International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Real Time Event Detection using Probabilistic Method and Analysing through Twitter

Shruti M G[1], Sameena Banu[2]

[1]Master of Technology , Department of Computer Science and Engineeing, Visvesvaraya Technological University, Khaja Bamda Nawaz College Of Engineering Kalaburagi, Karnataka, India

[2]Associate Professor, Department of Computer Science and Engineeing, Visvesvaraya Technological University, Khaja Bamda Nawaz College Of Engineering Kalaburagi, Karnataka, India

Abstract: ***A probabilistic model provides a way to detect multiple instances of real time events and to estimate the locationof targeted event like earthquakes, typhoons, traffic jams. Forthis, two models have been proposed named temporal and spatialmodels to detect real time events and estimate the targeted eventlocations respectively by dealing with sensor readingappropriately. Our work is based on the twitter-which is used todeal with sensor reading appropriately real time events andparticularly for location estimation. An important characteristic of twitter is its real-time nature. We investigate the real-timeinteraction of events such as earthquakes in twitter and propose an algorithm to monitor tweets and to detect a target event. As an application, we develop an earthquake reporting system for use in Japan. Because of the numerous earthquakes and the large number of Twitter users throughout the country, we can detect an earthquake with high probability (93 percent of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected) merely by monitoring tweets. Our system detects earthquakes promptly and notification is delivered much faster than JMA broadcast announcements.***

Keywords: probabilistic model, twitter analysis, tweets, particlefiltering, social networking.

1. Introduction

Our work is on detection of real time events like earthquakes, using online social network, particularly by use of micro blogging service. Twitter has been examined by some researchers, and notified it as, is one of the microblogging service. An important characteristic of Microblogging service is of its real time nature. So, twitter has been consider for our research, to detect the real time event, depending on the tweets posted. Tweet is a status update message that posts to our friends and colleagues, are socially connected around the world. For instance using Twitter people comment on various ongoing and real time events. Elections are talked about over Twitter. The tweets made by various people can have useful information when mined. Such knowledge can help in making well informed decisions pertaining to advertisements, marketing, promoting sales campaigns and so on. The tweet analysis in the real time also helps to know the important topics on which people are focusing. The process of acquiring business intelligence from the tweets of Twitter became active research area of late. Many researchers tired to build applications based on Twitter. That way research on Twitter became very important as provides insights to make businesses grow faster. Micro blogging is attributed to tweets concept in Twitter. This kind of blogging enables people to send text updates to online users besides expressing their

Responses to various real time events. From the twitter tweets and studying them we understood that it is possible to analyse tweets and gain business intelligence. Recently Sakaki*etal*. [1] presented an approach to analyse tweets for real time event detection. Inspired by this research, in this paper, we propose a framework that allows tweet analysis. We also built a probabilistic model that demonstrates the proof of concept. Our application can detect multiple instances of events from analyzing the tweets in spatio temporal domains. Our empirical results revealed that the proposed application is very useful to detect multiple instances of real time event and notify people so as to make necessary steps in the aftermath of events. A probabilistic model is a statistical analysis tool that estimates, on the basis of past (historical) data, the probability of an event occurring again. Different statistical tools are available, some of them are simple, and some are complicated, and often very specific for certain purposes. In analytical work, comparison of data, or sets of data has been done to quantify accuracy (bias) and precision. Fortunately, with a few simple statistical tools: the "*t*-test, the "*F*-test", and regression analysis, can quantify the accuracy and precision. Clearly, statistics area tool, not an aim. Simple inspection of data, without statistical treatment, by an experienced and dedicated analyst may be just as useful as statistical figures on the desk of the disinterested. The value of statistics lies with organizing and simplifying data, to permit some objective estimate showing that an analysis is under control or that a change has occurred. Equally important is that the results of these statistical procedures are recorded and can be retrieved.

2. Related Work

Social media has been active for many years and the users of it are growing exponentially. This section provides review of literature on related works such as real time event detection. Key word based topic search was proposed by Cataldi et al. [2] in order to identify emerging topics with respect to news and keywords. Twitter network features were investigated by some researchers [3]. Later on Haewoon et al. [4] crawled Twitter data and applied PageRank algorithm on it. Then Huber man et al. [5] analyzed 3 lakh Twitter users to discover interactions among friends. Characteristics of Twitter as social media were investigated by some

researchers [6]. With respect to the election of Germany Tumasjan et al. [7] analyzed Twitter tweets to predict winners in elections. Sentiment analysis concept was used in [8] for knowing public opinion. Some researchers investigated Twitter to relate it with mobile e-Learning [9]. In [10] some investigation is made to know the relationships between micro blogging and semantic web. Many applications came into existence in order to examine Twitter data and analyze the content for various benefits such as marketing, advertising and so on. The Twitter and its spatial aspects were studied by Backstop et al. [11] which is close to the study made in [1].both the researches [11] and [1] are close the present research in this paper. Another research was carried out on blogging for event detection through discovery of spatiotemporal patterns . Photographs from Flickerwere used in to map them with world map. The placeand event semantics were explored in . In the field ofsocial media location estimation related studies were found. Estimation of the location of an object under study hasmany real time utilities. To achieve this different type softGPS and infrared badges were used. Another technique used for location estimation is the particle filtersas explored in . The particle filter approach proved tobeefficient. RecentlySakaki et al. [1] presented an approach todetect real time events by analyzing tweets. Their research focuses on probabilistic spatiotemporal model in order tolocate the event place and event recognition. Particlefiltering is used in their search for exact locationestimation. As there were numerous earthquakes happeningin the world, they study was focusing on them. In this paperwe did related work but our approach can detect multipleinstances of the events at a time. This helps to have betterresults when it comes to notifying users so as to let themunderstand about real time events. This is achieved byanalyzing tweets. As the tweeting became part of socialmedia and that is the means to generate voluminous data over Internet, the tweet analysis was given importance in our paper too.

3. System Architecture

In this paper we proposed a framework for , the detection of multiple instances of events and the location of such events by analyzing tweets posted on Twitter ,describes the event has been occurred . Recent study on this kind of research was done in [1]. However it locates single instance of an event .We use the concepts and ideas from those authors and present similar kind of framework that detects and locates multiple instances of an event with some modifications. The framework proposed is as shown in Figure 1.

As shown in Figure 1, it is evident that the framework has various modules for performing tweet analysis and finally showing multiple instances of events along with location map. The functional requirements of the system are logically divided into the following modules such as Tweet Crawler, Classification, Event Detection and Location Estimation. Tweet Crawler Module is responsible to connect to the social networking site - Twitter and obtain tweets based on the keywords given. To achieve this it makes use of Twitter search API provided by Twitter for public use. Depends on the search word “earthquake”, it is possible to gather the tweets that consisting the search word ”earthquake”.

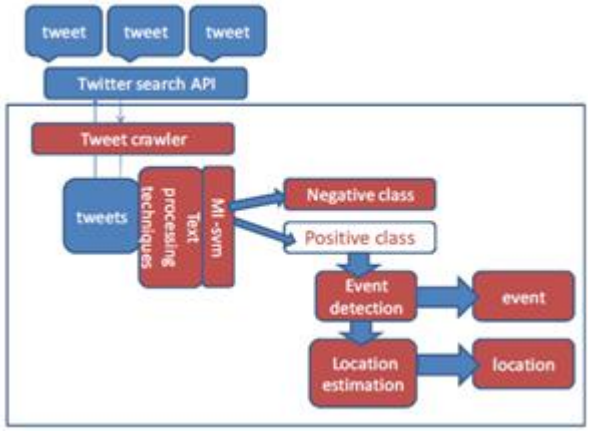

Figure 1: Architecture

This should be happens by categorizing the tweets into three groups of features i.e. statistical, keyword, word context features. In our application, keyword based feature has been adopted. Depends on this keyword based search, we should consider the tweets that are more relevant to the search word “earthquake”. Next, by adopting the tweets that are most relevant to the search word, those tweets should be processed. In this tweets processing, which are in natural language, we use text processing techniques like stop –word removal and stemming. After processing these tweets, we apply classification algorithm that separates the tweets into positive class or negative class.

We prepare three groups of features for each tweet as described below.

- Features A (statistical features): the number of words in a tweet message, and the position of the queryword within a tweet.
- Features B (keyword features): the words in a tweet.
- Features C (word context features): the words before and after the query word.

We can give an illustrative example of these features using the following sentence.

“**I am in Japan, earthquake right now!”**
(keyword: earthquake)

For this example, Features A, B, C are presented in Table 1.

TABLE 1
SVM Features of an Example Sentence

Feature Name	Features
Features A	7 words, the fifth word
Features B	I, am, in, Japan, earthquake, right, now
Features C	Japan, right

After the classification of tweets into positive class, these positive tweets will help us in detection of an event.

Temporal model: For this we make use of temporal model, which describes the time series of data. Each tweet has its own post time , depending upon the quantity of tweets about the target event, have to calculate the probability density function for our application , it is an exponential distribution

, because of chance of getting more than 10 earthquake events my happens within particular period. So this exponential distribution should be f(t;λ) = λe ¯λt where t>0 and λ>0here, λ can be the probability of post a tweet from time t to Δt. Assumes , n sensors, produce positive signals and the probability of all n sensors returning a false alarm is pfn. Therefore, the probability of event occurrence can be estimated as 1− pfn .so, the probability of an event occurrence at time t isp occur(t) = 1- pfn.We can calculate the probability of event occurrence if we set λ= 0.34 and pf=0.35.

***Location Estimation Module*:** After detecting events, the location estimation module is responsible for analyzing spatial and temporal information present in the tweets and estimates the locations of the real-time event. To locate the location of the event occurred, "particle filtering "algorithm has been proposed [7]. The algorithm is presented below in the algorithm sections, where in each steps have been given to describe it.

Social sensors are used in the process of event detection and the location from the tweets analyzed. In alerts are generated for a single instance of target event like earthquakes. With this system, users are sent alert messages in the form of tweets, about earthquakes in other places once event is detected. And administrator confirms that earthquake and report to all users in this site. With this alert public can take preventive steps to protect their lives and properties. In proposed system, we generate alerts for multiple instance of target events like earthquakes, typhoons, traffic jams. With this system, user sends messages about real time events in their places, once event is detected .It alerts public, so that the public and government can take preventative steps to protect their lives and property.

4. Algorithms

This sections deals with the algorithms which have been extensively used for this paper. Below are the algorithms with the description .

1. Text classification Algorithm :

Algorithm for text classification is

initialize $yi = YI$ for $i \in I$
REPEAT
compute SVM solution **w**; b for data set with imputed labels
compute outputs $fi = (\mathbf{w}; \mathbf{x}_I) + b$ for all **x***i* in positive bags
set $yi = \text{sgn}(fi)$ for every $i \in I$, $YI = 1$
FOR (every positive bag BI)
IF $(\sum i \in I\ (1 + yi)/2 == 0)$
compute $I^* = \text{argmax} i \in I\ fi$
set $yi^* = 1$
END
END
WHILE (imputed labels have changed)
OUTPUT (**w**; b)

In text classification, this F- test has most important to quantify the accuracy on different data sets. More Fmeasure, will gives most accurate data. So, Classification Module is responsible to take tweets obtained by crawler module and apply classification algorithm of data mining. To make classification of tweets whether they are related to event i.e either positive or not, Multi Instance – Support Vector Machine has been used [13]. Due the success of the Support Vector Machine (SVM) algorithm and the various positive theoretical results behind it, maximum margin methods have become extremely popular. So, SVM have been proposed for the MIL problem. SVM is used to find a hyperplane in the input space that separates the training data points with as big a margin as possible. The classifier is defined by the hyperplane normal w and the offset b, h = {w, b}. The margin, is defined as the smallest distance between a positive or negative point and this hyperplane. For example, we assume that the margin is at least 1, and shrink the size of the hyperplanenormal w. Data may actually not be separable, we also include slack variables €i for each point xi. The closest points to the hyperplane are called support vectors.

2. Particle Filter Algorithm :

Algorithm particle_ filter (X_{t-1} , u_t , z_t):
$\overline{X}_t = X_t = \emptyset$
for m = 1 to M do
sample $x_t^{[m]} \sim p(x_t \mid u_t, x_{t-1}^{[m]})$
$w_t^{[m]} = p(z_t \mid x_t^{[m]})$
$\overline{X}_t = \overline{X}_t + < x_t^{[m]}, w_t^{[m]} >$
endfor
for m = 1 to M do
draw i with probability $\propto w_t^{[i]}$
add $x_t^{[i]}$ to X_t
end for
return X_t

A particle filter is a probabilistic approximation algorithm implementing a Bayes filter, and a member of the family of sequential Monte Carlo methods.

For location estimation, it maintains a probability distribution for the location estimation at time t, designated as the belief Bel (xt) ={xt ,wt }, where i= 1 to n . Each xit is a discrete hypothesis related to the object location. The wit are nonnegative weights, called importance factors, which sum to one.

3. Sequential importance sampling (SIS) algorithm:

- Given a set of queries Q for a target event.
- Put a query Q using search API every s seconds and obtain tweets T.
- For each tweet t 2 T, obtain features A, B, and C. Apply the classification to obtain value vt ¼ f0; 1g.
- If the enough number of tweets comes(poccur in (1) exceeds 0.99 under the condition: 10 tweets in 10 minutes; _ ¼ 0:34; pf ¼ 0:35;) then proceed to step five (5).
- For each tweet t 2 T, we obtain the latitude and the longitude lt by using the associated GPS location, and by

making a query to Google Map for the registered location for user ut. Set lt ¼ null if neither functions.

- Calculate the estimated location of the event from lt; t 2 T using normal particle filtering, particle filtering with assigned weights, and particle filtering with weights and sampling.

The Sequential Importance Sampling (SIS) algorithm is a Monte Carlo method that forms the basis for particle filters. The SIS algorithm consists of recursive propagation of the weights and support points as each measurement is received sequentially.

5. Results

Experiments are made with the proposed application that demonstrates the concept of multiple instances of real time event detection though tweet analysis. Our experiments are mainly focused on detection of multiple instances of real time event like earthquake. The results in this paper are implemented as this is still under implementation.

For any given set of queries, first we process and sample the texts of the tweet. Then find the relevant tweet from T, by using processing techniques. And then apply the multiple instance learning algorithm for classification to obtain positive or negative class of tweets. If the enough number of positive tweets comes (poccurin(1) exceeds 0.99 under the condition: 10 tweets in 10 minutes;ʎ = 0:34; pf = 0:35;). Then we use the geo location for each tweet by GPS for the registered users .calculate the latitude by using particle filtering. Finally an email is sent to the registered users.

Put Number of days, the tweets are utilized for experiments The results reveal that the application is able to detect multiple instances of real time events and identify locations besides sending notifications to people concerned.

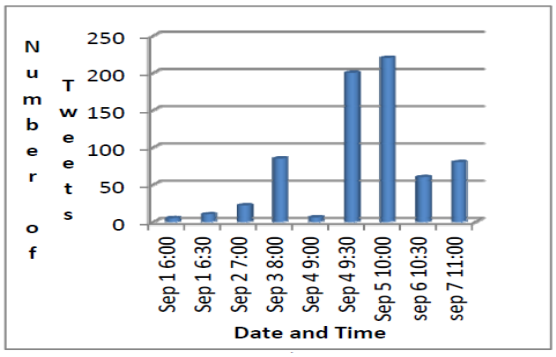

Figure 2

As can be seen in Figure 3, it is evident that, the real time tweets are taken for experiments. Especially observations are made on earthquake events and the statistics are plotted. The horizontal axis represents date and time of tweets while the vertical axis represents the number of tweets that are related to earthquakes.

6. Conclusion and Future Work

In this paper we studied Twitter as social medium. Thetweets are considered to have insights into real time eventssuch as earthquakes. Probablistic model is been implemented to predict the earthquake for multiple instances. Analysing with tweets help in detecting a real time event detection, also an event detection module has been implemented. Location are estimated with the help of API's and GPS. This can be implemented in any of the event detection such as traffic jams, tsunami, alerting systems. To find the more precise real time values the advanced algorithm can be approached to make it more realistic and more accurate in nature. Thus helps in improving the human power.

References

[1] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Tweet Analysis for Real-Time EventDetection and Earthquake ReportingSystem Development"
[2] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation," Proc. 10th Int'l Workshop Multimedia Data Mining(MDMKDD '10), pp. 1-10, 2010.
[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter:Understanding Microblogging Usage and Communities," Proc.NinthWebKDD and First SNA-KDD.
[4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, A Social Network or A News Media?" Proc. 19th Int'l Conf. World Wide Web (WWW '10), pp. 591-600, 2010.
[5] B. Huberman, D. Romero, and F. Wu, "Social Networks that Matter: Twitter Under the Microscope," ArXiv E-Prints, http:// arxiv.org/abs/0812.1045, Dec. 2008. Workshop Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07), pp. 56-65, 2007.
[6] G.L. Danah Boyd and S. Golder, "Tweet, Tweet, Retweet:Conversational Aspects of Retweeting on Twitter," Proc. 43rd Hawaii Int'l Conf. System Sciences (HICSS-43), 2010.
[7] MiodragBolic "Theory and implementation of particle filters "
[8] B. O'Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," Proc. Int'l AAAI Conf. Weblogs and Social Media, 2010
[9] M. Ebner and M. Schiefner, "Microblogging - More than Fun?" Proc. IADIS Mobile Learning Conf., pp. 155-159, 2008.
[10] A. Passant, T. Hastrup, U. Bojars, and J. Breslin, "Microblogging: A Semantic Web and Distributed Approach," Proc. Fourth Workshop Scripting for the Semantic Web (SFSW '08), http:// data.semanticweb.org/workshop/scripting/2008/paper/1 1, 2008.
[11] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial Variation in Search Engine Queries," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 357-366, 2008.
[12] Q. Mei, C. Liu, H. Su, and C. Zhai, "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 533-542, 2006.