Detect Hidden Road Hazards combining Multiple Social Media Data

Fang Jin  
Department of Computer Science  
Texas Tech University  
fang.jin@ttu.edu

Hongchao Liu  
Department of Civil, Environmental, and Construction Engineering  
Texas Tech University  
hongchao.liu@ttu.edu

Abstract—It is imperatively important to spot hidden road hazards which cause a high proportion of traffic incidents. Social media data is featured by its innumerable and up-to-date information, and provides a promising approach to road hazard spotting. However, this research area is not well studied yet. In this paper, we present our view of the important research issues, including challenges of mining spatio-temporal dataset, road hazards reasoning and etc.

I. VISION

For decades, government, industry and research have been dedicated to building an unhindered road system, improving driving safety, and minimizing traffic accidents. However, even with considerate planning and careful road design, there are still tons of accidents happening every second all over the world. While driver errors such as speeding, distracted driving and drunk driving are among the leading causes of automobile accidents, dangerous road conditions are also a significant contributor.

Spotting those hazardous road conditions timely, and taking suitable evasive action to avoid crashes is critically important to public traffic safety. Various government agencies are responsible for maintaining roads and ensuring they are safe for drivers, cyclists and pedestrians, including providing proper signage, lighting, pavement markings, signals and traffic control devices. They generally discover dangerous road conditions through individuals reporting or by conducting regular surveys of the roadways. Those traditional approaches play great roles to the road construction and public traffic safety. However, it is also suffering several disadvantages: (1) Lots of people do not report the dangerous conditions even they find some either because they are too busy to report or they do not think it really necessary to report. (2) Regular survey usually means physically driving testing, and is often not cost-effective. (3) Surveyors are usually familiar with the road conditions, and are hard to be aware of dangerous conditions which most driver will face. With the emergence of social media platforms, people develop the habit of posting their comments and moments to social media platforms. One of the important topics is their daily traffic information, themes such as traffic congestion, complains about traffic conditions, witnessed accidents, or horrible driving experiences, and even the true reason for traffic accidents. Those causal posts provide a good opportunity to expose public opinions regarding road design, hidden hazards, accidents causality, criminals escape and other critical issues that impact traffic safety. More importantly, with the combination of multiple information sources, text messages, photos, and videos, through multiple devices such as sensors, cameras, lasers, with supplemental data such as weather data, pavement conditions, news report, and other related broadcast, it is possible to identify the top accidents locations and major reasons behind, thus to provide valuable and feedback or suggestions to transport planning and management, performance measurement and quality evaluation.

The main concern of using social media data is whether social media can act as a credible information sources of sufficient quality to meet the needs of transport planners, operators, policy makers and all travelling public [1]. In other words, how to detect and identify the road hazards from the extremely noisy social media data. The second main concern is how to fuse the massive heterogeneous dataset appropriately, which involves large scale spatial temporal database design. In addition, it is also a challenge to integrate the refined text mining, geocoding, data fusion, and causality inference technology seamlessly. In the following section, we will discuss these research challenges in details.

II. CHALLENGES

There are major technical and societal challenges to detect and identify hidden road hazards from social media accurately.

a) Data screening: The overall world is generating large amount of posts, news, videos every second. Take Facebook for example, updated by October 2018, every 60 seconds, 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded on average ¹. The first non-trivial task is how to develop a sustainable and scalable infrastructure to collect and store multiple heterogeneous data. This is because of the complex data heterogeneities, which include text, photos, broadcast video, cameras, road maps, driving speed, and traffic networks.

The second non-trivial task is how to develop screen algorithms which are able to extract useful information regarding road safety as well as remove unrelated noisy messages. This task also hinges on how we consider a message is road safety related or not. In [2], they propose an unsupervised

dynamic query expansion algorithm to iteratively expand domain-related terms for Twitter data streams. This approach will help filter out most dominant traffic related Tweets when using a set of human selected traffic keywords as primary query seeds. However, in the road hazard detection study, Tweets discussion regarding one specific road section could be sparse which might severely affect the training process.

The third question is labeling, especially the dataset is heterogeneous and spanning multiple spatial and temporal granularities. It would be ideal to have transportation experts to collect all the related posts, conversations, reports, download all the traffic pictures, road situation broadcasts, and then discern if they are road hazards related. If yes, label each message, photo, video to one specific road location/section, and remove the redundant pieces, discard controversial and nonsense complains. Suppose we have thousands of millions of expert working on this, we can guarantee the data screening results would be detailed and accurate. However, it will be challenging to develop algorithms to map patches of information into specific locations and events.

b) Geocoding Tweets conversation: To resolve location expression from Tweets posts to the correct physical locations, is complicated by the fact that: (1) only a small portion of Tweets carry latitude/longitude coordinates, and they are usually tweeted from geo-tagging enabled smart phones. (2) some Twitter accounts are labeled with location information such as city, county, and sometimes finer-grained school name and address information from their profiles. Unfortunately, this type of location information generally has very weak correlation with the locations of some road hazards [3] that they were mentioning. Specifically, people might post a car accident, but this car accident location is totally different from the location on the users profiles. (3) The availability of data sources containing transport-related entities (e.g. road names, station formal and informal names, names of parking facilities) may constitute a valuable asset for identifying locations for transport-related messages [1]. However, directly using these road names, points of interest and etc., may be problematic since these names are usually not unique, and there are a lot of places around the world share the same name. This scenario of correct name resolution is even more complicated considering the Twitter message [4] is short (less than 280-character) [4]. Bry at al. [5] started the research to develop a world model for geospatial reasoning that can be used for SemanticWeb applications such as road crossings. Resolutions of geospatial locations closely depend on the level of detail provided.

c) Hazard reasoning: Road hazard reasoning lies on the following aspects. The first task is about knowledge representation. One approach would be using a neural networks (NN) to trigger another module once the NN recognizes a pattern precursor. Detecting critical information and triggering road hazard reasoning is complicated because it must invoke symbolic processors accurately and efficiently. This is in fact a problem that intermixes NNs and symbolics engines. Theoretically, there are innumerous reasons for road hazards, e.g., invisible stop signs, confusing road signs, wildlife crossing and etc. Sophisticated reasoning analysis is difficult considering such a vast reason space. Reasoning is often contextual dependent, and is often non-linear. A deduction is justifiable under one set of circumstances may be completely wrong in a different scenario, and is often based on incomplete information, or even conflicting information. For the countless pieces of jigsaw puzzle coming from multiple heterogeneous data sources, designing algorithms to discover the complete story for road hazards, or identifying the full picture for each road hazard is challenging. Storytelling algorithm was proposed in [6], aims to explicitly relate object sets that are disjoint by finding a chain of (approximate) redescriptions between the sets. Using physiological and behavioral measurements in a picture-based road hazard perception experiment to classify risky and safe drivers was proposed in [7]. Considering there are so many adversary circumstances, like incomplete information, conflicting information, accurate reasoning algorithms for road hazard detection using social media data is deteriorating challenging.

III. CONCLUSIONS

The incorporation of aggregated social media data such as Twitter, Facebook, Flicker, and Youtube will allow a flourishing of service designed around to enhance traffic safety. One of the most critical but has not yet fully exploited area is road hazard detection. The advantages of social media sources provide the opportunity to capture users driving experience, preserve elements of the road hazards associated context, extract related comments and evidence, garner qualitative data on large scale, and gather valuable information costless. New avenues of research would open efficient response to reliable data source collection, heterogeneous data fusion, accurate locations or road sections confirmation, and major concerns of road hazards detection. We believe that the spatio-temporal social media data mining work will provide a valuable feedback to transport planning and management, and further avoid potential car accidents.

REFERENCES