Mining Conversations of Geographically Changing Users

Liam McNamara
Uppsala University, Sweden
liam.mcnamara@it.uu.se

Christian Rohner
Uppsala University, Sweden
christian.rohner@it.uu.se

ABSTRACT

In recent disaster events, social media has proven to be an effective communication tool for affected people. The corpus of generated messages contains valuable information about the situation, needs, and locations of victims. We propose an approach to extract significant aspects of user discussions to better inform responders and enable an appropriate response.

The methodology combines location based division of users together with standard text mining (term frequency inverse document frequency) to identify important topics of conversation in a dynamic geographic network. We further suggest that both topics and movement patterns change during a disaster, which requires identification of new trends. When applied to an area that has suffered a disaster, this approach can provide 'sensemaking' through insights into where people are located, where they are going and what they communicate when moving.

Categories and Subject Descriptors

D.0 [**Software**]: General

Keywords

geographic, conversations

1. INTRODUCTION

In recent disaster events, social media has proven to be an effective communication tool. Combined with today's smartphone communication capabilities, they have the potential to revolutionise disaster management by harnessing the collective power of people and engaging them in the emergency preparation, response and recovery. Data collected from social media that is aggregated and presented to relief organisations in a concise and meaningful form can make disaster response more targeted, efficient and effective.

Information and communication are the cornerstone of any disaster response to facilitate 'sensemaking' and so inform decisions. They enable relief organisations and agencies to collaborate and so avoid gaps and overlaps in the concerted response, coordinate emergency services personnel, request aid from medics and victims to contact missing family/friends. A study by Olafsson on how technology played a role in the 2011 Pakistan flood response and how information management was utilised during the response concludes that techniques have not significantly evolved since 2005 and the response to the Pakistan earthquake [7]. Only connectivity has improved, and even that just slightly. The communication technology of responders varies from phones and short-range AM radios to satellite links and self-contained networked hubs. NetHope and TSF, two main crisis response organisations focusing on temporary telecommunication infrastructure, employ Broadband Global Area Network equipment and Network Relief Kit using WiFi access points bridged to a satellite data link to provide local Internet access. Other disasters have seen widespread social upheaval, but not necessarily total destruction of infrastructure, such as during the Japanese earthquake/tsunami [8]. This limited connectivity, even after serious disasters, offers the potential to use internetworking to spread and collect information.

Making data available and useful for disaster relief is a big challenge. Firstly, the data has to be collected, which may not be straightforward if the area's infrastructure has been damaged or impaired by the disaster (whether that disaster is natural, structural or political). The potential of using delay-tolerant opportunistic communication as an alternative to wired internet connection has been proposed to extend the amount of citizens that can reach social media [5]. Secondly, the collected data must be aggregated and processed so that important information can be extracted from the mass of data. Thirdly, it must be distributed and displayed in a clear manner to rescue teams and victims using appropriate tools such as visualisations (e.g., Google, GIS, OpenStreet, Ushahidi mapping platforms and disaster-compliant smartphone user-interfaces [2]). This calls for datamining processes that can extract relevant information, then present it through live and innovative mapping tools where data navigation is eased for non-technical users.

2. RELATED WORK

A great deal of recent research has focussed on the application of social media to disaster situations. Much of this interest has been driven by the unprecedented, and to some extent unforeseen, large-scale adoption of social media by citizens in these situations. Published case studies of social media usage in disaster events include statistical analyses of Twitter data in the aftermath of the 2011 earthquake and tsunami in Japan, 2010 earthquake in Chile and 2010 floods in Australia. These early studies find that social media is heavily used to spread news. Communities of interest form around relevant topics and reliable users (from local authorities to local media and even normal users). Thus, in the special situation of a disaster, interest-based networks (such as Twitter, where communities form around interests rather than purely social ties) are particularly interesting to victims to meet and organise themselves.

Post-hoc analysis of social media during these humanitarian disasters has been performed, including the Haitian [10] and Japanese earthquakes [8]. These disasters represent two very technologically different populations, yet both benefited from the application of social media. Rather than just post-hoc analysis, the inclusion

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

of social media into the actual disaster response offers rich potentials. Such as when agencies and volunteers monitor the networks to understand how the situation on the ground is developing. It could also use the direct involvement of victims asking for help or describing problems, such as a road being impassable or water contamination, facilitating organisations' efficient response. Such *crowdsourcing* of information from Twitter was described in [4].

The mining of social media updates for information has been performed for many purposes. Often this is performed through keyword or phrase frequency. To improve the accuracy and ease that automated systems process user updates *Tweak the Tweet* [9] was proposed. They describe the creation of structured tags to be used by crowdsourcing citizens when discussing disasters. Using a more formal language to communicate would undoubtably simplify the process of data collection, though relying on victims to strictly adhere to it may be asking too much.

Now that the usefulness of social media in disasters has been recognised, rather than just analysing peoples' usage, explicit development of applications and modifications to improve their functionality in such situations has begun. *Twimight* has been released which is a “disaster ready” Twitter application [5]. It acts like a normal Twitter application, until a disaster, where it then enables opportunistic exchange of tweets between Twitter peers. So if infrastructure is unavailable peers can still not only communicate with each other but they can ferry messages to Internet gateways and offload ferried messages. Such connectivity enhancing capabilities are crucial for enabling reliable social media interaction even during substantial connectivity problems, due to disasters or otherwise.

3. METHODOLOGY

We shall now introduce some fundamental processes that should be considered when attempting to mine geographic social media data sources for comparative analysis. We assume it is possible to access historical data covering the affected area. Whether this is publicly available, requested/purchased from aggregators or even from the social media organisations themselves. This historical information should contain a large volume of user identities, locations and conversations, which can then be compared to the live stream of updates coming from the disaster affected region to understand the discussion and how it differs from historical behaviour. Ideally, such a system would enable people to understand the current status of the population, how it has changed from the historical period, and more importantly, how it is currently evolving after the disaster. Historic geographic data will enable a rough understanding of the area's usual population density and activity. Social media information will naturally be biased towards younger more technologically literate members of the population, as they are likely to utilise social media, particularly during a disaster. However this nonprobability sample will still be of value to responders and inform them about the population as a whole. Large collections of geographic social media data will likely contain many updates from each user, from the many different locations that they visit. It may be of some value to take each update as a single data point, to understand how individuals move through the system. The evolution of users moving through different locations can then be used to gain insight into the flow of the population. Also, to avoid users with high update frequencies having undue bias on the dataset it could be preferable to condense each user to a single location, such as their more frequent, or most recent location.

3.1 Geographic Clustering

To form an idea of what many individuals are doing it becomes necessary to group them together and examine the overall behaviour.

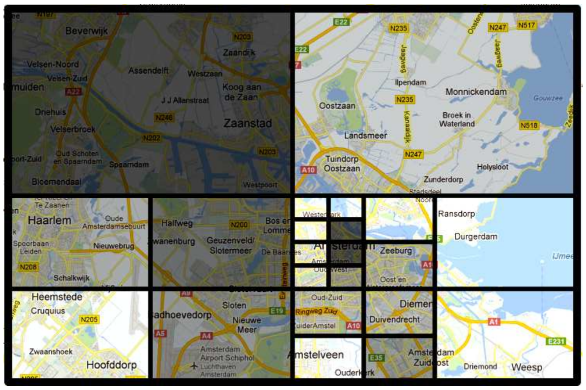

Figure 1: Quadtree of Amsterdam city, showing how a greater density of points leads to greater subdivision of the area. Areas of similar opacity contain a similar number of points.

When aggregating users in a geographic area an obvious approach would just be to divide the area into equally sized squares. However this will lead to problems if there are large density variations across the region. Sparse squares would contain much less useful information and possibly not be comparable with others. While users in dense squares would be considered at an unnecessarily coarse-grained level. Hence we propose using spatial *quadtrees* to allow natural subdivision of the area into roughly equally dense squares, as shown in Figure 1. Quadtrees also have the beneficial property of efficient search and point comparison in $O(\log n)$ time (where n is the number of points).

The quadtree can be constructed as follows, each data point (a latitude and longitude pair) is sequentially added to the geographic area, which is initially one large cell. If any cell exceeds its capacity C of data points, the area and its points are subdivided into 4 equally sized sub-cells. To avoid cells becoming too small a distance limit should be placed on them, under which they will not be subdivided even if their capacity is exceeded. This data structure can then be projected on to maps for responders to understand where people are located in the affected area. As shown in the figure, the user density of cells could be annotated to provide even more context.

When the area is appropriately divided into manageable regions, more useful comparative analysis can be performed using users' geographic movements through the city. The most important consideration at this point is that the post-disaster response will, by its very nature, be affected by specifics of the disaster and its effect on the city. There will likely be a large increase in peoples' desire for media usage to comment on and receive information about the disaster. Conversely, if people (or their connectivity) are seriously damaged by the disaster, they may be unrepresented in the generated data stream. Understanding where both these factors are occurring around the disaster will allow greater insight into its impact and how to respond to it. To this aim emergency response crews need as much situational information as possible.

3.2 Location Transitions

Measuring the density of user data from before and after provides a first level of analysis. Furthermore, examining how people are moving can be even more useful. If a user provides a geographic update from position $p1$ and then $p2$, it shows the user has moved between those points, providing some notion of location transition and connectedness. Indeed, if many people are moving between those points, it may demonstrate an important thoroughfare in an area. This can be important during disaster response when people are evacuating an area. If citizens are not transitioning between previously popular points, it may indicate there is some infrastructure damage, such as a bridge collapsing, a road being swept away or maybe even just traffic gridlock. An awareness of viable transport routes is important for delivering medical care and supplies.

Considering groups of users that are making similar transitions (rather than just being in a similar location) allows for more targeted consideration of their behaviour and intent. Mapping all of the cell-to-cell transitions performed by users gives a better method to classify users and allows even more specific aggregation.

3.3 tf-idf Mining

Thoroughly understanding which parts of a conversation are distinctive or important is an extremely challenging task, often requiring a sentient observer with an understanding of the context of the conversation. Though a statistical appreciation of infrequent words or phrases can still be a useful tool when searching for interesting information to help make decisions. A straightforward method for determining how important words are in a document compared to a corpus of documents is *term frequency-inverse document frequency* (tf-idf) [6]. It is used in text mining and information retrieval to understand relative proportionality of words, specifically it counts the number of times a word occurs in a document multiplied by the fraction of times that word occurs in a corpus of documents. Formally, the inverse document frequency is defined as:

$$idf(t) = log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right)$$

Where $|D|$ is the number of documents in the corpus, and the number of documents that contains term t is $|\{d \in D : t \in d\}|$. The idf is then combined with the term frequency tf in the document, to obtain a measure of how important that phrase in the document is:

$$tf\text{–}idf(t,d) = tf(t,d).idf(t)$$

A subgroup of user updates can thus be compared to an overall corpus of user updates to see if they deviate from what is being said by the rest of the population. Selection of the update subgroup and the overall corpus can be varied in order to expose different types of variation, whether they are between the userbase, spatial (to find dangerous areas) or temporal (to understand the situation's evolution). The combination of geographic social media updates and text mining allows observers to classify what specific groups are discussing in dynamic geographic networks. The inverse of this procedure could also be performed, by examining what movements are being performed by all users discussing a similar topic.

To avoid differentiation being made between the words 'quick' and 'quickly', all words should be *stemmed*. Stemming is the procedure of reducing words to their base or root form. We used the popular 'Snowball' stemming tool to create the stemmed corpus [1]. All words are also converted to their lower-case form and extraneous punctuation is removed.

4. ANALYSIS

To examine the proposed techniques, we apply them to a geospatial social media dataset. Specifically, a Twitter dataset collected during 2011 covering the metropolitan area of Amsterdam in the Netherlands. Unfortunately, this dataset does not contain coverage of a significant disaster. It is problematic collecting Twitter data that is individually geographically tagged both before and after an event with sufficient density for meaningful analysis. The inability to predict such disasters is also a confounding factor. This is an issue that we expect will not be as problematic in future as social media uptake and device location awareness become ever more prevalent.

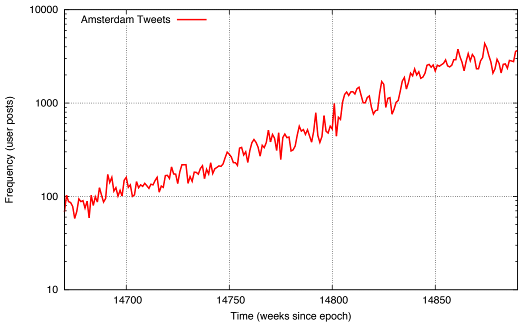

Figure 2: Total frequency of Tweets per week in the Amsterdam metropolitan area. Note the logarithmic Y-axis and thus the exponential growth.

This geographic area is defined as from latitude 52.3°N-52.5°N and longitude 4.6°E-5.1°E, an area that nearly 3 million people live in (Figure 1). The Netherlands has very high social media penetration, *comScore* rated it as the highest in the world in March 2011 [3]. The near exponential growth of Tweet volume in the dataset is shown in Figure 2, a property that is encouraging for the future usefulness of social media as an information source. We focussed on a four week period in May 2011.

4.1 Density

The most straightforward metric to consider in a geographic dataset is the density of users (or tweets), as performed in [8]. The distribution of people in a large city will be of paramount importance to the planning and execution of a disaster response. This could simply be graphically presented with an optional overlay of user density using varying opacity as in Figure 1. The total volume of tweets in the dataset was over 100,000, containing more than 8,000 users. The generated corpus of word frequencies contained over 123,000 unique words (or character strings) after the stemming procedure. The stemming was performed by a Dutch Snowball stemmer.

4.2 User Movement

Movement between cells was performed by nearly all users as they went about their lives. This movement is in fact only 'movement' between subsequent tweets, representing an under-sampling of true user movement, but potentially capturing only the important points of a user's path. As would be expected, when the quadtree cells were given lower capacity ($C < 50$), they would split more and so more consecutive user updates would be perceived as transitions between areas. This greater fidelity highlighted small scale user movement more than when cells were large, both are interesting, depending on what is being asked of the data. Movement distances were greater during the daytime, following expected rush hours and were generally between areas of residential areas and the city centre, particularly transport hubs.

The texts of all tweets was taken as the tf-idf corpus. When transitions within the same cell are considered all of their conversations

Term	tf-idf	Raw Count
schiphol	652.19	159
airport	282.25	58
13g8fe	164.03	28
amsterdam	153.50	79
amstelven	135.83	25
beekstrat	105.25	16
evert	104.82	16
loung	94.61	14

Table 1: Top distinctive words from people staying at Amsterdam Schiphol airport cell [52^o18N, 4^o45].

Term	tf-idf	Raw Count
amsterdam	89.38	46
3p13m3	71.18	13
central	68.55	14
other	58.62	16
station	51.92	13
stationsplein	45.12	8
caf	30.86	7
lunch	27.61	5

Table 2: Top distinctive words when moving from cells [52^o36N,4^o89E] (Amsterdam Old Town) to [52^o38N,4^o89E] (Centraal Station).

are reasonably similar, with the airport area having an immediately understandable lexicon, see the top tf-idf terms in Table 2. The string *13g8fe* is the Foursquare [1] code is caused by people ‘checking in’ and Tweeting about it, despite the lower number of occurrences, it is distinct across the corpus.

The terms revealed by the data in a popular cell-to-cell transition (from the Old Town to the station) is shown in Table 1. Many expected terms are present, *3p13m3* is again the Foursquare code for the Centraal Station. There are also conversations related to transport, eating and lunch, due to office workers using the city centre in the day. A more detailed breakdown of users moving into, out of and within a cell is shown in Figure 3. Foursquare codes can be seen, together with certain parks and entertainment venues as people discuss what they are doing and where they are going. It is only possible to separate these disparate groups in the Old Town, by considering user movements as well as what they are discussing.

5. CONCLUSIONS

This paper has presented ideas on how to use location tagged social media data to understand the dynamics of a population’s conversation in a large metropolitan area. Some potential of applying text mining and information retrieval techniques to geographic dynamics of social media data was also shown. Monitoring the conversations of the population could prove useful in understanding the needs of people after a disaster. Geographic social media can be used to not only characterise people’s conversations in particular areas, but the subdivision of these users according to their movements can provide more contextually relevant synopses. Specifically analysing subsets of the population that have specific movements can allow targeted mining of data, exposing the concerns of people performing similar journeys. This can be of particular interest if the ability to travel in an area has been compromised.

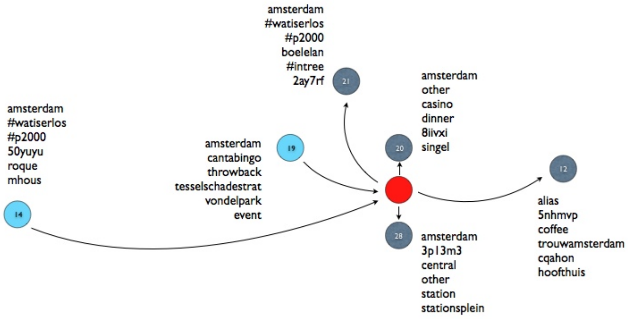

Figure 3: Transition conversation diagram from the Amsterdam Old Town [52^o36N,4^o89E].

The results are only indicative of the Amsterdam dataset, however we believe that such techniques would prove useful when applied to a currently occurring disaster. What is most important is to provide tools for responders that can immediately be applied to a given disaster situation to help their information gathering needs.

For future work we are going to apply these techniques to a dense geographic social media datasets that cover a significant disaster event to see how the conversations and movements change. We will also consider more advanced topic modelling and sentiment analysis to use the context of discussions to gain more information. Using more advanced text mining and natural language processing techniques to derive meaning from the vast collections of data is the most important direction to develop this work. The creation of software that can actively monitor social media streams and perform the sort of processing described in this paper would be an useful additional information source for emergency responders.

Acknolwedgements

We would like to thank Christian Doerr of TU Delft for the collection of to the Twitter data.

6. REFERENCES

[1] Snowball Stemmer. Retrieved from http://snowball.tartarus.org Jan 2012. Covered by BSD license.

[2] Ushahidi Open Source Project Website: http://ushahidi.com/. Retrieved Jan 2011.

[3] comScore. Press Release: The Netherlands Ranks #1 Worldwide in Penetration for Twitter and Linkedin. April 2011.

[4] Michael F Goodchild and J Alan Glennon. Crowdsourcing Geographic Information for Disaster Response: A Research Frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

[5] Theus Hossmann, Franck Legendre, Paolo Carta, Per Gunningberg, and Christian Rohner. Twitter in Disaster Mode: Opportunistic Communication and Distribution of Sensor Data in Emergencies. *ExtremeCom 2011 - The Amazon Expedition*, September 2011.

[6] Karen Spärck Jones. A Statistical Interpretation of term Specificity and its Application in Retrieval. *Journal of Documentation*, 1972.

[7] Gisli Olafsson. Information and Communication Technology Usage in the 2010 Pakistan Floods - A Case Study by NetHope. 2011.

[8] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet Trend Analysis in an Emergency Situation. In *Proceedings of the Special Workshop on Internet and Disasters*, SWID’11, pages 3:1–3:8, New York, NY, USA, 2011. ACM.

[9] Kate Starbird. Digital Volunteerism During Disaster: Crowdsourcing Information Processing. *Conference on Human Factors in Computing Systems*, 2011.

[10] Matthew Zook, Mark Graham, Taylor Shelton, and Sean Gorman. Volunteered Geographic Information & Crowdsourcing Disaster Relief: Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2010.

[1] Foursquare is a location-based social network that allows users to notify others of their presence at important locations.