A real-time social media aggregation tool: Reflections from five large-scale events

Jakob Rogstadius,
Vassilis Kostakos
M-ITI, University of Madeira
9000-390 Funchal, Portugal
{jakob,vk}@m-iti.org

Jim Laredo,
Maja Vukovic
IBM T.J. Watson Research Center
Hawthorne NY 10532, USA
{laredoj,maja}@us.ibm.com

Abstract. Social media is gaining interest among emergency response professionals and researchers as a source of real-time information about ongoing events. In this paper we summarize conclusions drawn while using a software prototype that collects, clusters and visualizes status updates from the social microblogging service Twitter in relation to five large-scale events during the spring of 2011. We identify different uses for gathered information and discuss metrics for prioritization of information with regard to these uses.

Introduction

In the period of time following a natural disaster or other large scale, situational awareness for individuals is usually limited to rich knowledge of their immediate surroundings, combined with sparse high level summaries provided by traditional media. During recent events such as earthquakes, elections, bushfires and terrorist attacks, people have begun to share their knowledge on a micro level with others through openly accessible online social media such as Twitter (Burns and Eltham, 2009; Slagh, 2010; Vieweg et al., 2010).

Now, more and more often, reports of incidents get published through social media before they reach traditional media. This trend suggests that in the future decentralized situational reports from the public will be an increasingly accessible

and important source of information for emergency responders. However, despite the timeliness of this appropriation of social media, it remains challenging for users to overview and navigate the torrent of information from large scale events. Additionally, the absence of up-to-date summaries and validity checks for each claim highlights the need for systems that give structure to the information flow surrounding an event, making it accessible for decision-making in real-time emergency response coordination.

In our ongoing research we develop a system (Figure 1) that leverages social media and crowdsourcing to improve real-time situational awareness during emergency response (Rogstadius et al., 2011). Parts of the proposed system's architecture have now been implemented as a prototype that focuses primarily on information deduplication and the exploration of news clusters. It collects status updates (tweets) from the social microblogging service Twitter in relation to selected current events and clusters these tweets into what we call stories, so that Twitter activity can be visualized and consumed on the basis of distinct pieces of information rather than individual tweets. In this paper, we reflect on the performance of our system so far and we draw conclusions that have implications for the development of future social media information management systems for emergency response use.

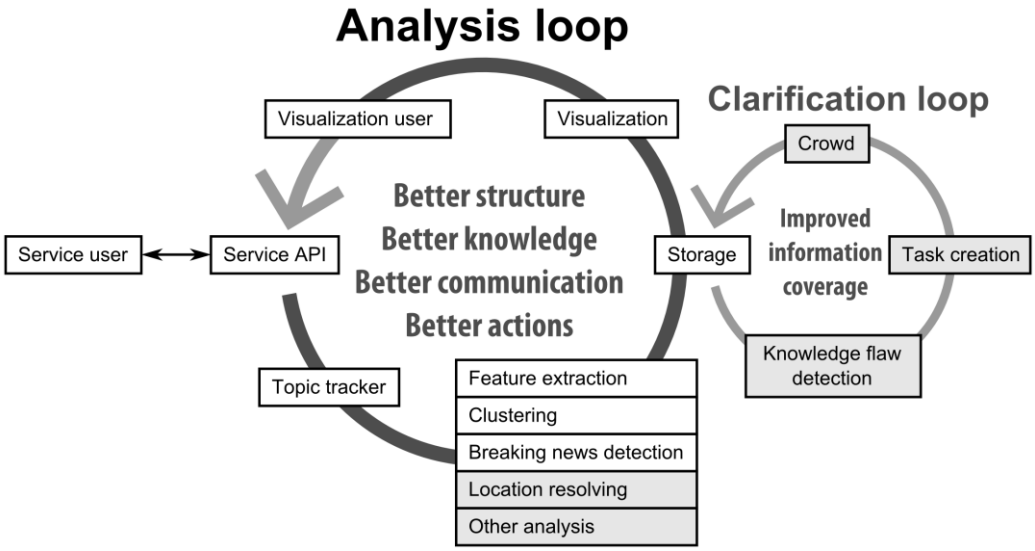

Figure 1. The information flow in the system (Rogstadius et al., 2010) incorporates two feedback loops. The prototype described in this paper implements most of the analysis loop, while grayed segments of the architecture remain to be implemented.

Our central use of textual stories is the primary difference between our system and previous systems that visualize Twitter activity, such as TweetDeck (2011), Twitris+ (Sheth et al., 2010) and Eddi (Bernstein et al., 2010). Another notable system is Ushahidi's Swift River (http://swiftly.org), which is still in the early stages of development without any published evaluation of the system's

performance in terms of reducing information overload. This system is noted for its potential in leveraging social media for emergency.

Description of system

Data collection and clustering of messages

Using our prototype system, an administrator can define the major events to track by describing them as sets of weighted keywords. We will refer to such a set of keywords as a topic. An admin-defined subset of the topic keywords are tracked using Twitter's streaming API, resulting in an incoming stream of tweets; each containing at least one of the tracked keywords and therefore considered potentially relevant to the tracked topics.

The tweets are then split into words, stemmed, and weighted against global word frequencies to generate a term-weighted word vector for each tweet. The cosine distance between the tweet vector and each admin-defined topic vector is then calculated, to determine the tweet's general "appropriateness" for each topic. If the tweet is similar enough and if it contains enough information (as determined by the length of its word vector), it is assigned to its most similar topic, else it is discarded.

Once assigned to a topic, tweets are clustered using a modified version of the algorithm presented by Petrović, Osborne and Lavrenko (2010). We refer to each cluster as a story, with each story typically containing one or many versions of a core claim (e.g. "#Syria; 60,000 protesters are gathering in Hama!" as well as "Thousands of protesters in the center of #Hama right now #Syria") and multiple retweets. Retweets, in addition to repeating a message, often contain added information or comments from the retweeting user. Therefore, not only the first tweet in a story is of interest from an information point of view, as many aspects of the event can be captured in later tweets. We make no explicit attempts to distinguish between source tweets and retweets (in fact, due to limitations in Twitter's API, it is possible that Twitter never sends us the original tweet or that it arrives after a retweet), but we do keep track of the time when the first tweet in a story was detected. The title that is assigned to a story is, somewhat simplified, the text of the centroid tweet in the cluster, or intuitively the tweet that is most representative of the story.

In a nutshell, the system enables administrators to define topics, and over time a number of stories emerge within each topic, driven by activity on Twitter.

User interface

Users of the prototype system can access the information through a web interface (Figure 2). On the left is a list of currently tracked topics and by clicking on a

topic the interface becomes populated with information related to that topic. Along the top is a graph presenting overall tweet activity (total number of tweets) over time and below that is a reverse-chronological list of detected stories. Stories are also overlaid as circles on the activity graph, with circle size corresponding to the number of tweets contained in the story. Dark blue segments on story bars and circles represent tweet volume over the past two hours.

Figure 2. The web interface for the prototype system.

By clicking on a story, either in the list or on the timeline, a viewer can bring up the message variations that together make up the story. Here, messages are grouped by the leading 20 characters in the message and sorted by descending group size, which effectively collapses retweets lacking added comments into a single item.

Stories containing few tweets are presented when those stories have very recently been detected, but as stories age, increasingly large stories get removed from the interface. This design decision was made to avoid cluttering while still maintaining a long-term summary of key events around the topic.

Discussion

We used the prototype to track five different large-scale events during the spring of 2011 for periods of four to eight weeks each and totaling 300,000 to 500,000 topic-related tweets per event, posted by 52,000 to 142,000 users. The events we tracked were the nuclear disaster at the Fukushima plant in Japan; the civil war following the public protests in Libya; the political protests in Syria; protests in various countries in the Middle East (multiple countries together) and the final and semi-finals of the UEFA Champions League. The events were selected for their high message volume, emergency-related nature (except Champions

League), as well as their expected long duration. In the following sections, we will informally evaluate how the system performed at giving us real-time and historic overviews of these events.

Degree of repetition of information in the dataset

Out of all the tweets we processed, 29-47% of tweets belonging to each topic were retweets. We define a retweet as a tweet containing the standard notation "RT @username", but as users can use other notations, actual numbers are likely higher. Furthermore, our system's processing pipeline classified as many as 60-75% of all incoming tweets (varying by topic) as being highly similar to at least one other tweet. The largest single cluster detected by our system (a NATO airstrike that killed Col. Gaddafi's son Saif al-Arab and three of Gaddafi's grandsons) contained 10920 tweets during 24 hours, plus a few smaller clusters on the same story that were incorrectly split off by the system. These numbers together strongly suggest that clustering can be a useful first step in coping with information overload during large scale events.

Retweeting as an indicator of importance

A system that attempts to extract salient information from a stream needs metrics for information prioritization and for noise removal. However, different metrics are suitable for different user groups and we see three general types of information that such a system needs to identify and work with.

- Information that enables members of the public to follow events and crises that they are not directly affected by. This includes links to news articles that summarize recent changes and real-time mentions of highly influential events.
- Information that helps people directly involved in the crisis to make decisions, e.g. victims and emergency responders. This can be, for instance, locations where help is provided after an earthquake, but also higher level information that has implications for the days to come, such as agreements made between parties involved in a conflict.
- Information too fine-grained to be picked up by more than a few individuals, which when combined with other information provides the necessary input data to a detailed event model.

Information for the general public

Previous work by Starbird and Palen (2010) has suggested that a large number of retweets is an indicator of high-level information that appeals to a broad audience during a crisis (Pearson correlation between total number of users in a story and

the total number of tweets is in our dataset always above 0.95). Our informal evaluation confirms that stories containing hundreds or thousands of tweets are indeed of greater interest to the general public than those containing only a few messages, though the exact size of one story relative to another appears to be quite random. This randomness could be due to limitations in our clustering algorithm or due to that we disagree with the general public on what is interesting. It could also indicate that Twitter's information propagation structure introduces randomness in ways similar to what Salganik, Dodds and Watts (2008) observed in the popularity of songs in an artificial music market. Though we lack a controlled study, our impression is that like in their work, stories containing information that is of little public interest are always small and stories of great public interest are always large, but the amount of Twitter activity around a story of medium-level public interest cannot be predicted well based on information content alone.

Most top stories in our dataset related to the conflict topics contained links, while the top stories related to the Champions League generally did not.

Information for domain experts

Information of value primarily to a local audience (particularly useful for crisis intervention) is not well captured by measuring the overall Twitter activity. Starbird and Palen (2010) suggest that a better metric is retweeting of information among users who are local to the event, but to get this information, they relied on manually examining the message history of each Twitter account to determine who the local users are. For our streaming scenario we have looked at automated metrics and the one we find to perform best is to keep track of the number of event-related tweets that have been posted by each user. The most frequently seen users are then treated as a set of domain experts, and message clusters are ranked by the number of expert users whose tweets are found within each cluster. This metric also acknowledges the fact that not all domain experts may be physically located on-site. Spearman's rank correlation scores between the total number of tweets (or users) in a cluster (story) and the number of expert users who contributed to the story is fairly low, around 0.6 for all topics. Such low correlations indicates that expert users tend to contribute to somewhat different stories than normal users, and support the argument that different target audiences seek to access different information from the type of information management system that we aim to develop.

Alternatively, instead of using purely computational techniques to identify domain expert tweeters, a crowdsourcing approach could be used to classify users based on their tweet history.

Information for event modeling

Our system cannot yet detect stories or individual tweets that contain who-what-when statements (e.g.”"The fourth division in the army are now bombing Alrastan from four sides with T-72 tanks #Homs #Syria”) that could contribute to a larger event model. However, aside from the text processing issues involved in parsing text that often contains errors in grammar and spelling, we have found a few Twitter users who invest significant effort on posting such updates. Often these posts are so fine-grained that they are never retweeted. Identifying such users and tracking their tweets may be a good future approach to collect data with minimal noise.

Coping with spam

We have observed that a common spamming approach on Twitter is to post messages containing a short message (e.g. “Cool pictures”), a large set of popular keywords in random order, together with a shortened URL. In relation to the political protests in the Middle East, we also observed what appeared to be an attempt at a form of denial-of-service attack, where a large number of messages containing only popular keywords would be posted from multiple accounts. Occasionally so many messages were posted that it became difficult to find legitimate posts containing these keywords using the search provided by the Twitter website.

By merging similar tweets into clusters, our prototype system represented this flood of spam messages as one or a few large stories and while the spam was noticeable, the system did not suffer from problems of information occlusion.

A second type of spam that we observed took the form of one or multiple users that posted the same story multiple times (up to hundreds). This spamming strategy is well dealt with by using number of unique users as an importance metric for stories, but not by using total number of tweets.

Differences in message types between periods of high and low activity

During the time of our system evaluation, protests in Syria were taking place primarily on Fridays and total tweet counts for this topic was on average twice as high on Fridays as during other days, with short peaks ten times higher than normal daily activity. To see if this volume difference reflected a difference in type of content, we looked at the weekday of the detection of each story as well as which users participated in sharing of the story. We found that only 11 of the top 50 stories among the general public discussed events on Fridays, while for domain experts this number was 19. This hints that the information shared by domain experts is indeed of a somewhat different nature than that shared by the

general public and that the two user groups would benefit from having access to separate rankings of information.

Conclusion

Social media is rapidly gaining interest in the emergency response community because of its ability to decentralize information creation, distribution and prioritization, as well as its potential to deliver highly detailed and timely reports of the situation on the ground. However, this information source often results in information overload due to its high volume and lack of structure.

In this paper we have discussed our findings from eight weeks of using a prototype system that collects, clusters and visualizes status updates from the social microblogging service Twitter. We conclude that clustering is an efficient method to reduce information overload when tracking large-scale events, by grouping together similar pieces of information from throughout the network into single stories, for which sharing characteristics become a form of associated meta-data. We also identify metrics that can be used for prioritizing information for different target audiences: total number of sharing users for the general public, and number of sharing domain experts for domain expert users. Finally, we discuss how well these metrics (as well as total tweets) cope with different popular spamming strategies.

Future work

Much work remains before the system is complete and ready for use in emergency situations. Primarily, location information needs to be extracted from stories, as well as references to entities (e.g. people and organizations). In addition, we plan to make extensive use of crowdsourcing to cope with problems that are computationally difficult to solve and to make the system as adaptable as it needs to be to function properly during crises.

Acknowledgments

This work is funded by an IBM Open Collaboration Research award and by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy).

References

Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S. and Chi, Ed. (2010): 'Eddi: interactive topic-based browsing of social status streams'. In *Proc. UIST 2010*, ACM Press, 2010, pp. 303-312.

Burns, A. and Eltham, B. (2009): 'Twitter free Iran: An evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 election crisis'. In *Record of the Communications Policy & Research Forum 2009*, Network Insight Institute (2009), pp. 298-310.

Petrović, S., Osborne, M. and Lavrenko, V. (2010): 'Streaming First Story Detection with application to Twitter'. In *Proc. HLT 2010*, Association for Computational Linguistics (2010), 181-189.

Rogstadius, J., Kostakos, V., Laredo, J. and Vukovic, M. (2011): 'Towards Real-time Emergency Response using Crowd Supported Analysis of Social Media'. *CHI 2011 Workshop on Crowdsourcing and Human Computation*, Vancouver, Canada (2011).

Salganik, M., Dodds P. S., Watts, D. (2006): 'Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market'. In *Science*, 311 (854), pp. 854-856 (2006).

Sheth, A., Purohit, H., Jadhav, A., Kapanipathi, P. and Chen, L. (2010): 'Understanding Events Through Analysis Of Social Media'. In *Proc. WWW 2011*, ACM Press (2010).

Slagh, C. L. (2010): *Managing chaos, 140 characters at a time: How the usage of social media in the 2010 Haiti crisis enhanced disaster relief*. Georgetown University, USA (2010).

Starbird, K. and Palen, L. (2010): 'Pass It On?: Retweeting in Mass Emergency'. In *Proc. ISCRAM 2010*, Seattle, USA (2010).

TweetDeck. (2011): http://www.tweetdeck.com.

Vieweg, S., Hughes, A., Starbird, K. and Palen, L. (2010): 'Microblogging during two natural hazards events: what Twitter may contribute to situational awareness'. In *Proc. CHI 2010*, ACM Press, 2010, pp. 1079-1088.