Identifying and Characterizing User Communities on Twitter during Crisis Events

Aditi Gupta*, Anupam Joshi†, Ponnurangam Kumaraguru*
*Indraprastha Institute of Information Technology, Delhi, India
†University of Maryland Baltimore County, Maryland, USA
{aditig, pk}@iiitd.ac.in, joshi@cs.umbc.edu

ABSTRACT

Twitter is a prominent online social media which is used to share information and opinions. Previous research has shown that current real world news topics and events dominate the discussions on Twitter. In this paper, we present a preliminary study to identify and characterize communities from a set of users who post messages on Twitter during crisis events. We present our work in progress by analyzing three major crisis events of 2011 as case studies (Hurricane Irene, Riots in England, and Earthquake in Virginia). Hurricane Irene alone, caused a damage of about 7-10 billion USD and claimed 56 lives. The aim of this paper is to identify the different user communities, and characterize them by the top central users. First, we defined a similarity metric between users based on their links, content posted and meta-data. Second, we applied spectral clustering to obtain communities of users formed during three different crisis events. Third, we evaluated the mechanism to identify top central users using degree centrality; we showed that the top users represent the topics and opinions of all the users in the community with 81% accuracy on an average. The top central people identified represent what the entire community shares. Therefore to understand a community, we need to monitor and analyze only these top users rather than all the users in a community.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information storage and retrieval—*Information Retrieval*; K.4.1 [**Computing Milieux**]: Computers and society—*Public policy issues*

General Terms

Experimentation, Measurement

Keywords

Community detection, Online social media, Crisis events

DUBMMSM'12, October 29, 2012, Maui, Hawaii, USA.
Copyright 2012 ACM 978-1-4503-1707-8/12/10 ...$15.00.

1. INTRODUCTION

Online social media provides people with a platform to disseminate ideas, learn information, explore knowledge, and express their opinions on diverse topics. There is a sudden rise in activity over the Internet and online social media, especially during crisis and emergency situations [5]. People log on to social media websites to check for updates about the event and also to share information about the event with others. In such situations, social media content provide a vast resource of unmonitored and unstructured but rich information about events. Since the data is generated in real time and by users, many of whom are some times directly or indirectly in the actual event; mining this content can yield quite useful knowledge about the crisis situation [4, 8].

Our work aims to identify, extract and characterize prominent components of topics and opinions that people share during crisis events on Twitter. Twitter, is a micro-blogging website on which users post messages (called tweets) of a maximum length of 140 characters.[1] Researchers have analyzed relevance of online social media, and in particular Twitter as news disseminating agent. Kwak et al. showed that 85% topics discussed on Twitter are related to news [3]. Users who post messages about events, have varied opinions, sentiments, and information content. The biggest challenge in analyzing data from any social media is the volume of content being generated. Millions of users post messages on Twitter everyday, and these numbers further surge up during a real-world event (in particular, high impact events). The motivation for our work is that, high volume of content and users on such social networks, makes manual monitoring of each message or user impossible. We propose a framework to cluster together users who are similar and then extract views of prominent people in each of these communities formed by identifying the top central users in each of them.[2] Community detection helps us in monitoring the broad subtopics and sub-communities of users that are formed during an event. Identifying top central users is useful as they can represent the entire community. Hence we need to monitor and analyze the opinions and sentiments of only these few user profiles rather than all the users in a community. To identify the communities of users, and top users in each of these community, we utilize techniques of clustering and social network analysis.

As case studies for the generic framework proposed in this paper, we analyze three major crisis events of 2011, the hur-

[1] www.twitter.com

[2] The words cluster, community and groups are used interchangeably in this paper.

ricane Irene, the riots in United Kingdom and the earthquake in Virginia. Hurricane Irene alone, caused a damage of about 7-10 billion+ USD and claimed 56 lives.[3] All these three real-life events, saw a vast number of message posts on online social media during the events itself, including Twitter. For instance, we collected over 1.1 million tweets related to England riots alone. The methodology and framework evaluated in the paper is generic and may be applied to any event. The aim is to identify different user communities, and characterize them by the properties of top central users in them. Communities in a network are groups of nodes such that, there are maximum number of edges inside the clusters than between clusters [7]. The main contributions of this paper are:

- We defined a novel similarity metric to compute similarities between users based on their links, content posted and meta data.
- We applied spectral clustering to obtain communities of users formed during three different crisis events.
- We showed most central people (by degree centrality) represent what users in the cluster are talking about with 81% accuracy on average.

In section 2, we present the theory and methodology followed in the paper. Section 3 gives the data description considered for the analysis. Section 4 describes the evaluation results. The last section consists of discussion and future work.

2. THEORY AND METHODOLOGY

Most of the previous work has focussed on detecting and analyzing communities based on explicit social links like friends, followers or retweets [9, 10]. We follow the approach proposed by Java et al. for clustering blog data by using spectral methods [2]. Java et al. adapted the N-Cut spectral clustering algorithm for image segmentation to online social media [7]. Adapting the approach proposed by them for blogs to Twitter, we formulate the communities based on different metrics like content, link and meta-data of user on Twitter. Figure 1 represents the architecture of the methodology followed in this paper. After extracting the communities of users, we also propose the use of centrality metrics of social network analysis in order to obtain top central users in each community. The identification of top users in each community helps us in characterizing and analyzing, the topics and opinions of each community, and who comprises of that community.

We perform clustering of users based upon the tweets posted, the link structure and the meta-data of the users on Twitter. A *user*user* similarity matrix U is constructed, where for all users u_i and u_j, the value $a_{i,j}$ is the sum of the following:

- **Content similarity** $C[i,j]$: We calculate the similarity of content between two users by computing number of common words, hashtags and URLs in all the tweets by u_i and u_j related to the particular event.[4]
- **Link similarity** $L[i,j]$: The similarity between two users, u_i and u_j with respect to links is computed based on how many times two users retweet, mention or reply to each other tweets or a common third person's tweet.
- **Meta-data similarity** $M[i,j]$: Twitter profile of a user contains an optional field called location. We compute similarity between two users based on the value in their location field. The field is a text box, the user may leave it blank, or fill it with a valid / invalid location. We used Yahoo PlaceFinder API[5] to check whether the given text corresponds to a valid location. From the values returned by the API, we check if the users have similarity at the level of country, state or city.

For each of the similarity metrics above, we normalize the score of each similarity using the maximum value of similarity score for each of the feature and then compute their sum (Equation 1).

$$U[i,j] = C[i,j] + L[i,j] + M[i,j] \qquad (1)$$

After collecting data from Twitter API, and constructing $U[i,j]$, the next step is to perform spectral clustering on this matrix. We performed the spectral clustering using the *specc* function of *R* statistical analysis package, by specifying the number of clusters as three. The number of clusters was selected as three intuitively, for our preliminary study. In future, we would like to apply standard techniques for selection of number of clusters. The users are divided in different communities, which is followed by identifying top users in each of the community. We use the degree centrality metric of social network analysis to identify the top users for the communities. We then evaluate, that with how much accuracy, the top central users represent the opinions and sentiments of the entire community.

3. DATA DESCRIPTION

Three events analyzed in this paper are England riots, hurricane Irene and earthquake in Virginia. All events occurred during the months of July and August, 2011. Riots in United Kingdom caused 5 deaths, in addition to 16 civilian and 186 police injuries. We selected tweets about the UK riots, based on the keywords related to them that emerged as trending topics on Twitter from 6th August to 11th August [1]. Similarly, we collected data for the other two events. Our second event under consideration is the earthquake of magnitude 5.8 that hit the Piedmont region of the U.S. state of Virginia. The third event is the hurricane Irene, which caused 55 deaths and a damage of US $10.1 billion. We used the *Streaming API* from Twitter to collect the tweets and user information corresponding to these three events. Table 1 presents the tweets and users of the three events analyzed in this paper.

Table 1: Data statistics for the three crisis events.

Event	Tweets	Users
England Riots	1,165,628	546,966
Hurricane Irene	90,237	55,718
Earthquake in Virginia	277,604	219,621

[3] http://www.nytimes.com/2011/08/31/us/31floods.html?pagewanted=all

[4] Before finding the common words among the tweets, we remove all the stop words from the tweet text.

[5] http://developer.yahoo.com/geo/placefinder/

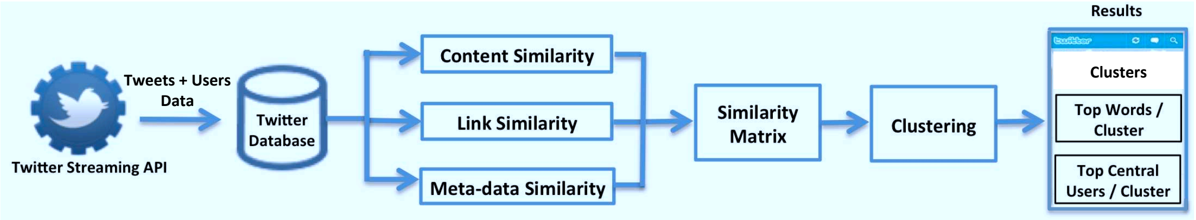

Figure 1: The architecture diagram for the methodology followed.

4. RESULTS AND EVALUATION

We performed two kinds of evaluation for the analysis. First, we measure the quality of communities formed using the modularity metric. Second, we evaluated if the top users identified using degree centrality metric represent the opinions and topic of the entire community.

4.1 Clustering Evaluation

For evaluating the quality of clusters obtained by spectral clustering, we computed the modularity score for each of the three events. Modularity score was proposed by Newman et al. as an index to measure how good a division of the network is in communities [6]. This score represents the fraction of edges that lie inside the communities minus the factor of expected value if the edges were placed at random, while maintaining the degree of nodes. Newman et al. also showed that value of modularity above 0.3 implies a good division of graph into communities.[6] We get the modularity score greater than 0.5 for all events (Hurricane Irene = 0.67, England Riots = 0.51 and Virginia Earthquake = 0.53), hence we can conclude that the division of graph into communities is of good quality.

4.2 Top Users Evaluation

Next, we evaluate if the top central people in a community based on the similarity metrics represent the opinion and topics of the entire community. The motivation behind proving this result is that, if we can show that the top users based on similarity represent what the entire community shares, then to analyze and characterize the network, we need to monitor and investigate only these few user profiles rather than all the users in a community. We used Mean Average Precision (MAP) evaluation metric to test the same. MAP for a set of queries is the mean of the average precision scores for each query. We extracted the top N words, according to the tf-idf [7] score, for all users in each of the community and then only by the top central users in each of the community. Hence, for computing MAP scores, the top N words by all users in the community form the ground truth and we compute the MAP for each community for the top N words by the central users. Figure 2 shows the results obtained, by MAP, we obtained 81% accuracy (on average; max: 96%) for all three events. We can conclude that the top central people represent what the entire community is saying with a high accuracy of 81%, which implies monitoring and analyzing these users can provide us with the opinions and topic of the entire community.

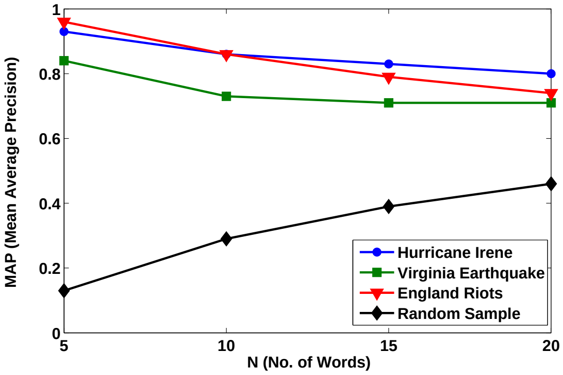

Figure 2: The graph shows the Mean Average Precision score for the top central users' detection for the three crisis events. We observe that while for the three events, MAP values are high; the corresponding values for a random sample are much lower. The graph is truncated to show the main effect. The values plateau after the points shown in the graph.

Next, we go a step further to check if the above result holds true only in case of tweets corresponding to events, or, in general with any set of tweets. To do this, we selected a random sample of messages from Twitter corresponding to the same time frame as the three events and having number of tweets equal to the mean average of the number of tweets in the three events. Figure 2 shows that the performance of MAP drastically reduces for the random sample of tweets, we get a MAP score of 32% on average. The conclusion that top users based on degree centrality represent the opinions and topic of the community holds true in case of events, since the tweets related to an event contain higher similarity in various properties like content of tweets and location of users.

4.3 Community Characterization

We now characterize the communities obtained for Hurricane Irene in depth to understand how the approach evaluated in this paper can be useful in real-life. Table 2 shows the top twenty words in the three clusters corresponding to the communities for all three events. We observed that while community 2 and 3 contain words related to information about the hurricane in different forms, community 1 consists of tweets by spammers on Twitter. It is a common practice by spammers to use trending hashtags to promote obscene

[6] The maximum value of modularity for a network is 1 and minimum is 0, a high value of modularity score indicates, good quality division of the network, i.e. more number of edges in the network are within the clusters than between clusters.

[7] Tf-idf stands for term frequency-inverse document frequency. It is a statistical measure which used to evaluate how important a word is to a document in a collection of documents.

Table 2: Top 20 words from each community (cluster) for the three crisis events.

Events	Community 1	Community 2	Community 3
Hurricane Irene	hurricane, irene, http, jada, song, brenda, butistillloveu, photos, icanhonestlysay, check, neverapologizefor, video, scandal, angelina, jolie, storm, nick, college, bahamas, puerto	irene, hurricane, http, coast, category, bahamas, east, puerto, storm, rico, winds, advisory, issued, forecast, heads, track, path, update, moving, number	hurricane, irene, http, puerto, coast, storm, rico, news, track, winds, bahamas, category, east, forecast, advisory, florida, update, issued, tropical, latest
England Riots	londonriots, http, ukriots, police, london, riots, people, hackney, rioters, news, fire, dont, riot, cameron, riotcleanup, birmingham, looting, croydon, night, stop	londonriots, http, ukriots, police, london, riots, people, rioters, cameron, hackney, news, fire, dont, riot, looting, prayforlondon, birmingham, manchesterriots, stop, riotcleanup	londonriots, http, ukriots, police, riots, london, people, rioters, hackney, cameron, riot, birmingham, news, fire, riotcleanup, dont, looting, make, manchesterriots, stop, rioting
Virginia Earthquake	earthquake, http, today, quake, east, coast, earth, richmond, washington, august, virginia, news, campbell, glen, monument, heard, norwegian, alaska, new york, video	earthquake, http, east, coast, virginia, washington, felt, magnitude, news, monument, feel, damage, today, nuclear, quake, epicenter, people, school, video	earthquake, http, coast, east, washington, felt, virginia, news, damage, today, magnitude, monument, feel, quake, people, video, area, breaking, shit, nuclear

or spam content, hence community 1 contains the words that were trending with respect to the hurricane, and rest of the words refer to spam content (marked in red) like *photos and videos of Angelina Jolie*. Using the technique of spectral clustering followed by top users detection using degree centrality measures, we were able to cluster together information sharing users and separate them from the spammers who exploit the trending topic terms to spread their content. By looking at the top words in each community, community 2 and 3 look very similar, hence we look at the corresponding top central users in cluster 2 and 3. By analyzing the screen name, user description and tweets by the top central users, we found that: community 2 is formed mostly of official news media, emergency and alert profiles (for e.g. *ShareWith911, metofficestorm, RES911CU, Neednewsnow*), while community 3 is formed with common or general users of Twitter (for e.g. *KerrieRamagano3, YaekoAvilez3837, AureliaHickman1*), none of whom had even provided any description on their Twitter profiles. Thus, even though community 2 and 3 had almost similar top words, our characterization shows that tweets in community 2 originate from more authentic and official profiles than community 3 and can be considered more trustworthy. Similar to hurricane Irene analysis, when we analyzed communities for Virginia earthquake in Table 2, we observed while the first cluster contains more words related to the location of the earthquake (*richmond, virginia, new york, alaska, norwegian*), community 2 and 3 emphasize on the damage and people affected (*damage, people, feel, school, nuclear*). We skip the other details, due to limited scope of this paper.

5. DISCUSSION AND FUTURE WORK

In this paper, we identified and evaluated communities of users on Twitter during three different crisis events. We applied spectral clustering technique to cluster users based on a similarity metric, taking into account the content, link and meta-data similarities of the users. We then used the degree centrality measure from social network analysis to extract top central users from each of the communities formed. We showed that the top central users represent the topics and opinions of the entire community with an average 81% accuracy. The results obtained with the case study of three events show the potential of applying spectral clustering and centrality measures, to identify community of users and the prominent top users in each community, during crisis events. We aim to identify various other crisis events to perform similar analysis. Also, we would like to strengthen our similarity metric construction by including more features and adding weights to all the features.

6. REFERENCES

[1] Gupta et al. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12.

[2] Java et al. Detecting Commmunities via Simultaneous Clustering of Graphs and Folksonomies. In *Proceedings of the Tenth Workshop on Web Mining and Web Usage Analysis (WebKDD)*. ACM, August 2008.

[3] Kwak et al. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10. ACM.

[4] Longueville et al. "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, LBSN, 2009.

[5] Mendoza et al. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10.

[6] Newman et al. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(5 Pt 2), Nov. 2004.

[7] Shi et al. Normalized cuts and image segmentation. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*.

[8] Vieweg et al. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI '10*.

[9] Wang et al. Detecting community kernels in large social networks. In *Proceedings of International Conference on Data Mining*, ICDM '11.

[10] Welch et al. Topical semantics of twitter links. In *Proceedings of the fourth international conference on Web search and data mining*, WSDM '11.