Real-Time Social Network Data Mining for Predicting the Path for a Disaster

Saloni Jain, Brett Adams Duncan, and Yanqing Zhang
Department of Computer Science, Georgia State University, Atlanta, USA
Email: {sjain5, bduncan7}@student.gsu.edu, yzhang@gsu.edu

Ning Zhong
Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi-City, Japan
Email: zhong@maebashi-it.ac.jp

Zejin Ding
Hewlett-Packard Company, 5555 Windward Pkwy, Alpharetta, USA
Email: dingzejin@gmail.com

***Abstract*—Traditional communication channels like news channels are not able to provide spontaneous information about disasters unlike social networks, namely, Twitter. This work proposes a framework by mining real-time disaster data from Twitter to predict the path; a disaster like a tornado will take. The users of Twitter act as the sensors, which provide useful information about the disaster by posting first-hand experience, warnings or location of a disaster. The steps involved in the framework are – data collection, data preprocessing, geo-location tagging data filtering and extrapolation of the disaster curve for prediction of susceptible locations. The framework is validated by analyzing the past events using regression with the government warnings. This framework has the potential to be developed into a full-fledged system to provide instantaneous warnings to people about disasters via news channels or broadcasts.**

***Index Terms*—data mining, disaster computing, real-time disaster prediction, regression**

I. Introduction

Social Media has become a very important tool to stay in touch with friends, to market products and services offered by companies and even to make announcements by government agencies and news channels. One of the social networking websites which has gained vast popularity is Twitter. This research work deals with the data obtained from Twitter, which is mined for getting useful information for a real-world scenario, mainly, disaster path prediction. It is discussed in the next section.

A. Twitter and Its Importance

Twitter is an Online Social Network (OSN) used by millions of people all over the world. It enables people to stay connected with their friends, family and colleagues. With advancement in technology, it has become easier to access Twitter using mobile devices like iPhones and iPads. Currently, Twitter has 288 million monthly active users with an average of 500 million tweets being sent per day [1].

Twitter has become an important resource for the field of Data Mining because of its many features. It has a varied variety of users, which can represent a sample of the entire population. The revolution of Information and Communication Technology (ICT) has made it possible for billions of people to access social networking sites ensuring that they have a wide reach of people. They can post messages on the go which ensures the real-time nature of the messages. Compared to emails, this "push" of information is almost instantaneous. Twitter also has a feature of searching or filtering messages which are interesting to a user using hashtags. Users have the freedom to follow or join groups that they like. It also caters for security for its users, where they can decide to post tweets publicly or privately.

Mostly, people post their trivial personal experiences but sometimes they post messages which contain information that are valuable on mining. This information can be about events like politics, traffic jams, riots, fires, earthquakes, storms, etc. Therefore, Twitter can also act as a non-traditional medium to obtain news as people can tweet information which is newsworthy. They can even create messages with news value, which can be used in early warning detection systems. However, the most important feature for this study is the real-time nature of the information dissipation in the Twitter network. It further becomes useful when 80 per cent of the users are mobile users [1] which can provide us with exact geo-location and more up-to-date information.

B. Data Mining for Disaster Management

Data Mining plays a crucial role in extracting useful information from Social Media. The reason is because information in social media contains personal trivial data which is not very enlightening or useful to a large group of people. It is used in many areas for analysis. For

Manuscript received July 23, 2015; revised November 11, 2015.

doi: 10.12720/jait.7.2.81-87

example, companies and organizations can perform sentiment analysis for their products and services [2], [3]. It can also help in detecting and predicting disasters [4] and events such as influenza [5]. This can be the basis of forming early warning systems, one of which was proposed [6]. The subsequent section talks about previous works in disaster management using Twitter.

II. Related Work

Twitter data were mined for real-time earthquake detection [4]. They created an application for earthquake reporting system in Japan. This system consists of two parts–event detection and the probabilistic spatiotemporal model of the event. The detection is performed by making a classifier using a Support Vector Machine. The features used are–keywords in a tweet, the number of words in the tweet and the context of target-event words. For creating a probabilistic spatiotemporal model, the authors assumed that the users are social sensors and their tweets are sensory information. This information is noisy because the users will not always tweet about the event. Some sensors can be very active and others might not be. Just like using physical sensors, these social sensors can be used for Kalman filtering and particle filtering. These are used for estimating the location in ubiquitous computing. They were able to detect 96 per cent of earthquakes reported by Japan Meteorological Agency.

Avvenuti *et al.* proposed a novel architecture for an early warning system and validated it with an implementation in [6]. They made use of social sensing where a group of people or a community provides similar information that might be obtained from a single sensor. The authors dealt with the issue of earthquake detection in Italy. The main steps involved in their study are – Data Acquisition, Data filtering, Event Detection, Damage Assessment and Early Warning. The keywords selected were Italian words for "earthquake" and "tremor". They used Streaming API of Twitter for up-to-date tweets. The filtering phase reduces noise by discarding retweets, replies, tweets containing blacklisted words and tweets by official channels. A more sophisticated filtering strategy was done by classifying tweets as useful and not useful. The features used are: URL, mentions, words, character, punctuation and slang/offensive words. Events are detected by temporal and spatial analysis. For temporal analysis, they created a novel burst-detection method which observes a peak of the number of messages in a time window. They extracted location from the content of the tweet for spatial analysis using TagMe [7]. Damage assessment was done by using a bigger set of general keywords, images and videos. The results obtained from the experiment were checked with official data to show that earthquakes with a magnitude equal or greater than 3.5 on Richter scale can be timely detected with 10 per cent False Positives.

There is a need for an automated disaster management system that can recommend suitable action patterns in case of a disaster. These can deal with informing about the shelter in case of a disaster emergency or maybe how to travel from one place to another. One of the methods was suggested by Nguyen *et al.* [8]. They built an earthquake semantic network using human activity on Twitter based on Web Ontology Language. A Twitter activity was defined by five attributes, namely–action, object, location, time and actor. The network is connected by the relationships–*Next* and *BecauseOf.* They also created automatic data for the network. This network was further used to recommend suitable actions in the face of a disaster. They found out that their learning model Conditional Random Field (CRF) outperformed the baseline method (which used syntactic parser with the linguistic pattern for training data) and the previous extraction method [9]. One of the problems with the works of Banerjee *et al.* [9] was that the list of actions and objects had to be prepared before extraction.

Another focus for disaster management is to make sure that information is spread as widely as possible. Social media is the fastest way of information diffusion where general population as well as government agencies can respond to requests for assistance, information and announcements. Retweeting on Twitter is the most efficient way to spread an original message beyond the author's network. Zhu et al. built a predictive model for finding the retweeting decision of a user [10]. They have found out the factors affecting the retweet decision. The features can be classified into three categories–contextual influence, network influence and time influence from which a set of features are found. A Monte-Carlo simulation was also performed for finding how the information propagates in Twitter network. Even though the information on social media is important for spreading awareness, credibility of the information might be a problem. Kongthon et al. analyzed the content of Twitter messages and the characteristics of Twitter users which can be used for better disaster management [11]. They used keyword analysis and rule based approach for classifying tweets into five categories–Situational announcements and alerts, support announcements, requests for assistance, requests for information and other. They also classified users on the basis of number of followers and retweets. It is of utmost importance to make sure that the information being used for disaster preparedness and response is current and true. Hence, all the factors must be considered.

Disaster Management is also useful for topical analysis. In the aftermath of the 2011 Japan tsunami [12], there was a commotion due to damage to conventional communication networks and power outages. Family members wanted to confirm safety of each other. There was a need to exchange demands and opinions. In [13], Murakami *et al.* discussed text mining techniques on tweets. They are used to find areas where there are a shortage of supplies and other peoples' needs. Two approaches are used; one that uses simple keyword matches, and another that uses syntactic pattern dictionaries. In the keyword match approach, the system found areas needing supplies, but there was noise in some of the tweets found in which the tweet contained a keyword but was unrelated to the disaster. In the syntactic approach, a syntactic pattern dictionary was used to

identify things that were in short supply. For example, "cannot buy <noun>," and "<noun> is sold out" are patterns that were used. Using this approach, the most frequent nouns that had shortages were water, battery, rice, gasoline, and toilet paper.

III. Methodology

This section presents the details of the experiment to trace the trajectory of a disaster and use it for issuing the forewarning to the susceptible people. The main steps followed are–Data acquisition, Data preprocessing, Extracting information from tweets, Clustering data points for filtering, Curve fitting to get the trajectory for the disaster, Extrapolation of the graph, Validation of the extrapolated graph. The details of the experiment are explained in the next sub-sections [14].

A. Event Studied – Alabama Tornado October 2014

Alabama in US was hit by a series of small tornadoes on 13th October 2014. Some of the states where tornado or strong winds were reported are Alabama, Louisiana, Florida, Georgia, Tennessee, Arkansas, Missouri and Illinois. Tornado reports are available on NOAA's National Weather Service Storm Prediction Center [15].

B. Data Acquisition

The role of collecting data is extremely important for any experiment as all further operations are performed on the data obtained in this step. The data should be able to represent all the information that is required for the cause, which in this experiment, is the prediction of the path of the tornado that hit Alabama in October 2014. There are two phases in this step – identifying an apt set of keywords and collecting data from Twitter using an appropriate interface.

1) Keyword selection

Twitter produces around 6000 tweets per second [16], thus increasing the probability of noise in the dataset. To have the relevant data for the experiment, the keywords should be selected carefully. They should neither be very specific nor too general. If the keywords are very specific to a certain event, it might limit the tweets which are extracted. On the other hand, using very general keywords like, #StrongWinds will lead to too much noise in the collected dataset. Therefore, it becomes extremely crucial to choose the correct set of keywords. The next phase is Data Collection which is discussed in the next sub-section.

2) Data collection from twitter

The system made use of Twitter Search API through the interface developed by Martin Hanksey called Twitter Archiving Google Spreadsheet TAGS v5.1 [17]. It requires authentication which has been mandated by Twitter for all its APIs. However, since it uses the Search API, there are some limitations. It can over-represent the more influential users which might lead to some bias in the data. Also, the API can access only a subset of all the tweets but we obtained a large number for performing the experiments. Streaming API would have given a more complete dataset which would have given a more accurate result because the tweets obtained would be in real-time. For building a framework, the Search API worked well. The event had already occurred rendering the Streaming API of not much use for the system. It is better for implementation for real-world applications.

C. Data Preprocessing

A set of more than 4000 tweets was used to run the experiment using the keywords mentioned in the results section. This set was taken for the time period of two days, namely – 12th and 13th October. This set can obviously have tweets which won't be useful for this study. It is important to eliminate those to save computation power and prevent noise in the data. The dataset should only contain English tweets as the event in consideration, i.e. the Alabama Tornado in October 2014, occurs in United States. Even if people from around the world tweeted about it, we will want to eliminate those if they are in another language because they will not provide us with much information. However, the system is just an initial framework, which can be extended to multiple languages.

The first step was to remove all non-English tweets. This was done by checking the language feature of a tweet. If it was not 'en' it was discarded. Along with getting the language of the tweet, we also obtained the geo-coordinates for those tweets since the system already made a call to the Twitter API. Location extraction is discussed in the next section. The next step was to remove spam tweets. A list of 165 spam words was compiled using numerous blacklists. The tweets which contained spam words were removed.

D. Time and Location Extraction

For predicting the path of a disaster, it is safe to assume that the disaster has at least already started. Once occurrence of the event is made certain, the possible path of the disaster needs to be traced using the dataset. This can be done using the extraction of time and location from the tweets. The location field can determine the where aspect and time field can determine the when aspect of a disaster. Thus, it is important to do temporal and spatial analysis which mainly deals with getting the time and the most accurate locations from a tweet.

1) Time extraction

It is fairly straight-forward to get the time of the tweet that was created. For the experiment, I decided to use the local time at which the tweet was created for understanding the association between the actual event and the predictions by the system. The UTC time can just as well be used. In fact, it will be much more useful if the event being considered is a global event.

2) Location extraction

Twitter provides its users with the option to geo-code their tweets which can offer the exact locations, especially in mobile devices. But, less than 1% of the data that was collected had geo-locations associated with them. This leaves a very small number of tweets to work with. Sakaki *et al.* proposed a workaround to this solution. He proposed that the location of the Twitter account can be used to get an approximate location [4]. However, this

is just an approximate location. The best way to get the most information from a tweet about its location is through the content of the tweet [6]. In the present work, the TagMe API developed at University of Pisa, Italy to get the "spots" from the content of tweets which can potentially be location names [7].

For the proposed experiment, location is very important. Hence, location was extracted from the three sources for a single tweet – (1) geo-coordinates, (2) the content of the tweet and finally (3) the location of the Twitter account. The next step is to filter out noise and keep only the samples which will be useful for getting the curve for trajectory prediction.

E. *Clustering the Data for Filtering*

It is important to find the useful points that will lead us to the solution. Clustering can help us achieve that. It will group nearby locations in the same clusters. The clusters that lie far away from other clusters can be eliminated reducing the noise and leading to a better polynomial curve. For this study, DBSCAN [18] was experimented with. In DBSCAN clustering, the clusters are based on the density of data points. They are areas of high density which are separated by areas of low density. It will have a set of core samples and a set of non-core samples. Core samples are those points which have at least a given number of minimum samples, *min_samples* within a specified distance, *eps*. This algorithm is beneficial for the experiment as it can eliminate outliers and also perform sampling in the process. Clusters that have lesser number of data points as compared to others should be removed. Also, data points which are not part of any clusters are removed as noise.

F. *Curve Fitting and Extrapolation for the Trajectory of the Disaster*

Latitude and longitude of the locations which were struck by disaster or will possibly be hit by disaster can be visualized as functions of time as the disaster progresses. The aim was to find a polynomial curve which will satisfy the existing points and also give a good prediction for the future points based on obtained values. For the simplicity of the equation, latitude and longitude were taken as independent of each other. Therefore, two polynomial functions of time are needed, one for latitude and one for longitude. It was experimented with linear regression and non-linear regression with degrees two, three and four. Higher degrees were also tried but they did not change the curve significantly.

The curve was first found for the data for dates 12th and 13th October 2014 separately for latitude and longitude. For finding the future locations that might be susceptible, the curve was extrapolated for future time values. A confidence level of 95% was kept and prediction bounds were also found. Validation was done by plotting the actual points of the future date: 14th October 2014 in this case and plotting a curve using that data. The plots are discussed in the results section. Using the extrapolated curve, future value pairs are obtained which are then reverse-geocoded to obtain the actual location addresses which will tell us the places which should be warned.

IV. RESULTS

This section discusses the results obtained at each step of the experiment.

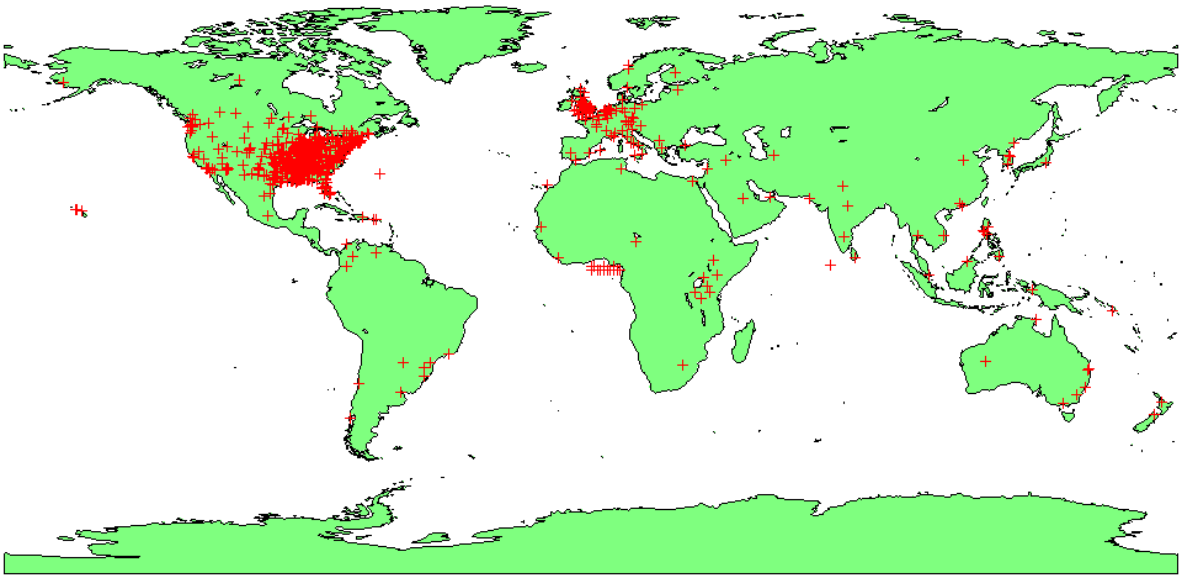


Figure 1. A world map containing all the locations that were extracted from the tweets containing the keywords.

A. *Data Acquisition*

As mentioned earlier, keyword selection is very important to capture all the relevant data. For this study, three keywords were selected–'*Tornado*', 'AlabamaTornado' and '*storm'*. The type of event can generally serve as good keywords since they'll capture more data and prevent the loss of relevant data. For my study, the training data was taken for the dates 12th and 13th October 2014. A total of 4147 tweets were extracted for these days. Some of the fields that were extracted are *–text, time, geo-coordinates and iso_language_code.* Similarly, future data which acted as the validation data was taken for the date 14th October 2014. A total of 2920

tweets were collected with the same fields. Tweets from 15th to 17th October 2014 were also collected but data till 14th October was observed as sufficiently enough for the experiment to prove the validity of the framework.

B. Data Preprocessing

The data was processed for getting only the English tweets and removing the spam tweets. The training data resulted in 3909 tweets which was a big enough number to work with. On the other hand, the validation data, that is, the data for 14th October 2014 resulted in 2390 tweets.

C. Time and Location Extraction

Using the locations and time obtained, a tuple is created containing the geo-coordinates for the exact point on earth, and local time. The extracted points are displayed on a world map in Fig. 1. The actual affected points are also present on the world map which was generated using locations from the tweets. However, there are some more points which do not reflect the actual data. There is a need to remove these points which is achieved by filtering.

D. Clustering the Data for Filtering

On experimentation, it was found that the best number of clusters to obtain is 3 where the cluster with minimum number of samples can be removed. This was obtained by keeping the *min_samples* value as 160 and *eps* = 2.8. The non-core samples were eliminated as they can be represented by more important core samples. The cluster with the minimum number of points was also removed. Using this method, the data points were reduced from 4018 to 2144 data points. The points used for curve fitting are shown in Fig. 2.

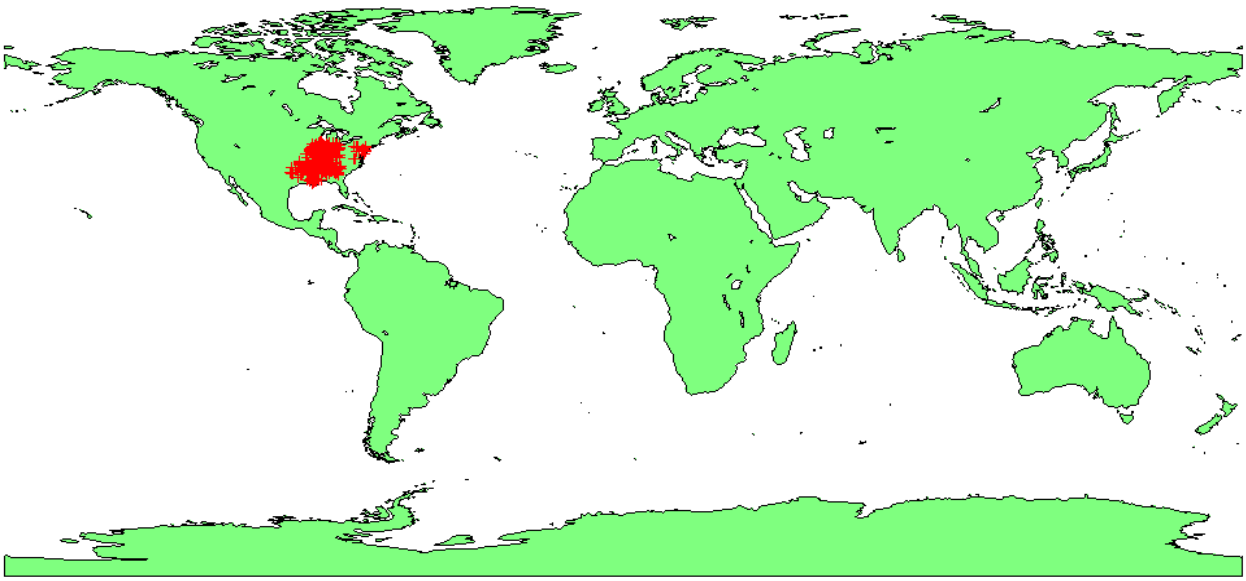


Figure 2. Points after complete filtering on a world map.

E. Curve Fitting and Extrapolation for the Trajectory of the Disaster

Results using linear and non-linear regression for extrapolation of the trajectory are discussed below. Since, latitude and longitude were assumed to be independent of each other; there are separate curves for them. Linear regression will give us a linear equation which will be the best fit for given data. Fig. 3a and Fig. 3b show the plot for Latitude vs. Time and Longitude vs. Time with prediction bounds with 95% confidence bounds. We can see that almost all the points are getting included in the prediction bounds. The detail about the curve is presented in Table I where p1 and p2 are the coefficients for the linear polynomial equation.

On extrapolation, it can be observed that the curves can represent most of the points of the validation data, i.e. the data of 14th October 2014. This can be seen in Fig. 4a and Fig. 4b. The blue dots represent the original data and the green dots represent the validation data.

Therefore, using the equation parameters of Table I, given a value for time, the values of corresponding latitude (La) and longitude (Lo) can be found with a confidence of 95%. These values are then reverse-geocoded to result out the locations. Some of the locations which were obtained were places in Texas, Arkansas, Missouri, Indiana, Kentucky and Michigan. These locations are the future values which need to be warned for the disaster. As a more future value of time is taken, the locations start coming out to be in Canada.

TABLE I. LINEAR MODEL FOR Y VS. TIME

y	Coefficients with 95% Confidence Bounds					
	p1			p2		
	Lower	Exact	Upper	Lower	Exact	Upper
La	2.646	3.323	3.999	-1.68E+05	-1.39E+05	-1.11E+05
Lo	0.009029	0.1594	0.3099	-88.2	-88.05	-87.9

Quadratic regression was also tried. However, on comparing linear and non-linear (quadratic) regression, it was seen that linear gave much more accurate results as they are comparable to the locations which were reflected in official government records. On visualizing the data, the points were linear in nature and hence it intuitively made sense that the linear curve was a better fit. Degrees

higher than 2, further worsened the fit of the curve and hence, were not studied further.

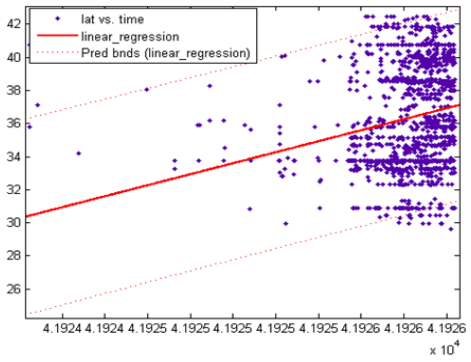


Figure 3a. Linear regression for Latitude vs. Time for training data.

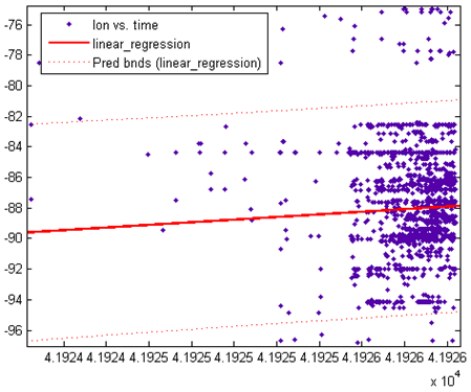


Figure 3b. Linear regression for Longitude vs. Time for training data.

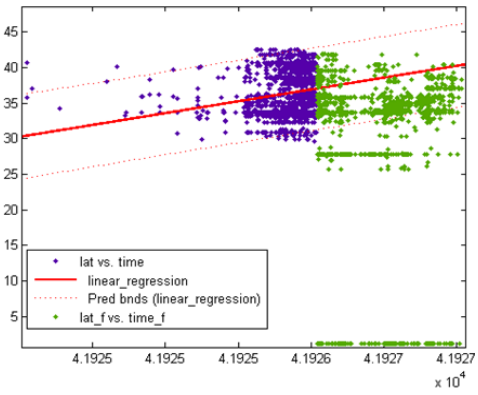


Figure 4a. Linear curve for training data extrapolated for Latitude vs. Time.

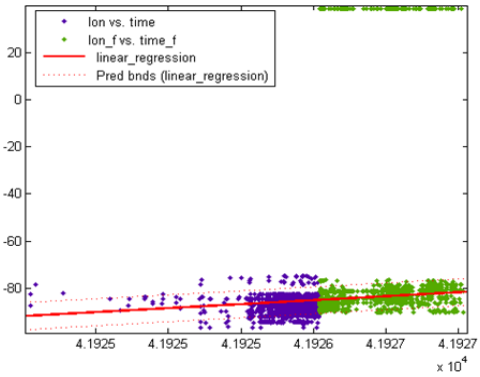


Figure 4b. Linear curve for training data extrapolated for Longitude vs. Time

V. Conclusion and Future Work

Earlier works in disaster management on Twitter dealt with disaster detection, early warning prediction system, dealing with aftermath of a disaster, etcetera. However, the problem of finding the trajectory has not been looked into much by the research community. This research work proposed a framework for finding out the trajectory of a disaster for real-time data extracted from Twitter. The steps that were involved are – 1) Data Acquisition, 2) Data Preprocessing, 3) Time and Location Extraction, 4) Filtering of Data using Clustering, 5) Curve Fitting and Extrapolation for finding the trajectory of the disaster. Finally, the framework was validated by looking at the validation data and comparing the locations obtained by the extrapolation of the curves. The locations were also compared to the official government records. This framework can be implemented in disaster management systems by refining some of the techniques discussed in the paper. A complete independent system can be implemented which will take care of detecting an event, predicting the location and then providing relief for the disaster struck areas.

The framework tested was a preliminary framework to prove the validity that the path of a disaster can be predicted. However, the efficiency of the path can be increased by considering the real-time nature of the data. Like mentioned, that can be done by using the Streaming API to collect data at predefined intervals. This data can be treated separately and then the results can be aggregated together to get the full picture of the disaster.

Acknowledgment

This work was supported in part by the NSF REU grant # 1156733.

References

[1] About. (January 18, 2015). [Online]. Available: https://about.twitter.com/company

[2] B. Agarwal, I. Xie, O. Vovsha, Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *LSM '11 Proc. the Workshop on Languages in Social Media*, 2011.

[3] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the OMG!" in *Proc. the Fifth International Conference on Weblogs and Social Media*, Barcelona, 2011.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. WWW 2010*, Raleigh, 2010.

[5] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Proc. the First International Workshop on Cyber-Physical Networking Systems*, 2011.

[6] M. Avvenuti, S. Cresci, M. N. L. Polla, A. Marchetti, and M. Tesconi, "Earthquake emergency management by social sensing," in *Proc. Pervasive Computing and Communications Workshops*, Budapest, 2014.

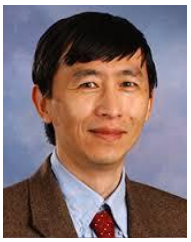
[7] P. Ferragina and U. Scaiella, "Fast and accurate annotation of short texts with Wikipedia pages," *IEEE Software*, vol. 29, pp. 70-75, 2012.

[8] T. M. Nguyen, K. Koshikawa, T. Kawamura, Y. Tahara, and A. Ohsuga, "Building earthquake semantic network by mining human activity from twitter," in *Proc. IEEE International Conference on Granular Computing*, 2011.

[9] N. Banerjee, D. Chakraborty, K. Dasgupta, A. Joshi, S. Mittal, S. Nagar, A. Rai, and S. Madan, "User interests in social media sites: An exploration with micro-blogs," in *Proc. CIKM*, 2009.
[10] J. Zhu, F. Xiong, D. Piao, Y. Liu, and Y. Zhang, "Statistically modelling the effectiveness of disaster information in social media," in *Proc. IEEE Global Humanitarian Technology Conference*, 2011.
[11] Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of social media during a natural disaster: A case study of 2011 Thai flood," in *Proc. PICMET '12: Technology Management for Emerging Technologies*, 2012.
[12] C. Taylor. Twitter Users React to Massive Quake, Tsunami In Japan. (March 10, 2011). [Online]. Available: http://mashable.com/2011/03/10/japan-tsunami/
[13] Murakami and T. Nasukawa, "Tweeting about the tsunami? - mining twitter for information on the Tohoku earthquake and tsunami," in *Proc. WWW 2012*, Lyon, France, 2012.
[14] Jain and Saloni, "Real-time social network data mining for predicting the path for a disaster," Thesis, Georgia State University, 2015.
[15] SPC Storm Reports - 20141013's Storm Reports (1200 UTC - 1159 UTC). NOAA's National Weather Service. [Online]. Available: http://www.spc.noaa.gov/exper/archive/event.php?date=20141013
[16] Twitter Usage Statistics. Internet Live Stats. [Online]. Available: http://www.internetlivestats.com/twitter-statistics/
[17] M. Hanksey. Twitter Archiving Google Spreadsheet TAGS v5. (February 15, 2013). [Online]. Available: https://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/
[18] "Scikit-learn: Machine Learning in Python," Pedregosa, et al. *JMLR*, vol. 12, pp. 2825-2830, 2011.

Saloni Jain was born in India on January 7th, 1992. She completed her Bachelors in Computer Science from Indraprastha Institute of Information Technology, Delhi, India in 2013. Currently, she is doing her Master's in Computer Science from Georgia State University, Atlanta, USA. She is expected to graduate in May 2015. Her research interests include machine learning, data mining and web intelligence. She was working as a Graduate Assistant in the Department of Student Accounts at Georgia State University. In summer of 2014, she worked as a Programming Intern at Moda Operandi, New York for building the administrative side of customer management system. She worked in the summer of 2012 as a Research Intern at Infosys, Pune, India which resulted in a white paper. The patent on the paper is still pending. She has also worked as a Programming Intern at NCAP, Delhi, India and IIIT-Delhi, India.

Brett Duncan Jr is a Master of Science student studying computer science at Georgia State University, Atlanta, USA. He received the Bachelor of Science degree in computer science with a minor in mathematics from Georgia State University in 2013. His research interests include data mining and machine learning.



Yanqing Zhang is a full Professor of the Computer Science Department at Georgia State University, Atlanta, USA. He received the Ph.D. degree in computer science from the University of South Florida in 1997. His research interests include hybrid intelligent systems, computational intelligence, machine learning, data mining, bioinformatics, brain informatics, health informatics, computational web intelligence, green computing, granular computing, Yin-Yang computation, nature-inspired computing, security, cloud computing. He is a member of the Bioinformatics and Bioengineering Technical Committee of the IEEE Computational Intelligence Society. He received Outstanding Academic Service Award at IEEE 7th International Conference on Bioinformatics & Bioengineering, Achievement Award of the 2007 World Congress in Computer Science, Computer Engineering and Applied Computing, and 2005 IEEE-Granular Computing Outstanding Service Award at 2005 IEEE International Conference on Granular Computing.

Ning Zhong is the head of the Knowledge Information Systems Laboratory and a professor in the Department of Life Science and Informatics, Maebashi Institute of Technology, Japan. He is also the director and an adjunct professor in the International Web Intelligence Consortium (WIC) Institute, Beijing University of Technology. His research interests include Web intelligence, brain informatics, data mining, granular computing, and intelligent information systems. Zhong has a PhD in the interdisciplinary course on advanced science and technology from the University of Tokyo.

Zejin Ding is a senior security researcher at the Software Security Research team of HP. His research includes application/network security, machine learning, social networks, and big data analysis. His main interests are in integrating data analysis methods to application security and social network areas, to create more intelligent systems to detect security vulnerabilities. He had actively spoke in various top security conferences, including Black Hat, RSA USA, (ISC) ²Security Congress, Security B-Sides (Las Vegas, Atlanta, Austin, Boston), GFIRST, Virus Bulletin, etc. His previous important studies were on identifying new threats on social networks, such as fake profiles on Facebook, and Twitter tweet and follower botnet, etc. Zejin worked at Barracuda Labs before and earned his Ph.D from Georgia State University in 2011.