Disaster Events Detection using Twitter Data

Hongwon Yun, *Member, KIMICS*

***Abstract*—Twitter is a microblogging service that allows its user to share short messages called tweets with each other. All the tweets are visible on a public timeline. These tweets have the valuable geospatial component and particularly time critical events. In this paper, our interest is in the rapid detection of disaster events such as tsunami, tornadoes, forest fires, and earthquakes. We describe the detection system of disaster events and show the way to detect a target event from Twitter data. This research examines the three disasters during the same time period and compares Twitter activity and Internet news on Google. A significant result from this research is that emergency detection could begin using microblogging service.**

***Index Terms*—Twitter, Tweets, Disaster Events, Disaster Detection.**

I. INTRODUCTION

Microblogging is a variation on blogging in which users write short posts to a special blog that are subsequently distributed to their friends and other observers. Twitter, a popular Microblogging service, has received much attention recently [1,2]. It allows its users to send short messages to others and these short messages are referred to as tweets and can be sent and retrieved through a variety of media. All the tweets are visible on a public timeline, where an asymmetric following system allows users to see their personal timeline for tweets they consider to be interesting [3]. These twitter data have the valuable geospatial component and particularly time critical events. According to TechCrunch, Twitter is now attracting 190 million visitors per month and generating 65 million Tweets a day. These numbers are up slightly from 180 million self-reported unique visitors per month back in April, and 50 million Tweets per day in February 2010 [8].

Twitter is able to send short message with mobile devices and retweet those messages to their followings, it would be use for emergency management activities. Twitter provides access to thoughts, opinions, activities, and experiences of several hundred millions of users in real time, with the option of sharing the user's location. This rich source of data is motivating a growing body of scientific as studying user motivations [9,10,11] and user collaboration [12,13,14]. Some researchers have focused specifically on crisis management and collective problem solving in mass emergency events [2,15,16,17]. According to Sakaki et al.'s study of Earthquake sensing, their recent paper results in the way to infer the epicenter of earthquake and the trajectories of typhoons using twitter data [2]. De Longueville, et al. [3] analyzed the twitter data related to a wild fire near the French city of Marseille and showed that twitter data were generally well synchronized to the temporal and dynamics of the Marseille fire event [9]. These are the motivation for this research resented here.

In this study, our interest is in the rapid assessment of hazard event based on features such as the keywords in tweets, the frequency of keywords, time and location. We describe the detection processes of disaster events and show how these data can be useful as an early detection indicator for emergency disaster. The remainder of this paper is organized as follows. Section 2 describes an overview of previous work in emergency events. In section 3 we introduce the collection, classifying, filtering and analysis of twitter data to detect emergency events. Preliminary evaluation is showed in Section 4. Finally, we summarize our research and give some future work directions in Section 5.

II. PREVIOUS WORKS

Twitter is currently one of the most popular microblogging platforms. Fig. 1. shows the geographical distribution of Twitter users across the world and the number of users in each continent. This map was generated by Akshay Java on April 15, 2007 [6]. Twitter is globally popular and the social network of its users crosses continental boundaries. We can recognize that Twitter is most popular in North America, Europe and Asia even though it is not newly data. It means that Twitter will be potential indicators of geographically specific events.

A tsunami is a series of waves generated by an impulsive disturbance in the ocean or in a small, connected body of water. These waves sometimes inflict severe damage to property and pose a threat to life in coastal communities [4]. Fig. 2. depicts a map of the Tsunami runup that is from National Oceanic and Atmospheric Administration. This Tsunami runup map

Manuscript received December 30, 2010; revised January 29, 2011; accepted February 1, 2011.

Hongwon Yun is with the Department of Information Technology, Silla University, Busan, 609-735, Korea (Email: hwyun@silla.ac.kr)

provides information on locations where tsunami effects occurred. A twitter user who is an information source is also a sensor for disasters and has a large number of followers. Twitter users are potentially the near real-time detection of a hazard such as Tsunami.

A tornado is defined as violently rotating column of air extending from a thunderstorm to the ground. These destructive forces of nature are found frequently all around the world. A tornado is part of a severe convective storm, and these storms occur all over the Earth, tornadoes are not limited to any specific geographic location [5] as shown in Fig. 3. We have no way at present to predict exactly which storms will spawn tornadoes or where they will touch down, however Twitter just can be useful for emergent real-time notification and help during a emergency disaster.

In Korea, the forests of three regions are Kangwon coastal, Woolyong coastal, and Hyung-Taewha coastal (eastern coastal region of Korea). These regions are vulnerable to fire because they have very low rainfall in the spring; and foehn and quasi-foehn winds abruptly interchange many times in a day. Under these meteorological conditions, wildfires spread rapidly and over large areas. Fig.4. shows large scale forest fires during the period 1980 – 1999 in South Korea [7]. Large scale forest fires cause huge damage to the environment and endanger human lives. Many countries are developing forest fire information systems to effectively manage and control forest fire. Twitter can be used as source of disaster events and we can find its possible role to support emergency disasters.

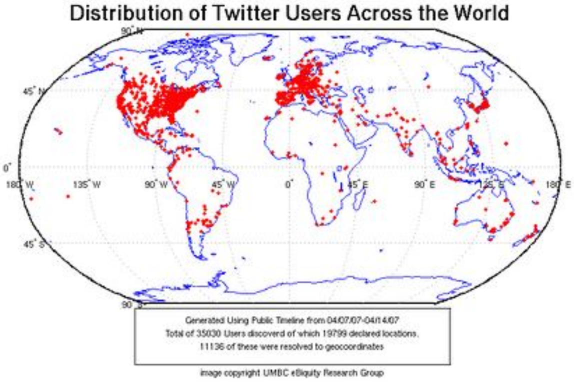
Distribution of Twitter Users Across the World
90° N
45° N
0°
45° S
90° S
180° W
135° W
90° W
45° W
0°
45° E
90° E
135° E
180° E
Generated Using Public Timeline from 04/07/07-04/14/07
Total of 35030 Users discoverd of which 19799 declared locations.
11136 of these were resolved to geocoordinates
image copyright UMBC eBiquity Research Group

Fig. l. Global map of twitter users

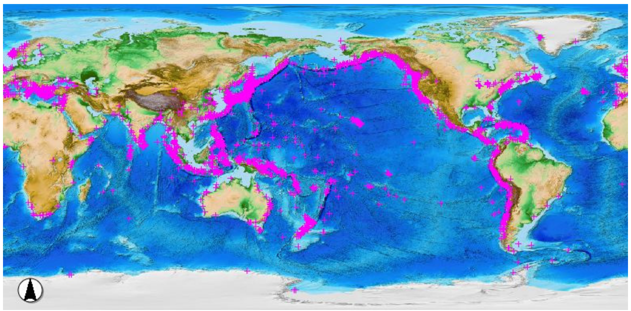

Fig. 2. Global historical Tsunami events and runups

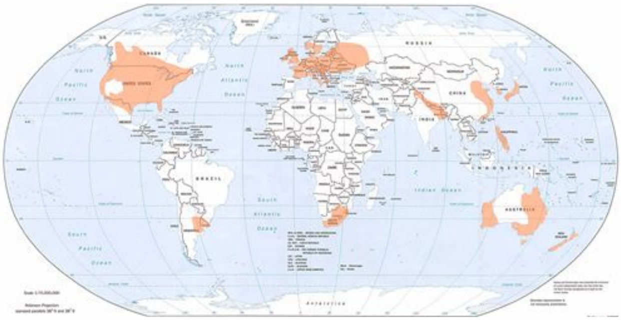

Fig. 3. Regions of the world with increased likelihood of experiencing tornadoes

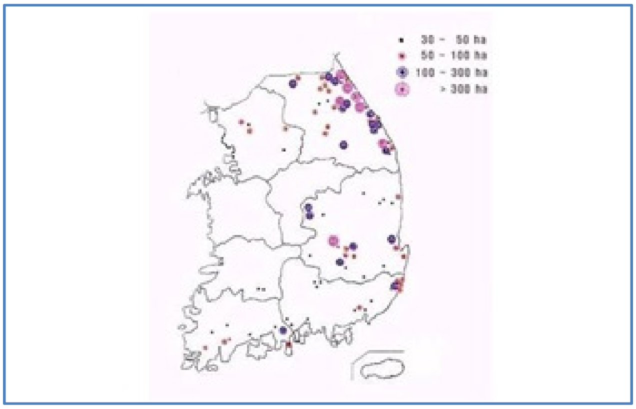

Fig. 4. Large scale forest fires in Republic of Korea

III. DISASTER DETECTION

Disaster events such as tsunamis, tornadoes, fire forests, and earthquakes are visible through Twitter data. In this section, we describe how to detect an emergency event from tweets. Our proposed emergency detection system is to analysis and detects disaster from twitter data. This system collects Twitter data continuously similar to real time mining to find only the specific terms such as tsunami, earthquake, fire forest and tornado. It is significant to pick up the specified keywords in huge volume of data. The whole process for emergency detection is presented in Fig. 5. Twitter users send tweets to their followers when the disaster events are occurred. Tweets subsequently distributed to their friends through the specific data center. The emergency detection system can pick up Twitter data related to disaster events. In this process, we use bag of disaster event words in order to search specific keywords such as tsunami, tornado and so on. To classify for disaster events, we use a classification tree as shown in Fig. 6. A certain probability model should be considered to determine a disaster event.

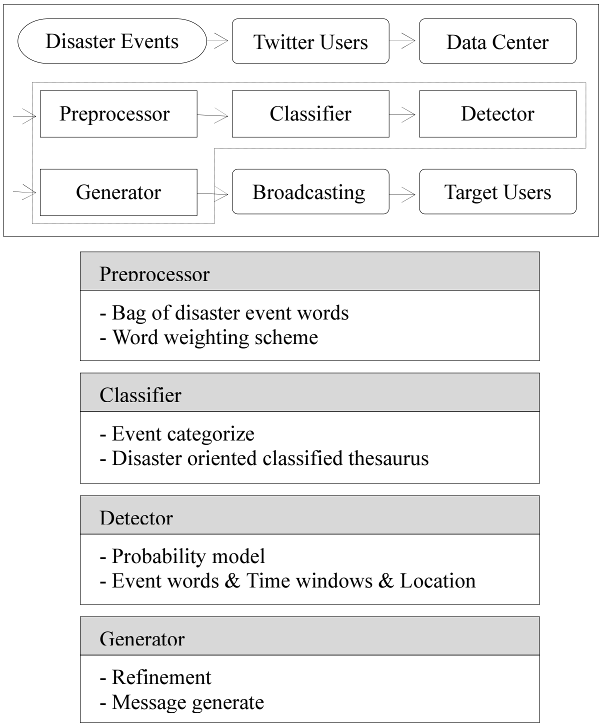

Fig. 5. Event detection processes from Twitter

We use an event specific database to search twitter messages. The disaster related database has bag of words. Consider a bag of words W that stores compact data related to disaster event words. Let W_i represent all words belonging to class i. Thus, the bag of words W is represented as $W = \{W_1, W_2, W_3, \ldots, W_n\}$ assuming there are n classes in the bag of words. For each class, W_i consists of a number of words I_j in the form of $I_j <a_{j1}, a_{j2}, a_{j3}, \ldots, a_{jm}>$, where a_{jk} represents the k^{th} word of I_j.

The input data will be classified into categorical one or none in preprocessing phase for tweet data classification. Fig. 6. shows the tree to classify for disaster event words. Each class has the top most used terms, for example, popular terms are “earthquake”, “help”, and “people”. We can decide what emergency event is occurred through a significant amount of tweet data to contain terms related to the event. This step is to classify tweets into the following categories: for instance, earthquake, tsunami, tornado and so on.

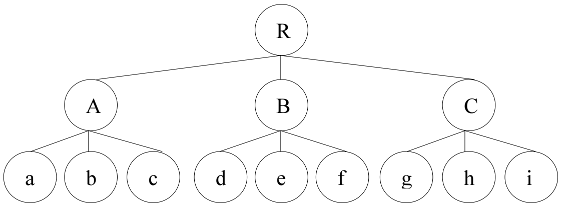

Fig. 6. Tree to classify for disaster event words and locations

To confirm the veracity of information on twitter, the number of tweets and redundancy of terms are significant factors. We can expect that the twitter activity increases proportionally to the significance of the event. Applying of weighting scheme is important for each disaster events.

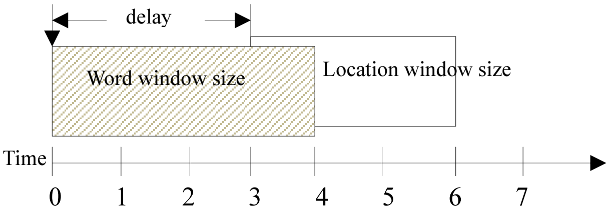

Fig. 7. Limitation on window size for finding disaster events

We use a word weighting scheme which is more specific and high frequency word have a higher weight value. Each word’s weight is calculated based on event specific terms such as tsunami and tornado.

Time series dataset contain data collected over a period of time and can be processed by a sliding window. Detection has the limitation on window size for finding emergency events. Each window in Fig. 7 represents some sequence of Twitter data where a window is a single event with emergency. The amount of tweets contained in the window may be various. A window is generated after an event word is detected and it is removed when the frequency is lower than the certain threshold.

TABLE I
THE SAMPLE VALUES OF EVENT WORDS

Time	*e*Class	*e*Window	*e*Count	*e*Acc	*e*Weight
1	e_1	-	0	0	0
2	e_1	*w*[2,12]	50	50	0.3
3	e_1	*w*[2,12]	80	130	0.4
…	…	…	…	…	…
12	e_1	*w*[2,12]	200	1000	0.25

TABLE II
THE SAMPLE VALUES OF EVENT LOCATIONS

Time	*l*Class	*l*Window(*d*=4)	*l*Count	*l*Acc	*l*Weight
6	l_1	*w*[6,12]	60	60	0.2
7	l_1	*w*[6,12]	60	120	0.3
8	l_1	*w*[6,12]	70	190	0.3
…	…	…	…	…	…
12	l_1	*w*[6,12]	150	800	0.25

The sample values of event words and locations are given on TABLE I and TABLE II respectively. Given the window size is 10 as *w*[2,12], the second with *e*Class is the representation for event category, and the counts for frequency computation is named for *e*Count as fourth column, the weigh values are applied from time =1 to 12 can be obtained as shown in TABLE I. The sample values of event locations can be obtained in the same manner as shown in TABLE II.

IV. PRELIMINARY EVALUATION

A. Data Collection

In this section we conduct a case study to confirm the importance of information based on Twitter during an emergency situation. We manually collected three data sets from the Twitter on two natural disaster and one forest fire. The one natural disaster was tornado that happened in Dallas, Texas and another one was tsunami that occurred near the Indonesia. These data related to tornado and fire forest were gathered that both tweeting on November 1, 2010. Another event is Indonesia tsunami on October 26, 2010. Indonesia has been hit by a massive tsunami that was born after a 7.7 earthquake underneath the ocean floor. We collected their data from the Twitter and the Google during a short duration of time between October 23 and 31. This research examines and compares during this same time period between Twitter activity and Internet news on Google. Data collection timeframes for each event were determined by the nature of the event.

B. Analysis of Sample Characteristics

We used the key word ‘Tornado’ to obtain the sample of tweets and specified the time line as November 1. When we use three different terms ‘Dallas’, ‘Tornado’, and ‘Nov 1’, it returned zero data. 28 tweets were found in our specified time line when we use two key words ‘Dallas’ and ‘Tornado’. 600 twitter data were gathered matching the search key word ‘Tornado’ exactly same specific time period. It is noticeable that the selected data by ‘Tornado’ include slang and places. The data for a particular emergency event contains domain specific words. However the word ‘Tornado’ contains more meaning irrelevant to the emergency. We have to explore not only a way to dynamically refine the corpus related to disaster but also a method to provide a higher accuracy.

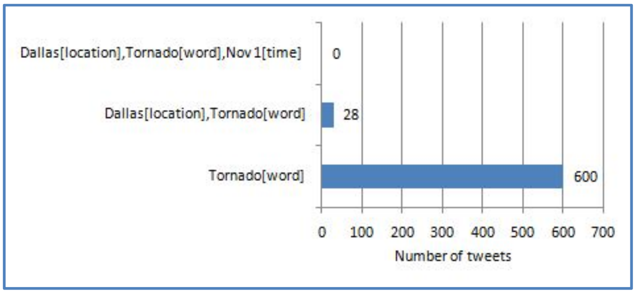

Fig. 8. Number of users that published tweets by key word “Tornado”

In order to collect appropriate sample tweets in the local language, we used the keyword ‘산불’, it means forest fire. 60 tweets were gathered from this search as shown in Fig. 9. We added the location keyword ‘포천’, it was as a role of filter, 14 tweets were remained. We applied further filter to the data to ensure only ‘산불’ were directly connected to the ‘포천’ and ‘오후’. Finally the remaining number of tweets was 3. To further investigate the possibility of detecting emergency based on tweets, we compared number of Google news and number of tweets related to “Indonesia Tsunami” during specified time period as shown in Fig. 10 and Fig. 11 respectively. In Fig. 10, Google news show gradual increase in number of news and Twitter as shown in Fig. 11 shows rapid increase in number of tweets. We can realize that it is possible to detect emergency disasters from Twitter data.

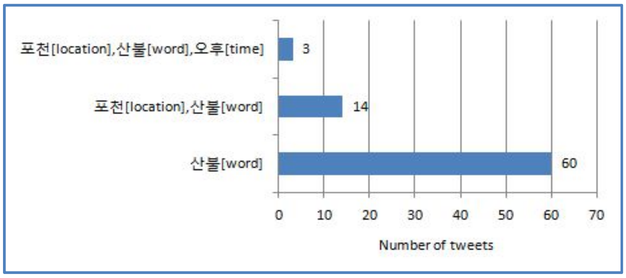

Fig. 9. Number of users that published tweets by keyword “산불(forest fire)”

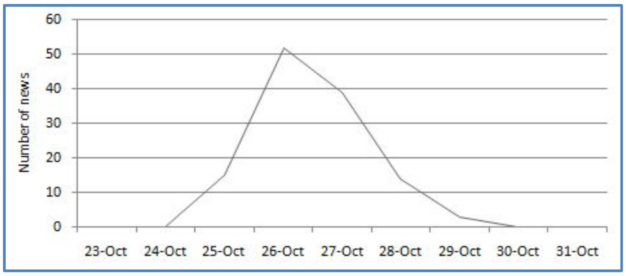

Fig. 10. Number of news by keyword “Indonesia Tsunami” on Google new

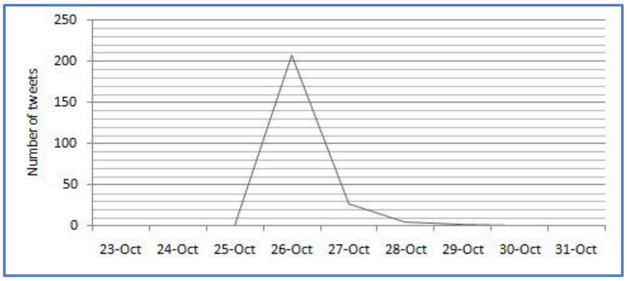

Fig. 11. Number of tweets by keyword “Indonesia Tsunami” on Twitter

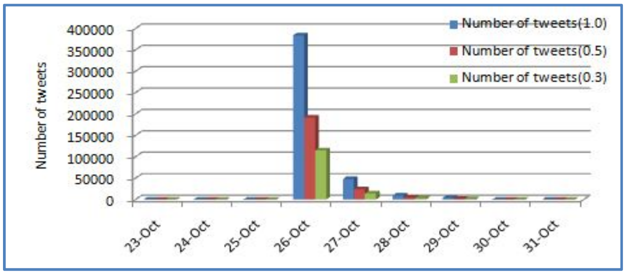

Fig. 12. Estimated number of tweets by keyword “Indonesia Tsunami”

A twitter can designate a tweet as a specific reply to another user, users begin these reply messages with the '@' symbol directly followed by the username. This retweet are a way of getting the attention of a specific user. We computed how many retweet occur in our data sets with a random sample of all followers during our entire data gathering timeframe to estimate total twitter messages. Maximum twitter messages were estimated that approximately 384 thousand tweets were sent at first date of tsunami, on October 26 as shown in Fig. 12.

V. CONCLUSIONS

Twitter is able to send short messages with mobile devices and easily distribute those messages to a wide user. Twitter is becoming popular and is being adopted by many. Some research addresses how social media such as Twitter is being used in emergency and mass convergence situations. This research focused on the features of Twitter use in emergent events. Our goal is in the rapid detection of disaster events such as tsunami, tornadoes, forest fires, and earthquakes. To achieve this objective, we examined and analyzed the keywords in tweets, the frequency of keywords, time, and location information from Twitter. And we described the detection system of disaster events and showed the way to detect a target event from Twitter data. The emergency detection system could pick up Twitter data related to disaster events. In this process, we used bag of disaster event words in order to search specific keywords such as tsunami, tornado and so on. To classify for disaster events, we used a classification tree and to determine a target event, we considered the weighting scheme. In this research, we examined the three disasters during the same time period and compared Twitter activity and Internet news on Google. A significant result from this research was that emergency detection could begin using microblogging service. We expect that similar microblogging technology could use in emergency warning, response and recovery.

REFERENCES

[1] A. Java, X. Song, T. Finin, and B. Tseng, "Why We twitter: An analysis of a microblogging community, *In Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*, 2007.

[2] T. Sakaki, M. Okazaki, Y. Matsuo, "Earthquake Shake Twiiter Users: Real-time Event Detection by Social Sensors," Proc. WWW 2010, pp. 851-860, April, 2010.

[3] B. Longueville, R. S. Smith, and G. Luraschi, ""OMG, from here, I can see the flames!": a use case of mining Location Based Social Networks to acquire spatio-temporal data on forest fires," Proc. ACM LBSN '09, pp. 73-80, November 2009.

[4] National Oceanic and Atmospheric Administration, Available: http://www.ngdc.noaa.gov/hazard/tsu.shtm

[5] National Oceanic and Atmospheric Administration, Available: http://www.ncdc.noaa.gov/oa/ncdc.html

[6] UMBC ebiquity, Available: http://ebiquity.umbc.edu/blogger/2007/04/15/global-distribution-of-twitter-users/

[7] Fire Situation in Republic of Korea, IFFN No. 26, January 2002, p. 61-65, Available: http://www.fire.uni-freiburg.de/iffn/country/kr/kr_2.htm

[8] TechCrunch, http://techcrunch.com/2010/06/08/twitter-190-million-users

[9] M. guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and Dissemination of Citizen Reported and Seismically Derived Earthquake Information via Social Network Technologies," IDA 2010, pp. 42-53, 2010.

[10] D. Zhao, M.B. Rosson, "How and why people Twitter: the role that micro-blogging plays in informal communication at work," In Proc. the ACM 2009 international conference on Supporting group work, pp. 243–252, 2009.

[11] A. Java, X. Song, T. Finin, B. Tseng, "Why We Twitter: An Analysis of a Microblogging Community," Advances in Web Mining and Web Usage Analysis, pp. 118–138, 2009.

[12] C. Honeycutt, S. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," Proc., the 42nd Hawaii International Conference on System Sciences, pp. 1–10 2009.

[13] J. Dixon, C.R. Tucker, "We use technology, but do we use technology? using existing technologies to communicate, collaborate, and provide support," Proc. the ACM SIGUCCS fall conference on User services conference, pp. 309–312, 2009.

[14] B. McNely, "Backchannel persistence and collaborative meaning-making," Proc. the 27th ACM international conference on Design of communication, pp. 297–304, 2009.

[15] K. Starbird, L. Palen, A. Hughes, S. Vieweg, "Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information," Proc. the ACM 2010 Conference on Computer Supported Cooperative Work, .2010

[16] A. Hughes, L. Palen, "Twitter Adoption and Use in Mass Convergence and Emergency Events," Proc. the 2009 Information Systems for Crisis Response and Management Conference, 2009.

[17] S. Vieweg, L. Palen, L. Sophia, A. Hughes, "Collective Intelligence in Disaster," An Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shootings, Proc. the Information Systems for Crisis Response and Management Conference, 2009.

Hongwon Yun
He received his B.S. and the Ph.D. degrees at the Department of Computer Science from Pusan National University, Korea, in 1986 and 1998, respectively. He is a professor at the Department of Information Technology, Silla University in Korea. His research interests include database, temporal database and social network.