Sub-Event Detection During Natural Hazards Using Features of Social Media Data

Dhekar Abhik, Durga Toshniwal
Department of Electronics and Computer Engineering,
Indian Institute of Technology Roorkee,
Roorkee, India
abhik_dhekar@hotmail.com, durgatoshniwal@gmail.com

ABSTRACT

Social networking sites such as Flickr, YouTube, Facebook, etc. contain a huge amount of user-contributed data for a variety of real-world events. These events can be some natural calamities such as earthquakes, floods, forest fires, etc. or some man-made hazards like riots. This work focuses on getting better knowledge about a natural hazard event using the data available from social networking sites. Rescue and relief activities in emergency situations can be enhanced by identifying sub-events of a particular event. Traditional topic discovery techniques used for event identification in news data cannot be used for social media data because social network data may be unstructured. To address this problem the features or metadata associated with social media data can be exploited. These features can be user-provided annotations (e.g., title, description) and automatically generated information (e.g., content creation time). Considerable improvement in performance is observed by using multiple features of social media data for sub-event detection rather than using individual feature. Proposed here is a two-step process. In the first step, clusters are formed from social network data using relevant features individually. Based on the significance of features weights are assigned to them. And in the second step all the clustering solutions formed in the first step are combined in a principal weighted manner to give the final clustering solution. Each cluster represents a sub-event for a particular natural hazard.

Categories and Subject Descriptors

H.3.3 [**INFORMATION SEARCH AND RETRIVAL**]: Clustering, K.4.3 [**ORGANIZATIONAL IMPACTS**]: Computer-supported collaborative work

General Terms

Design, Human Factors

Keywords

Sub-event detection; emergency-situation awareness; social-media; natural-hazards

1. INTRODUCTION

For the people looking for sharing their personal news and information, the social networking sites like Flickr, YouTube, Facebook, etc. has emerged as a popular destination. Due to this reason these sites holds a large amount of user contributed data for a wide variety of events ranging from popular, widely known organized events (e.g., a concert by a popular music band) to various natural hazards(e.g., earthquakes, hurricanes, wildfires, etc.). This work focuses on the natural hazards (e.g., earthquakes, floods, etc.).

During the emergency situation events it is important to acquire as much information about the event as possible. This can be helpful for the aid-team to carry out the rescue and relief activity management during such events. It can also be helpful for the people to remain updated about the ongoing situation. But it is not feasible for the command center or the news agency to have their correspondents cover the entire affected area for gathering information about the ongoing event for the entire period of time.

In this case they can rely on the data shared by the people on the social networking sites for getting updates about the event. Extracting information about the latest trends using the social networking sites is one of the latest topics under research. Various researchers have studied and proposed various approaches for extracting useful information from the social networking sites. For example, social media data can be used for predicting election results, getting reviews about some product, etc.

Similarly, using social media data for situation awareness during emergency situations is also one of the latest topics for research. For any major event occurring there are many sub-events associated with it. For example during a natural hazard like earthquake, there might be sub-events like a bridge getting damaged at one location and some famous building getting damaged at some other location. During flood, crest observed at different areas/cities can also be considered as sub-events. The problem discussed in this paper is to identify such sub-events in a particular natural hazard event using the data provided by the users on social networking site.

Identifying sub-events and their associated documents over social media sites is a challenging problem as the information provided by the social media users is inherently noisy and heterogeneous. We can say that the problem is much similar to the topic discovery task like the one used for news event identification in a continuous stream of news data. But for social networking site data, the approach used for traditional news articles cannot be used. Because news articles have structured text while social networking sites may have unstructured data.

Even though the data obtained from social networking sites presents the challenges, they also provide opportunities for using the metadata or features associated with them like title, description, location, upload or creation time, etc. Sub-event detection using individual feature can prove to be unreliable and noisy. Hence here more than one feature is taken into

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

consideration. Based on the type of feature, variety in similarity measure can be observed. Clusters are formed considering each feature individually in the first step. Weights are assigned to features based on their significance. Then the clusters formed in the first step are combined in a weighted manner to form combined clusters in second step. Each of these combined clusters now corresponds to the sub-event in the particular event.

In the next section we discuss the related works done in the field of using social media data for emergency situations and sub-event detection. Section 3 discusses the overview of the framework for social media data exploration. Section 4 discusses sub-event detection using the method for obtaining the clusters taking features of the social media data into consideration. Section 5 discusses the experiment and results obtained.

2. RELATED WORK

Citizen-driven information sharing system during emergency situations has attracted the attention of researchers. The work in [5] describes the role of public libraries in the relief of the 2004 and 2005 Gulf Coast hurricanes. Residents in affected area used the public libraries to check for updates and searching missing relatives. The study in [4] investigates how a local farmer community used a grass roots computer network during disease crises in the UK. The network provided a platform for the local farmers to exchange information, to communicate and provide emotional support. Reference [14] shows how an online forum was used for coordinating the donated goods distribution. The authors of [6] proposed the Emergency Response Grid – an information infrastructure that supports citizen-driven emergency response.

Social media in emergency cases is getting increasingly important, comparable to its intense utilization in private and commercial areas to communicate different situational, news and contextual information. Extensive research has been done on social media in disasters by studying the microblogs [15]. The study identifies the types of emergency messages related to emergency situations (Red River Floods 2009 and Oklahoma grassfires 2009). Reference [12] discuss about the impact of earthquake in Japan on twitter and presents an approach for earthquake detection using twitter. Reference [9] presents the study of use of a Chinese micro blogging system, Sina-Weibo immediately after a major disaster – the 2010 Yushu earthquake. Besides using textual messages, like microblogs [13], visual information can also be used. Flickr, for example is a valuable source of information to detect events, as the work in [10] shows.

Problem definition: From the data related to a particular natural hazard posted on various social networking sites, the goal is to find sub-events associated with the natural hazard.

The problem is similar to topic discovery and tracking of news events. The topic discovery and tracking found notable importance for discovering and organizing the news events in a continuous stream. However they are not suitable for social media data. This is because social network data might be unstructured. Reference [2] discusses the use of features of Flickr photographs for event detection. For sub-event detection in social media [8] presented an approach of using self-organizing maps for clustering of Flickr photos into clusters representing sub-events. Using self-organizing maps for large amount of data is computationally expensive and not suitable for real-time environment.

To find the similarity between social network documents, each textual feature (e.g.: title, description, tags, etc.) can be represented as *tf-idf* weight vector and the cosine similarity metric is used as the feature similarity metric [7]. Also while generating *tf-idf* weight vector, stop words are excluded and all the stems and synonyms of the word are considered same.

Dates are represented as values which are the number of days elapsed since the Unix epoch (i.e., since January 1st, 1970) and the similarity of two date values (say d_1 and d_2) is computed as,

if d_1 and d_2 are more than a year apart then similarity is taken as 0, else similarity is given by,

$$1 - \frac{|d_1 - d_2|}{y} \qquad (1)$$

where y is number of days in a year .

Location is another important feature in social media documents. In can be represented as textual representation or using geographical coordinates (i.e., latitude-longitude pairs). For textual representation of location, name of each location is used as a token or element and the Jaccard similarity metric is used as the feature similarity metric.

$$sim_{Jaccard} = \frac{N_{11}}{N_{10} + N_{01} + N_{11}} \qquad (2)$$

where N_{11} is the number of elements common for both the documents, N_{10} is the number of elements that occur for first document but not for second and N_{01} is the number of elements that occur for second document but not for the first.

To compute the proximity of two locations as geographical coordinates Haversine distance can be used. Suppose $L1 = (lat1,long1)$ and $L2 = (lat2,long2)$ are geographical coordinates of two locations, then similarity between them is given by $1\text{-}H(L1,L2)$, where $H(.)$ is the Haversine distance, a widely accepted metric for geographical distance [3].

$$X1 = \left(sin\left(\frac{lat2 - lat1}{2}\right)\right)^2 \qquad (3)$$

$$X2 = \cos(lat1)\cos(lat2)sin\left(\frac{long2 - long1}{2}\right)^2 \qquad (4)$$

$$d = H(L1, L2) = 2rarcsin(\sqrt{X1 + X2}) \qquad (5)$$

3. FRAMEWORK FOR SOCIAL MEDIA DATA EXPLORATION

During emergency situation getting information from different sources is necessary. Identifying sub-events helps in assessing the situation. Information provided by the users on social network can be used for situation awareness in emergency situation. However manually identifying the sub-events is a cumbersome. Figure 1 shows a framework which allows automatically analyzing data from different social networking sites in case of large-scale emergency situation [8]. The analysis is conducted using the metadata (e.g., tags and title) associated with the content found on social media platforms like YouTube, Flickr, Twitter, etc. An interface collects the streaming data from the social media sites.

First module of the framework collects the data from the social networking sites and performs pre-selection of the data. The streaming interface is the one which collects the data from social networking sites. For pre-selection we can use key-word based search for getting topic specific data. Here streaming interface and pre-selection modules can be easily implemented using APIs of the social networking site. Burst detection module discussed in [17] can be used to detect the occurrence of the event. For example, when we get a rapid increase in the occurrence of certain words related to natural hazards we can trigger the modules which records the data and sub-event detection module. For example if a burst is observed in the occurrence of word 'earthquake' within certain predetermined interval of time it can be said that an earthquake has occurred somewhere and the lower modules of the framework can start recording and analysis of data.

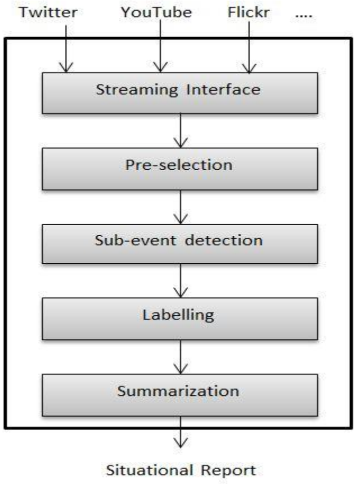

Figure 1 Social media data exploration framework

Next module, after pre-selection of the social media documents, is sub-event detection. Sub-events during a particular event are the smaller events separated by time or location. Sub-event detection is the main area under focus.

After the identification of sub-events it is necessary to label them. This is performed by labeling module. For labelling of sub-events the date and the location of centroid of cluster can be used. Also top few words of the *tf-idf* vector of textual features of the documents in the cluster can be used for labelling the sub-events.

The last module performs the summarization of the documents associated with the sub-events. And as a whole it gives the summarization of entire event which can be included into the situational report.

4. SUB-EVENT DETECTION

This paper focuses on the sub-event detection module of the social media data exploration framework discussed in earlier section. The approach discussed over here is a two-step process. In the first step clusters formed by taking different features of social media documents individually. In the second step the clustering solutions obtained in the first step are combined to give a single clustering solution. Each of these clusters obtained in final clustering solution represents a sub-event.

Using different features of the documents like the title, description, location, uploading or creating date and time, etc. and their similarity measures different set of clusters can be generated. Let us say $(F_1,\ldots,F_k)$ are the features of the documents and using their appropriate similarity measures different clustering solutions $(C_1,\ldots,C_k)$ can be formed as shown in the Figure 2. Here single pass centroid similarity technique [2] is used, which works as follows,

1. Given a threshold τ, a similarity function E and the data points to cluster $D_1,\ldots,D_n$, this algorithm considers each data point D_i in turn and computes its similarity $E(D_i,c_j)$ against each cluster c_j, for $j=1,\ldots,m$, where m is the number of clusters (initially $m=0$).

2. If no cluster is found with the centroid whose similarity to D_i is greater than τ, then a new cluster is formed containing data point D_i and with the centroid value as the value of D_i.

3. Otherwise, D_i is assigned to the cluster which gives maximum value for $E(D_i,c_j)$ and after adding D_i to cluster j new value of c_j is computed. Depending on the feature of data point being considered, the centroid for the cluster is either the average *tf-idf* score per term (for textual features such as title, description, tags), the average days (for date), or the mid-point (for location) of all data points in that cluster.

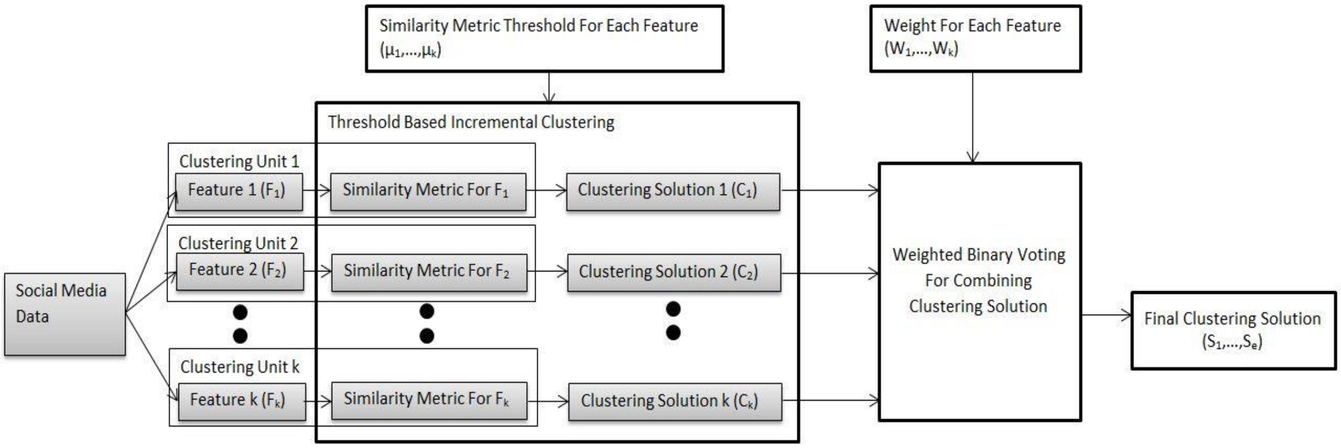
Similarity Metric Threshold For Each Feature
$(\mu_1,\ldots,\mu_k)$
Weight For Each Feature
$(W_1,\ldots,W_k)$
Threshold Based Incremental Clustering
Clustering Unit 1
Feature 1 (F_1)
Similarity Metric For F_1
Clustering Solution 1 (C_1)
Clustering Unit 2
Feature 2 (F_2)
Similarity Metric For F_2
Clustering Solution 2 (C_2)
Social Media Data
Clustering Unit k
Feature k (F_k)
Similarity Metric For F_k
Clustering Solution k (C_k)
Weighted Binary Voting For Combining Clustering Solution
Final Clustering Solution
$(S_1,\ldots,S_e)$

Figure 2 Conceptual view of overall process

For each of the clustering unit associated with each metadata or feature of the social media data the threshold parameters can be tuned during the training phase. Each clustering unit is trained using a labeled data set annotated with the sub-events of the event and the performance of each clustering unit is evaluated at different thresholds and the one which yields the highest performance is chosen. To measure the performance of the clustering unit Normalized Mutual Information (NMI) score is used. NMI measures how much information is shared between the two sets of partition. In the case under consideration it is useful to measure how much common information is shared in between the actual ground truth and the clustering result obtained using the clustering unit. Hence it measures the performance of the clustering unit.

The clustering solutions $(C_1,...,C_k)$ obtained from the clustering units can be seen as the voter which votes taking into consideration whether the pair of data points fall in the same cluster or not in their clustering solution. The function used in weighted binary vote works as follows,

1. For the pair of documents (D_i,D_j) and the clustering unit C, we can define a function $G_C(D_i,D_j)=1$, if D_i and D_j are in same cluster when clustered using the clustering unit C and $G_C(D_i,D_j)=0$, if D_i and D_j falls in different clusters.
2. Then we compute the score $\Sigma_C\, G_C(D_i,D_j)\cdot W_C$, where W_C is the weight of the clustering unit C.
3. The score computed in step 2 is used to determine the similarity between the documents while combining the clusters using single-pass incremental clustering with the threshold tuned in the similar manner as done in the clustering units.

At the end of this phase we get the set of clusters $(S_1,...,S_e)$ where each cluster corresponds to the sub-events in the particular event.

5. EXPERIMENTS AND RESULTS

In this study the data posted for Red River Floods 2009 and Mississippi River Floods 2011 was collected using the YouTube API. The data consists of sub-events observed at different cities located on the banks of these rivers. Mississippi river floods 2011 data consists of sub-events observed at different cities located on its banks including Memphis, Greenville, Vicksburg, Natchez and Helena collected for the period of May 2, 2011 to May 22, 2011. Similarly Red River Floods 2009 data consists of sub-events observed at places like Winnipeg, Grand Forks, Fargo and Moorhead from March 8, 2009 to April 27, 2009. The features associated with each video include title, description, date and location information. Location information in both textual representation and latitude-longitude representation has been considered.

For experiment MATLAB was used for implementing the single pass incremental clustering algorithm. To train the algorithm the Red River Floods data set was used.

During the training phase thresholds and weights are tuned. For training the algorithm, data is divided into two parts and is supplied in order of date. Half of the data is supplied for setting the thresholds. For setting the thresholds, clustering solutions obtained by each clustering unit is compared with the ground truth using the Normalized Mutual Information (NMI) score. Then the threshold value is varied in the range of [0,1] and the value at which it gives maximum NMI score is used as the threshold value for that clustering unit. Other half of the training data is used for setting the weights for the clustering units. The weights of the clustering units are set proportional to their NMI score such that the sum of the weights of all the clustering units equals to one.

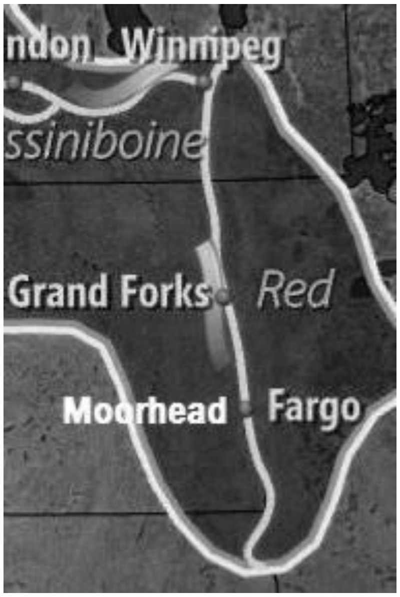

Figure 3 Red River and places on its banks

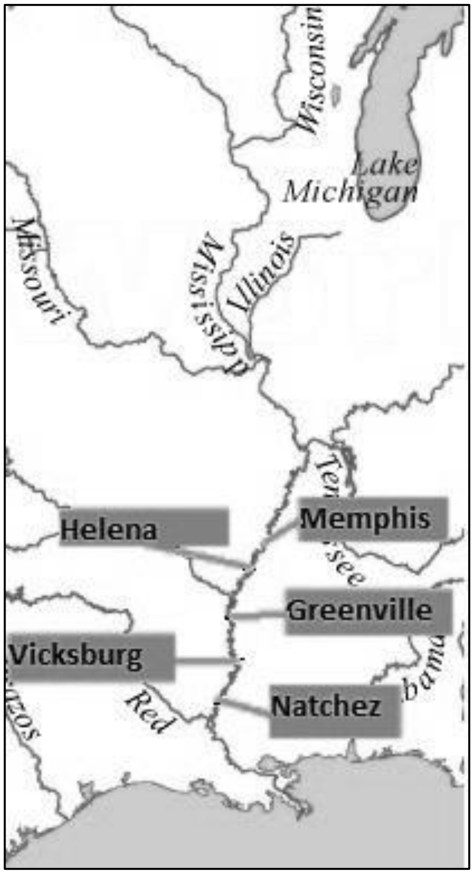

Figure 4 Mississippi River and places on its banks

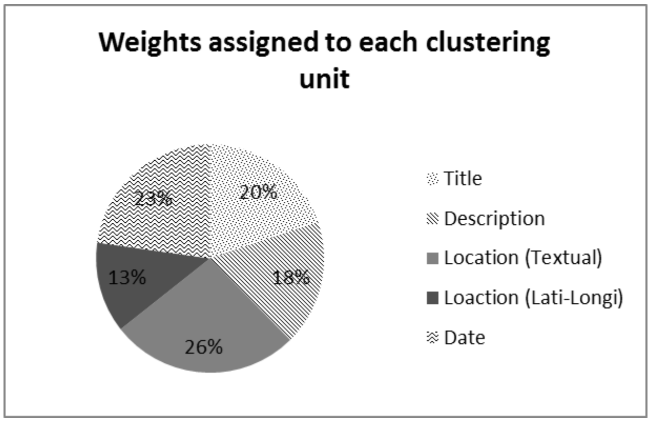

Figure 5 Suitable weights for each clustering unit

While Mississippi dataset was supplied to the algorithm in increasing order of date and the performance was evaluated by computing NMI score of the obtained results and the ground truth.

For evaluation of the approach comparison is made between the NMI scores obtained by the clusters formed by taking all the features individually and the ground truth with the NMI score obtained between the clusters formed by combining the clusters using above discussed technique and the ground truth.

The pie chart in Figure 5 shows the most suitable weights assigned to each clustering unit for given data. However these weights tend to change with change in type of data.

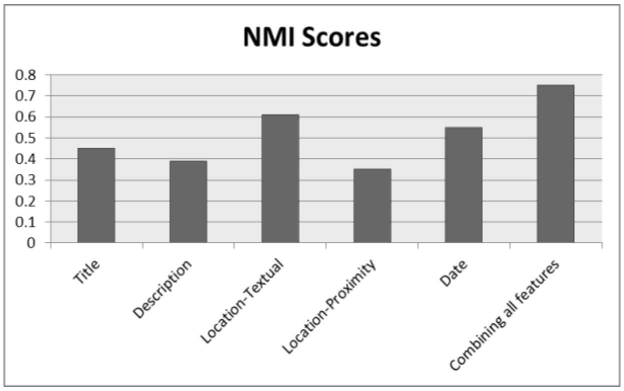

Figure 6 Comparison of NMI Scores taking different features individually and taking multiple features into account

Figure 6 shows the chart comparing the NMI scores of different clustering unit taking individual features. NMI score of 0.603 is obtained by taking textual location information and is highest amongst the performance obtained by taking each feature individually. While the last bar shows the NMI score of 0.7087 obtained by taking multiple features into account using the approach discussed in this paper. So it can be said that clustering the documents using multiple features of the social media document gives higher performance than taking individual features for detecting the sub-events in some natural hazard event.

Table 1 shows the labels of the top 6 clusters obtained by the sub-event detection module on using Mississippi River Floods 2011 data. Location and the date of the clusters are determined by the location and date of the centroid. Also it shows the top 4 words occurring in the *tf-idf* vector of textual features of the documents falling in the respective cluster.

Table 1 Sub-event location, date and top 4 terms occurring in the *tf-idf* vectors of documents

Location	**Date**	**Top 4 terms occurring in *tf-idf* vectors of textual features**
Memphis	5/10/2011	memphis, river, mississippi, part
Vicksburg	5/18/2011	vicksburg, ms, mississippi, river
Greenville	5/16/2011	greenville, mississippi, river, ms
Memphis	5/10/2011	memphis, mississippi, mud, island
Natchez	5/19/2011	natchez, hill, lousiana, flow
Helena	5/12/2011	helena, bridge, cross, look

For comparison of the proposed approach in this work with the technique proposed in [8] of using self-organizing maps, same data of Mississippi River Floods 2011 was used and NMI scores of obtained clusters with ground truth was compared in both cases. The NMI scores obtained by the technique proposed in [8] was 0.5328 while the NMI score obtained by the approach proposed in this work was 0.7087.

6. CONCLUSION

Research shows that social media data contains useful information about an event. And by facilitating the searching of sub-events of a particular event it can be helpful to the relief and rescue activities during natural hazards. Manually detecting sub-events is a difficult task. Traditional topic detection techniques for news articles cannot be used for sub-event detection with social network data. Other existing sub-event detection techniques for social network data are computationally expensive. Also from the results it can be concluded that we cannot rely only on any one particular feature for social network data. Here the approach discussed helps in sub-event detection considering multiple features of social media data into account.

YouTube data was used for experiments but the approach discussed in this paper is not constrained only to the YouTube data. It also be used with other social media data. However based on the type of features of the data, the similarity measures and weights of the feature based clustering units tend to change.

The limitation of this approach is that the sub-event detection module trained for a particular type of natural hazard cannot perform so well with other natural hazards as different natural hazards has different levels of granularity. That is some hazards have sub-events that are closely located in time and space while for other sub-events may be located at greater distance in time and space.

Work has to be done in determining some technique which can facilitate sub-event detection at different levels of granularity.

7. REFERENCES

[1] J. Allan. *Topic Detection and Tracking - Event-based Information Organization*. Kluwer Academic Publisher, 2002.

[2] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. *Twelfth International Workshop on the Web and Databases*, pp. 107-111, Providence, Rhode Island, USA, June 2009.

[3] Finding distances based on Latitude and Longitude using Haversine Formula.

http://andrew.hedges.name/experiments/haversine/

[4] C. Hagar, and C. Haythornthwaite. Crisis, Farming & Community. *The Journal of Community Informatics*, Volume 1, Issue 3, 41-52, 2005.

[5] P. T. Jaeger, L. A. Langa, C. R. McClure, and J. C. Bertot. The 2004 and 2005 Gulf Coast Hurricanes: Evolving Roles and Lessons Learned for Public Libraries in Disaster Preparedness and Community Services. *Public Library Quarterly*, 25, 3/4, 199-214, 2007.

[6] P. T. Jaeger, K. R. Fleischmann, J. Preece, B. Shneiderman, and, P. F. Wu. Community response grids: using information technology to help communities respond to bioterror emergencies. *Boisecurity and Bioterrorism: Biodefense Stratergy*, Volume 5 (4), 335-346, December 2007.

[7] G. Kumaram, and J.Allan. Text classification and named entities for new event detection. *Proceedings of the 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 297-304, 2004.

[8] D. Pohl, A. Bouchachia, and H. Hellwagner. Automatic Sub-Event Detection in Emergency Management Using Social Media. *Proceedings of the 21st international conference companion on World Wide Web*, 683-686, Lyon, France, April 2012.

[9] Y. Qu, C. Huang, P. Zhang, and J. Zhang. Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. *Proceedings of the ACM 2011 conference on Computer Supported Cooperative Work (CSCW)*, 25-34, Hangzhou, China, March 2011.

[10] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*, 103–110, New York, USA, 2007.

[11] Strehl, J. Ghosh, and C. Cardie. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 583-617, vol. 3, 2002.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. *Proceedings of the 19th World Wide Web Conference*, 851-860, Raleigh, NC, USA, April 2010.

[13] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg. Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work (CSCW)*, 241-250, Savannah, Georgia, USA, February 2010.

[14] C. Torrey, M. Burke, M. Lee, A. Dey, S. Fussell, and S. Kiesler. Connected Giving: Ordinary People Coordinating Disaster Relief on the Internet. *Proceedings of Hawaii International Conference on Systems Science 2007*, 217, Big Island, USA, January 2007.

[15] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of the 28th international conference on Human factors in Computing Systems (CHI)*, 1079-1088, Atlanta, Georgia, USA, April 2010.

[16] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K.M. Anderson. Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets During Mass Emergency. *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 49-57, Barcelona, Spain, July 2011.

[17] Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. 2012. Using Social Media to Enhance Emergency Situation Awareness. In *IEEE Intelligent Systems*, vol. 99, PrePrints, February 2012.