Monitoring and Analysis of Dynamic Traffic Analyzer using Twitter

[1]B Sureesha, [2]V Priyadarshini
[1,2]SRKR Engineering College, Andhra Pradesh, India

Abstract
Internet sites are source of info for event detection, with specific mention of the road traffic activity blockage and accidents or earth-quack sensing system. In this paper, we present a real-time monitoring system intended for traffic occasion detection coming from Twitter stream analysis. The system fetches tweets coming from Twitter as per a several search criteria; methods tweets, by applying textual content mining methods; last but not least works the classification of twitter posts. The goal is to assign suitable class packaging to every tweet, because related with an activity of traffic event or perhaps not. The traffic recognition system or framework was utilized for real-time monitoring of various areas of the street network, taking into account detection of traffic occasions just almost in actual time, regularly before on-line traffic news sites. All of us employed the support vector machine like a classification unit; furthermore, we accomplished a great accuracy value of ninety five. 85% by attempting a binary classification issue. All of us were also capable to discriminate if traffic is triggered by an external celebration or not, by resolving a multiclass classification issue and obtaining accuracy worth of 90. 89%.

Keywords
Traffic Event Detection, Tweet Classification, Text Mining, Social Sensing

I. Introduction

Twitter is prone to malicious tweets containing URLs for spam, phishing, and malware distribution. Conventional Twitter spam detection schemes utilize account of features such as the ratio of tweets containing URLs and the account creation date, or relation features in the Twitter graph. These detection schemes are ineffective against feature fabrications or consume much time and resources. Conventional suspicious URL detection schemes utilize several features including lexical features of URLs, URL redirection, HTML content, and dynamic behavior. However, evading techniques such as time-based evasion and crawler evasion exist. In this paper, we propose an intelligent system based on text mining and machine learning algorithms, for real time detection of traffic events from Twitter stream analysis. The system, after a feasibility study, has been designed and developed from the ground as an event-driven infrastructure, built on a Service Oriented Architecture (SOA). The system exploits available technologies based on state-of-the-art techniques for text analysis and pattern classification. These technologies and techniques have been analyzed, tuned, adapted, and integrated in order to build the intelligent system.

In particular, we present an experimental study, which has been performed for determining the most effective among different state-of-the-art approaches for text classification. The chosen approach was integrated into the final system and used for the on-the-field real-time detection of traffic events. In the existing system attackers use shortened malicious URLs that redirect Twitter users to external attack servers. To cope with malicious tweets, several Twitter spam detection schemes have been proposed. These schemes can be classified into account feature-based, relation feature-based, and message feature based schemes. Account feature-based schemes use the distinguishing features of spam accounts such as the ratio of tweets containing URLs, the account creation date, and the number of followers and friends. However, malicious users can easily fabricate these account features. The relation feature-based schemes rely on more robust features that malicious users cannot easily fabricate such as the distance and connectivity apparent in the Twitter graph. Extracting these relation features from a Twitter graph, however, requires a significant amount of time and resources as a Twitter graph is tremendous in size. The message feature-based scheme focused on the lexical features of messages.

However, spammers can easily change the shape of their messages. A number of suspicious URL detection schemes have also been introduced. With reference to current approaches for using social media to extract useful information for event detection, we need to distinguish between small-scale events and large-scale events. Small-scale events (e.g., traffic, car crashes, fires, or local manifestations) usually have a small number of SUMs related to them, belong to a precise geographic location, and are concentrated in a small time interval. On the other hand, large scale events (e.g., earthquakes, tornados, or the election of a president) are characterized by a huge number of SUMs, and by a wider temporal and geographic coverage. Consequently, due to the smaller number of SUMs related to small-scale events, small-scale event detection is a non-trivial task. Several works in the literature deal with event detection from social networks. Many works deal with large-scale event detection, and only a few works focus on small-scale event.

Regarding small-scale event detection, the detection of fires in a factory from Twitter stream analysis, by using standard NLP techniques and a Naive Bayes (NB) classifier in this project, we focus on a particular small-scale event, i.e., road traffic, and we aim to detect and analyze traffic events by processing users' SUMs belonging to a certain area and written in the Italian language. To this aim, we propose a system able to fetch, elaborate, and classify SUMs as related to a road traffic event or not.

II. Motivation

1. Tweets are created in real-time.
2. Twitter has 140-character-message limit and the popularity of mobile applications of twitter, users tweet and ability to retweet instantly [1].
3. Tweets have a broad coverage over events. Every user can report news that is happening around him or her.
4. Tweets are not isolated; instead it contains the rich information.
5. Tweets cover nearly all aspect of daily life such as breaking news, local events and personal feelings.

III. Objective

The vital objectives of proposed system are as follows: Design a real-time detection system for traffic analysis. The aim is to assign suitable class label to every tweet, as related with an activity of traffic event or not. It performs a multi-class classification, which recognizes non-traffic, traffic due to congestion or crash, and traffic due to external events. It detects the traffic events in real-time and It is developed as an event-ambitious infrastructure, built on an SOA architecture.

IV. Literature Survey

A. What's happening: A Survey of Tweets Event Detection

Twitter is now one of the main modes for spreading of ideas and information throughout the Web. Tweets discuss different style, ideas, events, and so on. This gave rise to an increasing interest in examining tweets by the data mining community. Twitter is, in nature, a good resource for identifying events in real-time.

In this survey paper, authors have presented four challenges of tweets event detection: health epidemics identification, natural events detection, trending topics detection, and sentiment analysis. These challenges are based mainly on clustering and classification. We review these approaches by providing a description of each one. These last years have been marked by the emergence of micro blogs. Their rates of activity reached some levels without precedent. Hundreds of millions of users are registered in these micro blogs as Twitter. They exchange and tell their last thoughts, moods or activities by tweets in some words [1].

B. ET: Events from Tweets

Social media sites some of which are Twitter and Face book have emerged as popular tools for people to express their ideas on various topics. The huge amount of data provided by these media is greatly valuable for mining trending topics and events. In this paper, we build an adequate, scalable system to detect events from tweets (ET). Our approach detects events by analyzing their textual and temporal components. ET does not require any target entity or domain knowledge to be stated; it automatically detects events from a set of tweets.

The key components of ET are:

1. An extraction scheme for event representative keywords
2. An adequate storage mechanism to store their appearance patterns, and
3. A hierarchical clustering technique based on the common co-occurring features of keywords.

Authors presented a scalable and adequate system, called ET, to detect real world events from a set of micro blogs/tweets. The key feature of this system is the adequate use of content similarity and appearance similarity among keywords, to cluster the related keywords. We demonstrate the adequateness of this combination in our experiments. ET does not need any human expertise or knowledge from other sources like Wikipedia, but still provides very accurate results. ET is evaluated on two different data sets from two different domains and it produces great results for both of them in terms of the precision [2].

C. Measurement and Analysis of Online Social Networks:

The online social networking sites like Orkut, YouTube, and Flickr are out of the most popular sites on the Internet. Users from these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and the contacts. The vogue of these sites provides an opportunity to study the characteristics of online social network graphs at an immense scale. Knowing these graphs is vital, both to improve the current systems and to design the new applications of online social networks. This paper shows a large scale measurement study and scrutiny of the structure of multiple online social networks. We scrutinize data gathered from four vogue online social networks: Flickr, YouTube, Live Journal, and Orkut. We crawled the publicly accessible user links on each of the site, obtaining a huge portion of each social network's graph [3]. Our data set contains over 11.3 million users and had 328 million links.

We suppose that this is the first study to examine multiple online social networks at scale. Our results explain the power law, small world, and scale free properties of online social networks. We find that the in degree of user nodes tends to match the out degree; that the networks have a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network. Lastly, the implications of these structural properties for the design of social network based systems. Presented an analysis of the structural properties of online social networks using data sets collected from four vogue sites. Our data shows that social networks are structurally different from previously studied networks, specifically the Web. Social networks have a much higher fraction of symmetric links and also display much higher levels of local clustering. We have outlined how these properties may affect the algorithms and applications designed for the social networks [4].

D. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors

Twitter, a vogue micro blogging service, has received much attention recently. A significant characteristic of Twitter is its real-time nature. For instance, when an earthquake occurs, people make many Twitter posts (tweets) related to the earthquake, which facilitates detection of earthquake occurrence promptly, simply by observing the tweets. As described in this paper, we scrutinize the real-time interaction of events such as earthquakes, in Twitter, and suggest an algorithm to monitor tweets and to find a target event. To find a target event, we arrange a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. Later, we produce a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location.

We then consider each Twitter user as a sensor and apply Kalman filtering and particle filtering, which are generally used for location estimation in ubiquitous/pervasive computing [5]. The particle filter works better than other compared methods in judging the centers of earthquakes and the trajectories of typhoons. As an application, we construct an earthquake reporting system in Japan. Because of the numerous earthquakes and the huge number of Twitter users throughout the country, we can investigate an earthquake by monitoring tweets with high probability (96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more is detected). Our system finds earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast by the JMA [6].

E. Text Detection and Recognition on Traffic Panels from Street-Level Imagery Using Visual Appearance

Traffic sign detection and recognition has been completely studied for a long time. Yet, traffic panel finding and recognition still remains a challenge in computer vision due to its different types and the immense variability of the information illustrated in them. This paper presents a method to detect traffic panels in street level images and to recognize the information contained on them, as an application to intelligent transportation systems (ITS) [7]. The main purpose can be, to make an automatic inventory of the traffic panels located in a road to support road maintenance and to help drivers. Our proposal extracts local descriptors at some interest key points after applying blue and white color segmentation. Then, images are represented as a "bag of visual words" and classified using Naïve Bayes or support vector machines. This visual appearance categorization method is a new methodology for traffic panel detection in the state of the art [8].

Lastly, our own text identification and recognition method is applied on those images where a traffic panel has been identified, so automatically read and save the information illustrated in the panels. We suggest a language model partially based on a dynamic dictionary for a finite geographical area using a reverse geo coding service. Experimental results on real images from Google Street View prove the efficiency of the suggested method and give a way to use street level images for different applications on ITS [9].

V. System Architecture

Fig. 1: System Architecture

Existing system propose an intelligent system, based on text mining and machine learning algorithms, for real-time detection of traffic events from Twitter stream analysis. The system, after a feasibility study, has been designed and developed from the ground as an event-driven infrastructure, built on a Service Oriented Architecture (SOA) [1]. The system exploits available technologies based on state-of-the-art techniques for text analysis and pattern classification [4]. These technologies and techniques have been analyzed, adapted, and added with existing in order to build the intelligent system [1]. In particular, system present an experimental study, which has been performed for determining the most effective among different state-of-the-art approaches for text classification. The chosen approach is added into the final system and then used for the on-the-field real-time detection of traffic events [1].

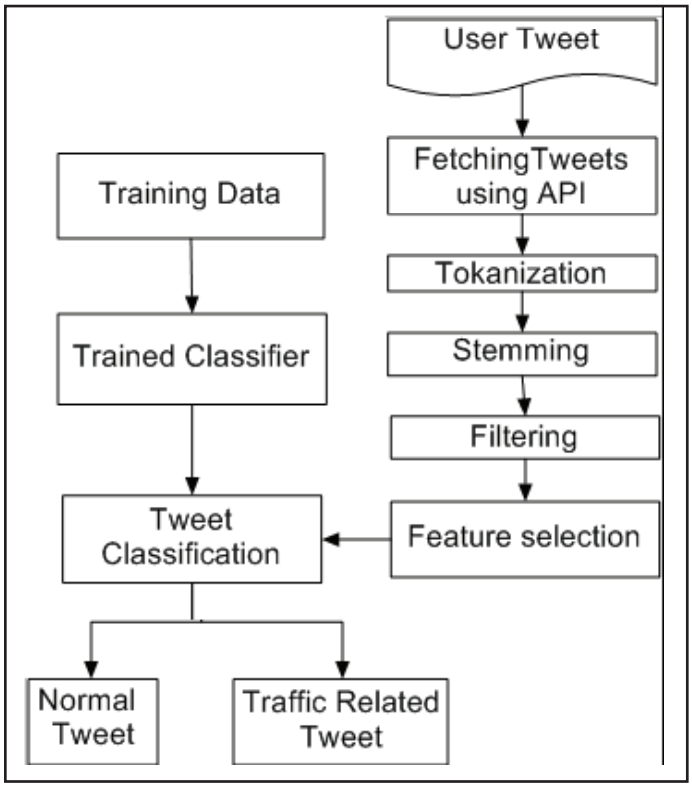

Fig. 2: System Architecture for Traffic Detection From Twitter Stream Analysis

In this section, our traffic detection system based on Twitter streams analysis is presented. The system architecture is service-oriented and event-driven, and is composed of three main modules, namely:

- Extraction of SUMs and Pre-processing
- Elaboration of SUMs
- Classification of SUMs

The purpose of the proposed system is to fetch SUMs from Twitter and process SUMs by using text mining steps, and to assign the appropriate class label to each SUM. Finally, as shown in Fig. 1, by analyzing the classified SUMs, the system is able to notify the presence of a traffic event.

1. Fetch of Sums and Pre-processing

The first module, Fetch of SUMs and Pre-processing‖, extracts raw tweets from the Twitter stream, based on one or more search criteria. Each fetched raw tweet contains: the user id, the timestamp, the geographic coordinate, retweet flag along with the text of the tweet.

2. Elaboration of Sums

The second processing module is Elaboration of SUMs. This is devoted to transforming the set of pre-processed SUMs, i.e., a set of strings, in a set of numeric vectors to be elaborated by the Classification of SUMs‖ module. To this aim, some text mining

techniques are applied in sequence to the pre-processed SUMs. In the following, the text mining steps performed in this module are described in detail:

(i). Tokenization is typically the first step of the text mining process. This process is used for transforming a stream of characters into a stream of processing units called tokens [1]. During this step, other operations are usually performed, such as removal of punctuation and other non-text characters [8], and normalization of symbols.

(ii). Stop-word filtering eliminates stop-words, the words which provide little or no information to the text analysis.

(iii). Stemming is the process of reducing each word (i.e., token) to its stem or root form, by removing its suffix. The purpose of this step is to group words with the same theme having closely related semantics.

(d). Stem filtering consists in reducing the number of stems of each SUM. In particular, each SUM is filtered by removing from the set of stems the ones not belonging to the set of relevant stems.

3. Classification of Sums

Last module is, Classification of SUMs. This module assigns each elaborated SUM a class label related to traffic events. Thus, the output of this module is a collection of N labeled SUMs.

Proposed Clustering Algorithm:
Input: Training Dataset T, Test dataset D,
Output: Clustered Tweet set.
Method:
1. Initially train the classifier using semi-supervised traffic related training dataset.
2. Fetch user tweets from tweeter account
3. Store in DB
4. For each tweet in DB
5. Calculate the similarity using Euclidean distance with trained data.
6. If (similarity > Threshold)
7. Add tweet to traffic related tweet set
8. Else
9. Add to normal tweet set.
10. End if
11. End for
12. Return classified tweets

VI. Result Analysis

In this system we are used three types of classes for SUM classification which are updated by user i.e. traffic related, Non traffic related and Traffic due to External event classification is done by using Navi Bayes classifier The first two class traffic related and non traffic related is also called 2Dataset and whole classes i.e. traffic related, on traffic related and Traffic due to External event is also called as 3Dataset. In this section we perform classification of SUM by the applying of NB Classifier, SVM and Text mining Technique. Some source words are used to fetching the SUM which is related to Traffic Event i.e. traffic, busy, jam, crush, queue, stuck, slowdown, signal etc. After classification of SUM its place in its desired class and our system send notification to suspicious user to knowing him about traffic status.

VII. Conclusion and Future Scope

In this paper, we have proposed a system for real-time detection of traffic-related events from Twitter stream analysis. The system, built on a SOA, is able to fetch and classify streams of tweets and to notify the users of the presence of traffic events. Furthermore, the system is also able to discriminate if a traffic event is due to an external cause, such as football match, procession and manifestation, or not.

As future work, we are planning to integrate our system with an application for analyzing the official traffic news web sites, so as to capture traffic condition notifications in real-time. Thus, our system will be able to signal traffic-related events in the worst case at the same time of the notifications on the web sites. Further, we are investigating the integration of our system into a more complex traffic detection infrastructure. This infrastructure may include both advanced physical sensors and social sensors such as streams of tweets. In particular, social sensors may provide a low-cost wide coverage of the road network, especially in those areas (e.g., urban and suburban) where traditional traffic sensors are missing.

References

[1] F. Atefeh, W. Khreich,"A survey of techniques for event detection in Twitter," Comput. Intell., Vol. 31, No. 1, pp. 132– 164, 2015.

[2] P. Ruchi, K. Kamalakar,"ET: Events from tweets," In Proc. 22nd Int. Conf. World Wide Web Comput., Rio de Janeiro, Brazil, 2013, pp. 613–620.

[3] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, B. Bhattacharjee,"Measurement and analysis of online social networks," In Proc. 7th ACM SIGCOMM Conf. Internet Meas., San Diego, CA, USA, 2007, pp. 29–42.

[4] The Smarty project. [Online]. Available: http://www.smarty.toscana.it/

[5] T. Sakaki, M. Okazaki, Y.Matsuo,"Tweet analysis for realtime event detection and earthquake reporting system development," IEEE Trans. Knowl. Data Eng., Vol. 25, No. 4, pp. 919–931, Apr. 2013.

[6] M. Krstajic, C. Rohrdantz, M. Hund, A. Weiler,"Getting there first: Real-time detection of real-world incidents on Twitter", in Proc. 2nd IEEE Work Interactive Vis. Text Anal.— Task-Driven Anal. Soc. Media IEEE VisWeek," Seattle, WA, USA, 2012.

[7] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power, "Using social media to enhance emergency situation awareness," IEEE Intell. Syst., Vol. 27, No. 6, pp. 52–59, Nov./Dec. 2012.

[8] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, K. Nawa,"Real-time event extraction for driving information from social sensors," In Proc. IEEE Int. Conf. CYBER, Bangkok, Thailand, pp. 221– 226, 2012.

[9] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, P. Chaovalit,"Social-based traffic information extraction and classification," In Proc. 11th Int. Conf. ITST, St. Petersburg, Russia, pp. 107– 112, 2011.

[10] P. Agarwal, R. Vaithiyanathan, S. Sharma, G. Shro, "Catching the long-tail: Extracting local news events from Twitter," In Proc. 6th AAAI ICWSM, Dublin, Ireland, pp. 379–382, Jun. 2012.