Text vs. Images: On the Viability of Social Media to Assess Earthquake Damage

Yuan Liang, James Caverlee
Department of Computer
Science and Engineering
Texas A&M University
College Station, TX, USA - 77843
{yliang, caverlee}@cse.tamu.edu

John Mander
Department of Civil Engineering
Texas A&M University
College Station, TX, USA - 77843
jmander@civil.tamu.edu

ABSTRACT

In this paper, we investigate the potential of social media to provide rapid insights into the location and extent of damage associated with two recent earthquakes – the 2011 Tohoku earthquake in Japan and the 2011 Christchurch earthquake in New Zealand. Concretely, we (i) assess and model the spatial coverage of social media; and (ii) study the density and dynamics of social media in the aftermath of these two earthquakes. We examine the difference between text tweets and media tweets (containing links to images and videos), and investigate tweet density, re-tweet density, and user tweeting count to estimate the epicenter and to model the intensity attenuation of each earthquake. We find that media tweets provide more valuable location information, and that the relationship between social media activity vs. loss/damage attenuation suggests that social media following a catastrophic event can provide rapid insight into the extent of damage.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; J.m [**Computer Applications**]: Miscellaneous

Keywords

Social media, damage assessment, attenuation pattern

1. INTRODUCTION

Social media, when coupled with extremely granular spatio-temporal information (e.g., timestamps and GPS-style geocodes), offers the tantalizing promise of a minute-by-minute and region-by-region account of a disaster as it unfolds. Indeed, recent work has illustrated this promise through automated methods to aggregate Twitter posts for detecting the epicenter and trajectory of an earthquake [8], for detecting earthquakes and building a predictive system to notify people at-risk [7], and for constructing "theme cycles" from geo-located blog posts for assessing the public's response to Hurricane Katrina [6].

These efforts to leverage geo-located social media are encouraging and inspire our investigation into the potential of social media to provide rapid insights into the location and extent of damage associated with earthquakes. Our overall research goal is to investigate the capacity of social media for conveying damage information, which is an important step for providing responders with rapid insight into the extent of damage to be expected in the field and the locations of greatest damage, which are both necessary for deciding how to best deploy the limited emergency response and recovery resources during the initial moments of an earthquake.

In this initial study, we assess the quality, coverage, and capacity of two types of social media: text-only tweets, which are typically short and require little effort to post, and media-containing tweets, which include links to either images or videos and are intuitively more expensive in the sense that the person posting must expend effort to capture the picture or video. We report on our initial investigation through an examination of the 2011 Tohoku earthquake in Japan and the 2011 Christchurch earthquake in New Zealand. Our study suggests that media tweets provide more valuable location information than text tweets, and that both provide comparable evidence of the linear intensity attenuation function for earthquakes, indicating a similar ability to serve as the foundation of rapid damage assessment for earthquakes.

2. RELATED WORK

Recently, with the thriving development of social network services, scientists have begun to study the use of social media on large-scale crises, and apply it to detect, track, summarize and assess them. For example, [4] and [9] examined the social life of micro-blogged information and show how social media can be used for summarizing hazards. By studying 106 million tweets generated, [2] found the majority (over 85%) of detected topics are headline news or persistent news, indicating Twitter plays a more important role as an information source.

Location sharing services have also attracted increasing attention in the last couple of years, and recently have been studied for emergency events. [4] analyzed the temporal, spatial and social dynamics of tweets during a fire emergency, and discussed how the location-based social network can be a source to collect information during emergencies. [8] treated every user as a sensor, and applied Kalman filter to the signals generated by these human-powered sensors to locate an earthquakes' epicenter and to predict the trajectory of the resulting typhoons. [1] explored the relationship between the spatial pattern of geolocated SMS (Short Message Service) messages and the building damage.

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

Images are increasingly playing a more significant role in disaster detection and summarization. [3] described the evolution of Flickr's role during disaster response and recovery efforts, and discussed this evolutionary growth pattern as a community forum for disaster-related activities. [10] extracted images semantic information under translation model, and use a time-line to summarize the 2011 Tohoku earthquake.

3. DATA COLLECTION

For our study, we collected two Twitter-based datasets associated with the March 2011 Tohoku earthquake in Japan and the February 2011 Christchurch earthquake in New Zealand. For each, we identified earthquake-related tweets from an ongoing crawl hosted in our lab that collects around 3 millions geo-located tweets per day. To identify tweets related to each event, we first selected several keywords with the largest counts co-occuring with the seed words "earthquake", "地震", and then filtered some of them according to tf-idf. For Japan, we identified 75 earthquake-related keywords (17 in English, 58 in Japanese); for New Zealand, we identified 15 earthquake-related keywords. Based on these keywords, we selected all tweets containing at least one of these keywords within the earthquake time-window. The details for the two collected data sets are listed in Table 1.

Table 1: Data Sets

Event	Time Frame	Selected Terms	#Tweet
JPEQ	03/11/2011 -03/15/2011	earthquake, epicenter, eqjp, honshu, fukushima	207,876
NZEQ	02/20/2011 -02/30/2011	earthquake, christchurch, new zealand, victim, rescue	38,699

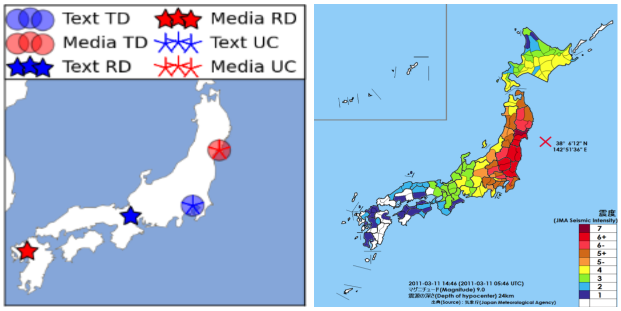
The Tohuku earthquake dataset (JPEQ) contains 207,876 tweets, of which 35.41% contain a URL (which links to an image, video, or webpage). The Christchurch earthquake dataset (NZEQ) contains 38,699 tweets, of which 32.55% contain a URL. We consider each URL-containing tweet as a media tweet, whereas we consider all other tweets as text tweets. On inspection, a random sample of media tweets were found to overwhelmingly include on-site pictures of earthquake damage.

4. APPROACH AND FINDINGS

To begin our examination of these two kinds of geo-located social media, we construct a series of investigations intended to assess the quality, coverage, and capacity of social media in the aftermath of the two earthquakes.

- Epicenter estimation: Firstly, we assess the quality of the two kinds of tweets for estimating the actual epicenter of each earthquake. We consider as input to this task three different features, including the tweet density, the re-tweet density, and the average tweets count per user.

- Intensity attenuation: Based on the detected epicenter, we further model the three features versus the radius from the epicenter to study the intensity attenuation pattern for earthquakes. It is well known that for any given specific earthquake there exists an "attenuation relationship" that relates the shaking intensity with respect to the distance to the earthquake's epicenter [5]. Do we witness such a pattern in text and media tweets? And which factors model intensity attenuation the best?



(a) Estimated Epicenter (b) Actual Epicenter

Figure 1: Location Estimation Using Different Features

- Spread speed: We integrate the temporal and spatial features of tweets to examine the speed of propagation of text and media tweets. This spread speed is important for understanding the influence of social media for disaster communication.

4.1 Epicenter Estimation

In our first study, we investigate the capacity of social media to estimate the epicenter of each earthquake. We consider three different tweet-based features for epicenter estimation and compare across both text and media tweets. For this case, we bucket all tweets by applying a grid overlaid on the bounding box of Japan and New Zealand. We use a grid width of 0.01 degrees, which corresponds to about 1.11 km. For each grid cell, we compute the following three features based on the geo-located tweets:

- Tweet density (TD). The first feature is the number of tweets for each grid cell divided by the area of the cell. In this way, we identify the density of tweets for each cell. Intuitively, this feature captures the assumption that people are more likely to post a text or media tweet in the regions that are more severely damaged.

- Re-tweet density (RD). The second feature is the number of re-tweets for each grid cell divided by the area of the cell. Perhaps severely damaged areas re-tweet more. Or perhaps areas outside of the most damaged regions re-tweet based on first-hand accounts from closer to the epicenter.

- User tweeting count (UC). The third feature measures the average number of tweets per user (number of tweets divided by the number of users) from a particular grid cell. Our intuition here is that users who are in a damaged region may tend to be engaged with the event longer than those who are not so close, so they might emit more tweets than those who are outside the damaged region.

For each feature and for each type of tweet (text versus media), we identify the grid cell with the maximum value as the detected epicenter. In Figure 1, the estimated epicenters using the text tweets density and text re-tweets density is located around Tokyo, which has the largest population density in Japan. This fact indicates that the count of text tweets can be easily affected by the population, so they are not good evidence for epicenter estimation. The re-tweet density feature locates the epicenter to the areas which are not badly damaged, which suggests that people who are in the safe places tend to re-tweet posts of others but not generate original content.

In contrast, media tweet density and users' media tweeting count perform the best; the epicenter detected by these features are located in most severe region in the earthquake, and are closest to the actual epicenter, which is labeled by a red cross in Figure 1b. These results suggest that media tweets perform better on epicenter estimation. We attribute this to the fact that media tweets are more likely to happen in the local place of a region of crisis because the scene in the images should be actually observed, while text tweet can happen anywhere no matter where the scene it describes really happens. Therefore, the location information for media tweets is more credible than that of text tweets.

Table 2: The Euclidean Distance Between Estimated Epicenter and Actual Epicenter (Degree)

Event	TD		RD		UC	
	text	media	text	media	text	media
JPEQ	3.747	**1.080**	6.950	12.927	3.747	**1.080**
NZEQ	7.066	**0.100**	3.095	3.111	7.066	**0.100**

In Table 2, we measure the distance between the estimated epicenter and the actual epicenter (from Wikipedia) with Euclidean distance. We find that for both JPEQ and NZEQ, the media tweets perform better than text tweets. And the tweet density and user tweeting count achieve the same good results. For JPEQ, since the epicenter is located in the ocean, the detected epicenter is located in the most severely affected region of Japan (shown in Figure 1b), so the smallest distance is about 1 degree. For NZEQ, the detected center is close to the actual center, with 0.1 degree difference.

Together, these results show that tweet density and user tweeting count are the best features for identifying the epicenter of an earthquake, that re-tweets tend to happen in the regions that are less affected, and that location information of media tweet is more credible than that of text tweets.

4.2 Intensity Attenuation

Based on the detected epicenter, we next examine the relationship between text and media tweets on intensity attenuation. It is well known that for any given specific earthquake there exists an "attenuation relationship" that relates the shaking intensity with respect to the distance to the earthquake's epicenter. For this study, we reconsider the same three feature as before – tweet density, re-tweet density and users' tweeting count – as well as the two different types of tweets (text and media).

Rather than consider a simple grid, we consider increasing concentric circles around the epicenter. Given an epicenter o for a certain region, we extract the features for the circle region centered at o with radius r. Then we increase r by 0.01 degree (about 1.11 km), and extract the features for the ring area outside the inner circle. At last, we observe the values of features against the radius r. Given the detected epicenter, which has the largest tweet density, the values for the three features versus the radius from the epicenter are shown in Figure 2.

From Figure 2a, we can see that in the areas close to the epicenter, the log density of tweets is linearly related with the radius r, which means the density decreases following a power law. This linear relationship is consistent with the previous seismic intensity research [5] on the power law decay in intensity of earthquakes. This suggest that tweet density could be used as a proxy for actual seismic readings toward constructing rapid damage assessments based purely on social media content. We can also see that text and media tweets have similar trends, indicating that both are suitable for earthquake intensity estimation.

Figure 2b shows the relationship between re-tweet density and distance from the epicenter. Interestingly the re-tweet densities firstly stay stable, then decrease exponentially when the distance from the epicenter exceeds 10 km. With respect to the results from Figure 2a, we know that in the nearest 10 km region, the re-tweet rate (re-tweet density/tweet density) increases with the distance from the center, because the re-tweet density stays constant and the tweet density decreases. This result is consistent with the previous finding in epicenter estimation that people tend to re-tweet more from the (more distant) less damaged area. These more distant re-tweeters are serving as a communications hub spreading the posts from the more direct observers.

Figure 2c shows the results that the tweeting count per user versus the distance from the center. Surprisingly, the tweeting count per user is not affected by the distances. For most regions they stay constant, but burst in certain regions. User tweeting counts appear to largely depend on particular population and media centers (e.g., where newspapers and government agencies are located), and so it is unrelated to the radius from the epicenter.

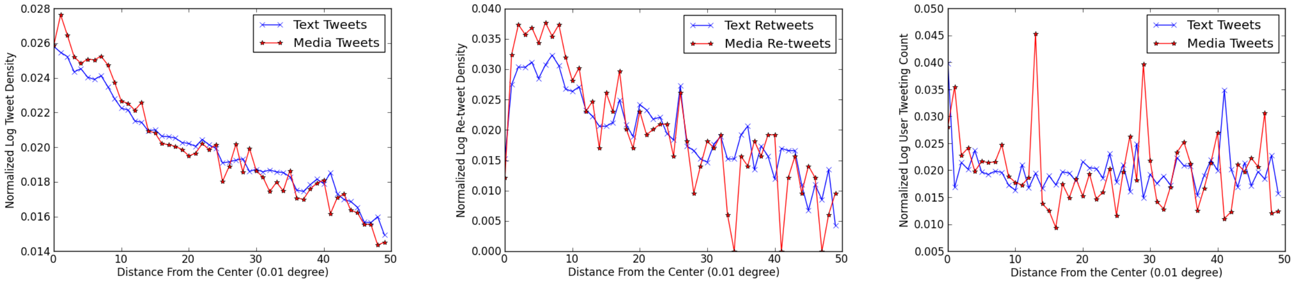
4.3 Spread Speed

Finally, we consider the dynamics of tweets: how fast does social media spread in the aftermath of an earthquake? We integrate the temporal and spatial features of tweets to examine the speed of propagation of text and media tweets. For each minute from the onset of each earthquake, we compute the average distance of posted tweets from the epicenter, allowing us to compute the average spread distance versus time. The spread distance in the first 90 minutes is shown in Figure 3.

For text tweets, we see that the spread distance and time are linearly related. As time passes, tweets spread more rapidly in terms of distance. In contrast, we see that media tweets are less frequent and have a more chaotic spread. Applying linear regression to the two, we get the correlation coefficients $r = 0.626$, slope $l = 0.693$ for media tweet, and $r = 0.910$, $l = 1.020$ for text tweet. The correlation coefficient shows that spread distance of text tweets is linearly related to the passing time, indicating a constant spread speed for text tweets. And the spread speed (represented by slope) of text tweets is about 1.020 degrees per minute (about 113.34 km per minute), which is much faster than media tweets, which have a speed of 0.639 degree per minute (71.01 km per minute). Consistent with the results in Figure 3, the correlation coefficient of media tweets suggest they are more chaotic than text tweets.

5. CONCLUSION

Based on this investigation into geo-located text and media tweets in the 2011 Tohoku earthquake and the 2011 Christchurch earthquake, we have seen encouraging evidence. First, we find media tweets provide more valuable location

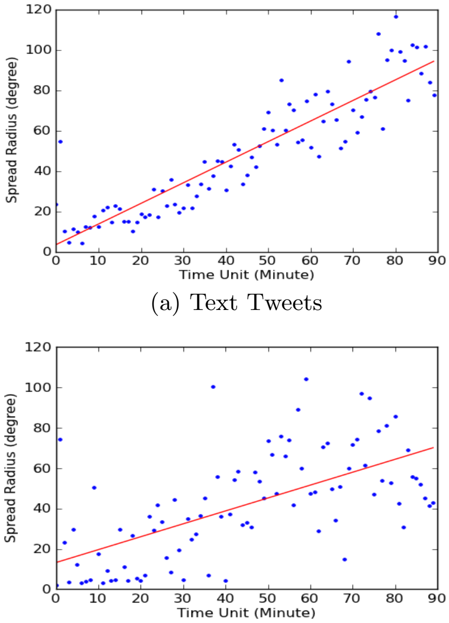


(a) Log Tweet Density

(b) Log Re-tweet Density

(c) Log Tweeting Count Per User

Figure 2: The Densities Versus the Radius in JPEQ



(a) Text Tweets

(b) Media Tweets

Figure 3: The Spread Speed of Tweets in JPEQ

information than text tweets, and thus play a more important role in epicenter detection. Second, they both provide comparable evidence of the linear intensity attenuation function for earthquakes, indicating a similar ability to serve as the foundation of rapid damage assessment for earthquakes. The findings of a relationship between social media activity vs. density attenuation suggests that social media following a catastrophic event can provide a rapid insight into the extent of damage to be expected in the field, and that this relationship can then be used to infer the locations of severest damage, as well as where to best deploy emergency response and recovery resources.

In our continuing work, we are investigating this connection through partnerships with structural engineers. Do we indeed find that the estimated social media attenuation function does in fact link to observed structural damage after an earthquake? Can we refine existing models of damage in [5] by incorporating evidence from social media? These and related questions motivate our ongoing investigation into the linkage between social media and traditional methods of post-disaster damage assessment.

6. ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-1138646. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

7. REFERENCES

[1] C. Corbane, G. Lemoine, and M. Kauffmann. Analysis of the relationship between the distribution of building damage and crisis reports in earthquake-affected haiti. *Nat. Hazards Earth Syst. Sci.*, 12:255–265, 2012.

[2] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[3] S. B. Liu, L. Palen, J. Sutton, A. L. Hughes, and S. Vieweg. In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster. In *ISCRAM*, 2008.

[4] B. D. 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. In *Workshop on Location Based Social Networks*, 2009.

[5] J. Mander and Y. Huang. Damage, death and downtime risk attenuation in the 2011 christchurch earthquake. In *Proceedings of the Annual Conference of the New Zealand Society of Earthquake Engineering*, 2012.

[6] Q. Mei, C. Liu, H. Suz, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, 2006.

[7] M. Okazaki and M. Y. Semantic twitter: analyzing tweets for real-time event notification recent trends and developments in social software. *Lecture Notes in Computer Science*, 6045:63–74, 2011.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, 2010.

[9] K. Starbird, L. Palen, A. Hughes, and S. Vieweg. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *CSCW*, 2010.

[10] S. Xu, L. Kong, and Y. Zhang. A picture paints a thousand words: a method of generating image-text timelines. In *CIKM*, 2012.