Identifying and Prediction of Events of Critical Public Domain using Social Sensor Big Data

Samatha p.k[1], Prof. Dr.Mohamed Rafi[2]

[1,2]*Department of studies in computer science & Engineering ,University BDT College of Engineering (A constuent college of VTU, Belgavi), Davangere, Karnataka-577001,India*

Abstract- **Public infrastructure systems provide many of the services that are critical to the health, functioning, and security of society. Many of these infrastructures, however, lack continuous physical sensor monitoring to be able to detect failure events or damage that has occurred to these systems. We propose the use of social sensor big data to detect these events. We focus on two main infrastructure systems, transportation and energy, and use data from Twitter streams to detect damage to bridges, highways, gas lines, and power infrastructure.**

Through a three-step filtering approach and assignment to geographical cells, we are able to filter out noise in this data to produce relevant geo-located tweets identifying failure events. Applying the strategy to real-world data, we demonstrate the ability of our approach to utilize social sensor big data to detect damage and failure events in these critical public infrastructures.

Index terms- Social Sensors, Big Data, Data Processing, Critical Infrastructure, Event Detection

INTRODUCTION

This includes energy systems that power nearly all devices, controls, and equipment, as well as transportation systems that enable the movement of people and goods across both short and long distances. Failure of or damage that has occurred to these infrastructures, whether from deterioration and aging, or from severe loads due to hazards such as natural disasters, poses significant risks to populations around the world.

Detecting these damage or failure events is critical both to minimize the negative impacts of these events, e.g., by rerouting vehicles away from failed bridges, and to accelerate our ability to recover from these events, e.g., by locating the extent of power outages for deployment of repair crews. Many of these infrastructures, however, lack continuous physical sensor monitoring to be able to detect these damage or failure events. Bridges, for example, are generally subject to only yearly inspections, and very few are instrumented with physical sensors that would be able to detect damage that may occur at any time. In addition, infrastructures that contain monitoring capabilities, such as energy systems, may have extensive networks of physical sensors at a centralized level, but less so at the distribution level. Thus, while power plants are closely monitored, maps of outages rely on individual reports.

In this paper, we propose the use of social sensors to detect damage and failure events of critical public infrastructure. Recently, there has been an exploration of the use of data from social sensors to detect events for which physical sensors are lacking. This includes the use of Twitter data streams to detect natural disasters (Sakaki et al., 2010) or the use of texts to manage emergency response (Caragea et al., 2011). In this paper, we use the LITMUS framework – a framework designed to detect landslides using a multi-service composition approach (Musaev et al., 2014a, 2014b) – to detect public infrastructure failure events. We focus on two main systems: transportation (bridges and highways) and energy (gas lines and power). The rest of the paper is organized as follows. Section 2 provides an overview of the approach used to detect infrastructure failure events using social sensor data.

LITERATURE SURVEY

In case of emergencies (e.g., earthquakes, flooding), rapid responses are needed in order to address victims' requests for help. Social media used around crises involves self-organizing behavior that can produce accurate results

[1] Often in advance of official communications. This allows affected population to send tweets or text

messages, and hence, make them heard. The ability to classify tweets and text messages automatically, together with the ability to deliver the relevant information to the appropriate personnel are essential for enabling the personnel to timely and efficiently work to address the most urgent needs, and to understand the emergency situation better. In this study, we developed a reusable information technology infrastructure, called Enhanced Messaging for the Emergency Response Sector (EMERSE). The components of EMERSE are: (i) an iPhone application; (ii) a Twitter crawler component; (iii) machine translation; and (iv) Automatic message classification. While each component is important in itself and deserves a detailed analysis, in this paper we focused on the automatic classification component, which classifies and aggregates tweets and text messages about the Haiti disaster relief so that they can be easily accessed by non-governmental organizations, relief workers, people in Haiti, and their friends and families.

They propose and evaluate a probabilistic frame work

[2] For estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. By augmenting the massive human-powered sensing capabilities of Twitter and related microblogging services with content-derived location information, this framework can overcome the sparsity of geoenabled features in these services and enable new location based personalized information services, the targeting of regional advertisements, and so on. Three of the key features of the proposed approach are:

(i) its reliance purely on tweet content, meaning no need for user IP information, private login information, or external knowledge bases; (ii) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (iii) a lattice-based neighborhood smoothing model for refining a user's location estimate. The system estimates k possible locations for each user in descending order of confidence. On average we find that the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100 miles of their actual location

People in the locality of earthquakes are publishing anecdotal information about the shaking within seconds of their occurrences via social network technologies, such as Twitter. In contrast, depending on the size and location of the earthquake, scientific alerts can take between two to twenty minutes to publish. We describe TED (Twitter Earthquake Detector)

[3] A system that adopts social network technologies to augment earthquake response products and the delivery of hazard information. The TED system analyzes data from these social networks for multiple purposes: 1) to integrate citizen reports of earthquakes with corresponding scientific reports 2) to infer the public level of interest in an earthquake for tailoring outputs disseminated via social network technologies and 3) to explore the possibility of rapid detection of a probable earthquake, within seconds of its occurrence, helping to fill the gap between the earthquake origin time and the presence of quantitative scientific data.

Little research exists on one of the most common, oldest, and most utilized forms of online social geographic information

[4] The "location" field found in most virtual community user profiles. We performed the first in-depth study of user behavior with regard to the location field in Twitter user profiles. We found that 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools. When users did input their location, they almost never specified it at a scale any more detailed than their city. In order to determine whether or not natural user behaviors have a real effect on the "locatability" of users, we performed a simple machine learning experiment to determine whether we can identify a user's location by only looking at what that user tweets. We found that a user's country and state can in fact be determined easily with decent accuracy, indicating that users implicitly reveal location information, with or without realizing it. Implications for location-based services and privacy are discussed

Micro blogging sites such as Twitter can play a vital role in spreading information during "natural" or man-made disasters

[5] But the volume and velocity of tweets posted during crises today tend to be extremely high, making it hard for disaster-affected communities and professional emergency responders to process the information in a timely manner. Furthermore, posts tend to vary highly in terms of their subjects and

usefulness; from messages that are entirely off-topic or personal in nature, to messages containing critical information that augments situational awareness. Finding actionable information can accelerate disaster response and alleviate both property and human losses. In this paper, we describe automatic methods for extracting information from microblog posts. Specifically, we focus on extracting valuable "information nuggets", brief, self-contained information items relevant to disaster response. Our methods leverage machine learning methods for classifying posts and information extraction. Our results, validated over one large disaster-related dataset, reveal that a careful design can yield an effective system, paving the way for more sophisticated data analysis and visualization systems

METHODOLOGY

An overview of the approach is shown in Figure 1. The sensor data source is Twitter. For the results presented in this paper, these are tweets pulled over the period of one month.

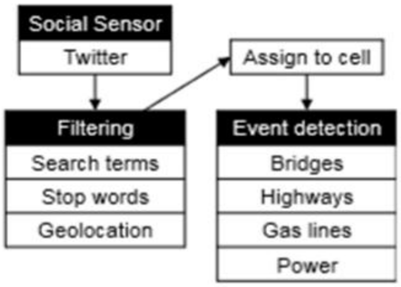


Figure 1: Overview of data, filtering, and event detection approach.

We use October 2018 as our evaluation period. It is noted that data from any other time period can be used within this framework. To detect infrastructure damage or failure events, all Twitter data is run through a series of filters to obtain a subset of relevant data. This filtering is done in three phases. First, we filter by search terms, which we have developed for various events of interest, e.g., "bridge collapse" to detect damage to bridge infrastructure. Second, as social sensor data is often noisy, with items containing the search terms but unrelated to the event of interest, data is filtered using stop words. Using a simple exclusion rule based on the presence of stop words, this filters out the irrelevant data. An example for detecting bridge collapses is the stop word "friendship" that refers to the collapse of a bridge or connection between two people. Third, data is filtered based on geolocation. Although most social networks enable users to geotag their locations, e.g., when they send a tweet, studies have shown that less than 0.42% of tweets use this functionality (Cheng et al., 2010). In addition, users may purposely input incorrect location information in their Twitter profiles (Hecht et al., 2011). As geolocating tweets is an important component in being able to identify specific infrastructure damage events, including their location, the data must be additionally filtered. In this study, the Stanford coreNLP toolkit (Manning et al., 2014) is used along with geocoding (Google, 2016) to geolocate the tweet. This assigns each filtered tweet to latitude and longitude and corresponding 2.5-minute by 2.5-minute cell as proposed in Musaev et al., 2014, based on a grid mapped to the surface of the Earth. Once all relevant tweets are mapped to their respective cells, all tweets in a single cell are assessed to identify the infrastructure damage and failure events. In this paper, we focus on the results for tweets relating to damage detection in four infrastructures: bridge, highway, gas line, and power infrastructure.

IMPLEMENTATION

Detection of damage and failure events to public infrastructure is implemented using data mining machine learning technique such as filtering, Decision tree classification techniques. Here we are implemented collaborative filtering technique to filter on search items based on critical information which is present in the content of twitter dataset. Classification technique is used to classify the identified filtered element into group which are related to each other based on the subject which we are considered. Then cluster is going to form the group of similar elements like on user's group and depends on subject group etc...,

The motivation for collaborative filtering comes from the idea that people often get the best recommendations from someone with tastes similar to themselves. Collaborative filtering encompasses techniques for matching people with similar interests and making recommendations on this basis.

Collaborative filtering algorithms often require (1) users' active participation, (2) an easy way to represent users' interests, and (3) algorithms that are able to match people with similar interests.

Typically, the workflow of a collaborative filtering system is:

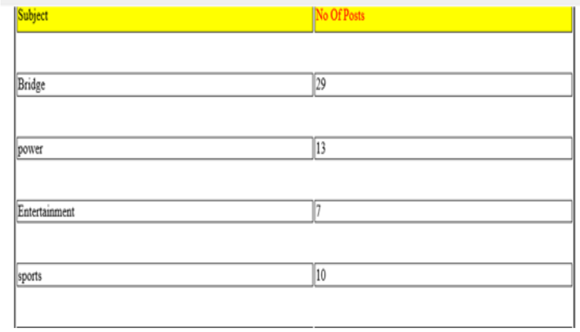
A user expresses his or her preferences by rating items (e.g. books, movies or CDs) of the system. These ratings can be viewed as an approximate representation of the user's interest in the corresponding domain. The system matches this user's ratings against other users' and finds the people with most "similar" tastes. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user (presumably the absence of rating is often considered as the unfamiliarity of an item).

Classification is technique to categorize our data into a desired and distinct number of classes where we can assign label to each class. Here are used decision tree classification technique to make decision on available data items and classify them according to critical information. Decision Tree is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data.

RESULT

Result on the studies is carried out with large number of data set collected from twitter dataset. Classification is done on the dataset with filtering resulting in different categories of data is available on different subjects .Here some of critical terms are considered like Bridges, Transports, Gas links like this we are considered a dataset processed according requirements and resulted in 98 percent accuracy using machine learning technique.

Example of bridges gases and sports of datasets categorized.



Subject	No Of Posts
Bridge	29
power	13
Entertainment	7
sports	10

CONCLUSION

Detection of damage and failure events to public infrastructure is critical to the ability of communities around the world to minimize the risks associated with both natural and man-made disasters and to recover more quickly and efficiently from the negative effects of these hazards. As many of our public infrastructure systems are not physically monitored to the degree necessary to provide relevant, detailed information about the states of these systems in real time, social sensor data is used to perform this assessment and detect damage events. In this paper, we describe an approach to use social sensor big data to identify public infrastructure damage events. This includes a three step filtering approach, whereby data is first filtered using search terms relevant to the event of interest. Next, noise in the data is filtered out using an exclusion rule based on the presence of stop words. Finally, data is filtered based on geolocation, resulting in each relevant filtered data item being assigned to a 2.5-minute by 2.5-minute cell in a grid mapped to the surface of the Earth.

REFERENCES

[1] Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, A.H., Giles, L., Jansen, B.J., Yen, J., 2011. Classifying text messages for the Haiti earthquake. In ISCRAM '11, Lisbon, Portugal.

[2] Cheng, Z., Caverlee, J., Lee, K., 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In CIKM'10, Toronto, Canada. Google, https://developers.google.com/maps/documentation/geocoding/intro, accessed on 2/5/2016.

[3] Guy, M., Earle, P., Ostrum, C., Gruchalla, K., Horvath, S., 2010. Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In IDA'10, Tuscon, Arizona.

[4] Hecht, B., Hong, L., Suh, B., Chi, E.H., 2011. Tweets from Justin Bieber's heart: The dynamics of the "location" field in user profiles. In CHI '11, Vancouver, Canada.

[5] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Extracting information nuggets from disaster related messages in social media. In ISCRAM '13, Baden-Baden, Germany.