Dhoorjaty (2004) using eight detector validation tests, which contrasted the performance of different sensor models and identified hardware problems to correct errors in the loop detector data. A novel event-driven architecture was proposed to deal with the continuous events created by sensors (Dunkel, Fernández, Ortiz, & Ossowski, 2011; Terroso-Sáenz, Valdés-Vela, Sotomayor-Martínez, Toledo-Moreo, & Gómez-Skarmeta, 2012). In terms of video image processors, spatio-temporal Markov random field algorithm (Kamijo, Matsushita, Ikeuchi, & Sakauchi, 2000) and kalman filtering-based approach (Veeraraghavan, Schrater, & Papanikolopoulos, 2005) were developed to automatically monitor traffic scenes and detect accidents at intersections. Li and Porikli (2004) and Porikli and Li (2004) detected highway traffic events in different illumination conditions (i.e., sunny, cloudy, and dark) based on an unsupervised, low-latency traffic congestion estimation algorithm.

However, it costs a lot to install physical sensors at a large scale to sense the traffic dynamics of the city. For example, as reported by Leduc (2008), the average cost to install and maintain an inductive loop detector at an intersection ranged from $9500 to $16,700 annually. Usually, such sensors are sparsely located along major highways other than local arterials. Traffic events may take place anywhere at any time (e.g., car crash on arterial road). Thus, the physical sensor-based method may not be an efficient way to timely detect traffic events due to its high cost, sparsity and limited spatial coverage. Crowdsourcing,

* Corresponding author.
E-mail address: snli@ryerson.ca (S. Li).

https://doi.org/10.1016/j.compenvurbsys.2018.06.006
Received 14 January 2018; Received in revised form 3 May 2018; Accepted 19 June 2018

which refers to a low-cost process of solving a problem through obtaining contributions from a large group of people via online communities (Doan, Ramakrishnan, & Halevy, 2011; Howe, 2006), is likely to be an available solution.

With the wide use of smart phones and mobile devices, crowdsourcing becomes a promising alternative approach to feasibly collect traffic related data with spatial and temporal information (Zheng et al., 2016). These alternative information sources include user-shared traffic information, vehicle and individual GPS trajectories, Bluetooth data, and cellular network, which provide traffic information implicitly through the users' movements. WAZE (Waze Mobile, 2018) is a real-time traffic monitoring and navigation system, which alerts users of abnormal traffic conditions and gives them the best route based on user shared reports of pre-defined categories (e.g., accidents, road hazards, and traffic jams). GPS trajectories data is another typical crowdsourcing information to detect traffic events. Kamran and Haas (2007) utilized the real-time GPS data collected from vehicles to identify the abnormal traffic pattern on motorways based on a multilevel approach. A hierarchical analysis was further conducted to determine the precise location of traffic events. Similarly, INRIX data, which were obtained from individual trajectories, was also used to recognize occurrence of traffic events through a Bayesian structure equation model (Park & Haghani, 2016). Furthermore, a mobile application, named WreckWatch, polled smartphone system sensors (e.g., GPS receiver and accelerometers) and context data (e.g., speed) to automatically detect traffic events (White, Thompson, Turner, Dougherty, & Schmidt, 2011). The emergency notifications were sent to the first responders through WreckWatch to improve situational awareness. Martchouk, Mannering, and Bullock (2011) measured the travel time variability due to adverse weather and traffic breakdown by collecting Bluetooth probe data on freeway segments in Indianapolis. Demissie, de Almeida Correia, and Bento (2013) analyzed the correlation between the cellular networks handover counts and traffic volumes to reveal traffic status by training an Artificial Neural Network (ANN). Such cellular based monitoring contains high measurement errors for the accuracy that are usually proportional to cell size. It was reported that the median error of cellular network positioning was up to 600 m (Zandbergen, 2009). The previously mentioned crowdsourcing data often belong to private operators and the quality of data sharing is often a challenging issue due to the privacy matters. Besides, they do not necessarily enable real-time data processing, which is required for traffic event detection.

Geosocial media platforms allow users to compose and post short statements about their perceptions and/or experiences with geolocations, which plays an increasingly important role in our daily lives (Kelley, 2013). The free-cost features enable users to easily share a variety of information to the public, including photos, video, and blogs, with more spatial and temporal overage compared to physical sensors (Kaplan & Haenlein, 2010). For example, Facebook, Twitter and Weibo (a Chinese version of Twitter), usually have hundreds of millions users, which can generate a large amount of posts attached with timestamps, geolocations, and text contents. In other words, wherever there is a user there is a potential for geosocial media data. Further, geosocial media data can be used to not only identify when and where traffic anomalies take place (i.e., traffic pattern), but also explain the reasons behind the traffic anomalies in a real-time manner due to the abundant semantics of geosocial media content. This provides a significant advantage of geosocial media data over GPS data in detecting traffic patterns (Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017), especially in detecting traffic events rather than condition along a road segment. Therefore, it is likely to be an effective way to extract useful information from these abundant messages to detect traffic events.

Specially, Twitter is one of the most popular geosocial media sites all over the most part of the world, from which many studies have been done on detecting traffic information. It provides a free approach to acquire public tweets through open API, such as REST APIs and Streaming APIs (Twitter Inc., 2018). The Twitter REST APIs provide the ability to search by certain keywords or accounts from sample of recent tweets published in the past 7 days. Streaming APIs enable developers to collect real-time tweets with a set of bounding boxes or comma-separated list of phrases. The acquired tweets are often tagged with a pair of longitude and latitude coordinates (if the location-based service is turned on), a timestamp, and a short message limited to 140 characters (280 after November 7, 2017). Recently, Twitter has been adopted as a powerful data source to detect disasters (Dingli, Mercieca, Spina, & Galea, 2015; Kryvasheyeu et al., 2016; Sakaki, Okazaki, & Matsuo, 2010), predict election results (Metaxas & Mustafaraj, 2012) and crimes (X. Wang, Gerber, & Brown, 2012; Zhao, Chen, Lu, & Ramakrishnan, 2015), spread breaking news (Amer-Yahia et al., 2012; Phuvipadawat & Murata, 2010; Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009) and identify small-scale geosocial events (R. Lee & Sumiya, 2010; Watanabe, Ochi, Okabe, & Onai, 2011), which have happened in the real world.

Traffic is also a popular topic people would like to discuss in their daily lives (Novaco & Gonzalez, 2009). Thus, they tend to post traffic related information via social media networks when there is an accident, car crash, roadwork, or road closure. Geosocial media, especially Twitter, proves to be a valuable source in generating a wide range of traffic related information to detect traffic events to support traffic planning and management (Gal-Tzur et al., 2014; Gal-Tzur, Grant-Muller, Minkov, & Nocera, 2014; Grant-Muller et al., 2014, 2015). An exploratory study conducted in Northern Virginia, which analyzed the correlation between traffic patterns and traffic-related Twitter concentration from a spatiotemporal perspective, revealed that 77.4% of traffic-related Twitter concentrations could be justified by local traffic surge (Zhang et al., 2016). Another similar study showed that tweets tended to be posted within 5-h of the event that they referred to, and were most often sent between 10 and 25 miles of the event's location (Mai & Hranac, 2013). Zhang, Tang, Wang, and Wang (2015) indicated that the spatial distribution of traffic related tweets were clustered mostly within 800 m around traffic incidents in Seattle downtown area. These previous studies have proved that geosocial media data is a valuable data source to detect traffic events.

Up to now, various Twitter based applications have been developed to detect traffic events in a cost-effective way. For example, both Steds (Fu, Lu, et al., 2015) and Butterfly (Fu, Zhong, Lu, & Boedihardjo, 2015) were proposed as novel query expansion methods based on apriori algorithm for extracting traffic related tweets, which were then ranked to better summarize the detected events. TEDS (Liu, Fu, Lu, Chen, & Wang, 2014) adopted spatio-temporal analysis and a wavelet analysis model for traffic events detection. STAR (Semwal, Patil, Galhotra, Arora, & Unny, 2015) analyzed the relationship between co-occurring problems and their causes to train a classifier to predict severe problems for the next day. TrafficWatch (Nguyen, Liu, Rivera, & Chen, 2016), Traffic Observatory (Ribeiro Jr. et al., 2012) and TEDAS (R. Li, Lei, Khadiwala, & Chang, 2012) followed a process of preprocessing tweets to create tokens, identifying traffic related tweets using classification methods, and geocoding them to determine the exact location of events. Moreover, semantic web technologies were applied to interpret the underlying reasons behind traffic events in Dub-Star (Daly, Lécué, & Bicer, 2013) and STAR-CITY (Lécué et al., 2014).

However, applying geosocial media data to detect traffic events still faces many challenges. For example, Twitter messages are restricted in length and written by anyone. Therefore, tweets include large amount of informal, irregular, and abbreviated words, as well as a large number of spelling and grammatical errors, improper sentence structures, and mixed languages. In addition, Twitter streams contain large amount of meaningless messages (Hurlock & Wilson, 2011), polluted content (K. Lee, Eoff, & Caverlee, 2011), and rumors (Castillo, Mendoza, & Poblete, 2011), which negatively affect the performance of the detection algorithms.

Existing studies (Atefeh & Khreich, 2015; Bontcheva & Rout, 2014; Garg & Kumar, 2016; Goswami & Kumar, 2016; Hasan, Orgun, &