What's Going On? How Twitter and Online News Can Work in Synergy to Increase Situational Awareness

Christian Rohrdantz[*]
University of Konstanz
Germany

Miloš Krstajić[†]
University of Konstanz
Germany

Mennatallah El Assady[‡]
University of Konstanz
Germany

Daniel A. Keim[§]
University of Konstanz
Germany

ABSTRACT

Recent studies have shown that micro-blogging services such as Twitter can outperform online news aggregators in reaction time to breaking events. On the other hand, disadvantages of Twitter, such as low signal-to-noise ratio, inaccuracy of geographical data, and lack of in-depth analysis prevent it from becoming a reliable news source. In this paper, we investigate how the different information sources Twitter and online news can be integrated to gain a better understanding of current incidents for emergency response management. Our initial results suggest that the combination enables both a quick event detection and a long-term tracking of the topical and geo-spatial implications of an event. To achieve situational awareness we suggest a graph-based topic analysis approach building on top of Latent Dirichlet Allocation, as well as a new glyph design for plotting categorical data on a map, which can be used to compare news topics.

INTRODUCTION

Recently, research on social media has received a lot of attention in information retrieval, knowledge discovery, data mining, and visualization communities. Micro-blogging services such as Twitter are being actively used by millions of users, who exchange information on a wide range of topics in real-time. Among these topics, people often comment on real-world events as they happen, thus creating a potentially valuable emergency response tool. An important line of research is trying to better understand the usage of Twitter as a reliable source of news information. Studies (e.g. [9]) have shown that in some cases Twitter can deliver breaking news faster than traditional media channels. However, Twitter has several fundamental drawbacks compared to news portals:

Tweets are limited to 140 characters and the use of language adapts to this constraint, i.e. tweets often contain cryptic abbreviations, ungrammatical language, and they lack details. As a consequence, state-of-the-art text mining algorithms (e.g. for topic modeling) do not perform well on tweets. The content of news articles is much richer and easier to process, and thus they can provide much more insightful and detailed information about an incident. Twitter users generate a lot of noise and generally irrelevant information. According to Naaman et al. [6] only 20% of the users can be categorized as *Informers* in contrast to 80% of *Meformers* who "typically post messages relating to themselves or their thoughts" [6]. Breaking news is often associated with a geographic location. Although, approaches for extracting geo-information from the tweets exist and the users can embed their GPS coordinates to geographically identify their tweets, currently, this feature is rarely used or it is not reliable enough. According to Cheng et al. [3] only 0.42% of all tweets are associated with a latitude and longitude and for 74% of all Twitter users the user locations "are overly general (e.g., California), missing altogether, or nonsensical (e.g., Wonderland)" [3].

Therefore, we can assume that Twitter, as a news source, in general can deliver breaking news information more quickly than traditional media, at the expense of lacking depth and precision. In this paper, we make use of the competition between social media and other online news channels combining them and exploiting their particular strengths for the task of emergency response and management in cases like natural disasters, accidents, terrorist attacks, and other catastrophes. More concretely, we present the concept of combining data from Twitter and news aggregators to help the analyst to get a faster and above all a better understanding of emergency events and achieve long-term situational awareness [8]. First, we describe the data we use and the key components of our system: (1) Event Detection and Information Fusion, (2) Geospatial Analysis using Circular Glyphs, and (3) Graph-based Topic Keyword Exploration. Next, we present initial results for an analysis use case that demonstrate the good performance of our techniques. Finally, we conclude with a brief discussion and outlook on future work.

SYSTEM

The following subsections describe the core components of our system.

Data

The Streaming API of Twitter with the so-called "gardenhose" level, enables us to collect 10% of the public Twitter live stream. The stream is processed using an extended version of a native XML database [10]. In addition, we use the Europe Media Monitor (EMM) news stream [1] which is available through a public website[1]. The EMM monitors about 2500 international news sources and enriches the news articles with metadata such as geo-tags.

[*] e-mail: christian.rohrdantz@uni-konstanz.de
[†] e-mail: milos.krstajic@uni-konstanz.de
[‡] e-mail: mennatallah.el-assady@uni-konstanz.de@uni-konstanz.de
[§] e-mail: daniel.keim@uni-konstanz.de

Published at the 2nd IEEE Workshop on Interactive Visual Text Analytics
Task-Driven Analysis of Social Media as part of the IEEE VisWeek 2012,
October 15th, 2012, Seattle, Washington, USA.

[1] http://emm.jrc.it

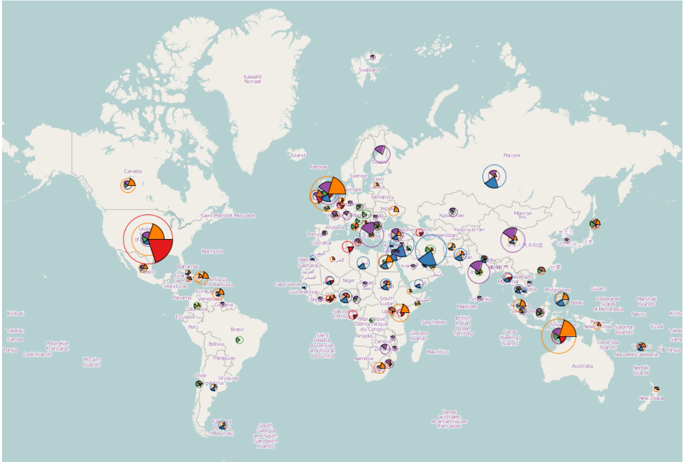

Figure 1. Map with topics learned for the whole news data on February 6th, 2012. The high-resolution screenshot is zoomable in the pdf version of the paper. The following keywords are the top topic descriptors according to the applied latent dirichlet allocation approach including one manually assigned label in italics that summarizes the further keywords: giants, powell, super, patriots, bowl ⇒ *super bowl*, iran, israel, government, syria, president ⇒ *political conflicts*, company, market, information, companies, profile ⇒ *economy*, year, government, percent, billion, country ⇒ *debt crisis*, police, people, year, years, told ⇒ *misc*.

Event Detection and Information Fusion

In recent years, a variety of event detection algorithms have been introduced for Twitter, see [2] for an overview. We decided to customize our own event detection algorithm, which is an extended online version of the approach described in [7], but in principle this component can be exchanged using any of the other existing algorithms. Once the algorithm detects an event of interest on Twitter, we complement the corresponding tweets with related information from the news. This is done using a keyword search within the news stream. The keywords stem from the event tweets and have to be contained in the news article headline or meta-data tags. From all news articles retrieved using the keyword queries, we use the fulltext for topic modeling and the geo-tags for further analyses.

Glyphs for geographical topic comparison

From the news data we can use the geo-tags indicating the location of the news source and the location of the news destination, i.e. the location that the news talks about. The destination location is more interesting for emergency response. Figure 1 gives a first impression of the geo-analytic component of the system representing the whole news data of a day. For such an overview visualization we aggregate more detailed location information to the country level. For each country appearing as a news destination, we create a glyph that shows the topic distribution of the documents that talk about the country. The topics for each country are conveyed using a glyph-overlay over a world map. Figure 2 illustrates the computation of the glyph. The topics are represented through circle segments. The angle width of each segment is identical, a topic can be easily identified by both color and orientation of the segment. This redundant coding enables the analyst to readily scan the geographical distribution of a certain topic. The radius of a segment is determined in such a way that the segment area corresponds to the number of documents belonging to the topic. This fact makes the glyph highly scalable, because with an increasing amount of data the circle radius will increase only at a logarithmic scale. For one certain location the sizes of the different circle segments are hard to compare, because of the different orientations. To overcome this drawback, for each topic we additionally insert a smaller circle into the glyph with the radius and color of the topic segment. With this visual aid the sizes of different topics at one location can be readily compared.

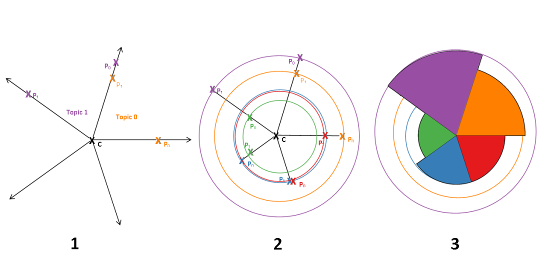

Figure 2. Steps for creating the glyphs.

Another advantage of the glyph is that it can be used for pixel placement, where pixels represent single documents and can be overlaid onto the segment of its topic. The color of pixels can then be changed to represent an additional data dimension like, for example, document sentiments through color hue or topic probabilities through color saturation.

Graph-based Topic Keyword Exploration

For topic modeling we use the Latent Dirichlet Allocation (LDA) method provided by MALLET[2]. As an output of the LDA we get for each topic a ranked list of relevant keywords. The number of topics to be generated with LDA is a user-specified parameter, however, it is not quite clear beforehand how many topics to request. Moreover, when the topics are returned, it is also not quite clear if and how they are related or connected. In order to face these two disadvantages we suggest to add another analysis layer on top of the LDA topic modeling, performing a graph-based exploration of the top keywords for the different topics. To do so, for each topic we consider a smaller set of keywords, which have the strongest association with the topic according to a statistical significance test. In the graph each keyword is a node and two keywords are connected with an edge if and only if they co-occur within documents statistically significantly more often than by chance. To measure the statistical significance, in both cases, we use the likelihood ratio test. This hypothesis test has been successfully applied before to investigate word collocations in [5], where also more details about the method are provided. The significance value is then used for weighting the edge between the corresponding keyword nodes. We visualize and further explore the keyword graph with the graph analysis tool VISONE[3]. In Figure 3 the edge connections reveal four clusters and it becomes evident that three of the clusters (the red, green, and purple one) correspond to one topic each, while the blue and orange topics are merged into

[2] MAchine Learning for LanguagE Toolkit (MALLET) http://mallet.cs.umass.edu/topics.php revised on 04/09/2012

[3] http://visone.info/html/about.html

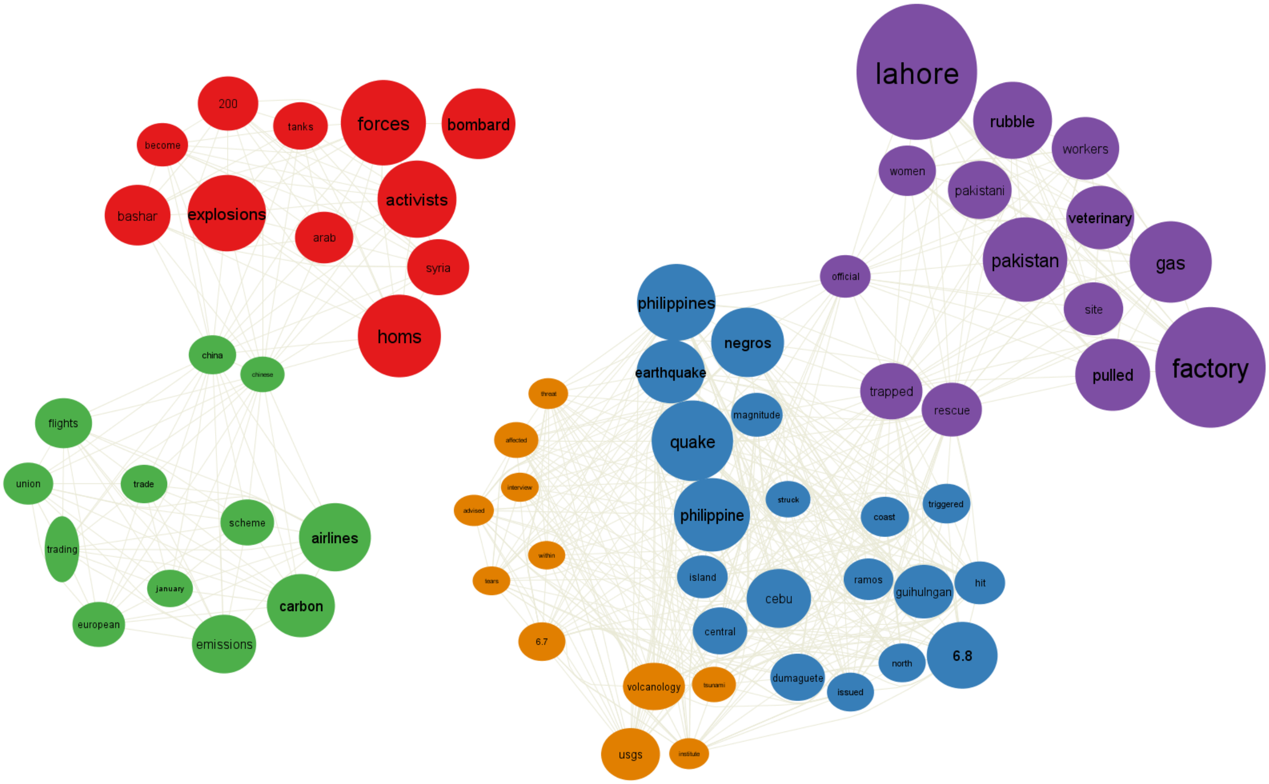

Figure 3. Keywords as nodes with a stress-minimizing layout to the edges (significant co-occurrence of two keywords), node colors represent LDA topics. The high-resolution screenshot is zoomable in the pdf version. Events: (1) China bans airlines from paying EU carbon emission charge, (2) Syrian forces bombard Homs, 15 killed, (3) Pakistan factory collapses in gas blast, 3 dead, many trapped, (4) Earthquake hits the Philippines, at least 43 dead, dozens missing, (5) Aftershocks rattle Philippines, false tsunami alert.

one cluster. In the next section (Use Case), we describe that this actually makes sense, because while the other topics are disjoint, the blue topic (Earthquake hits the Phillipines) is closely related to the orange topic (Background information on the Earthquake). Another analysis option in VISONE is to apply a method that divides the graph into hierarchical clusters. One such clustering method we have applied is the Girvan Newman Clustering (GNC) [4]. The result is provided in Figure 4 and again shows the close connection of the blue and orange clusters.

USE CASE: WHAT HAPPENED ON FEB 6TH, 2012?

Knowing that there were some important real-world events on Feb 6th, 2012, we first run the tool on the unfiltered news data to get an impression of the geographical spread and size of the major news topics, see Figure 1. Even though all found topics are relevant at that day, they were not newly upcoming and too general to point to particular emergency events. Typically, events have an earlier and stronger representation within Twitter than within news. Therefore, we use our customized real-time capable event detection method on Twitter to find out about events on Feb 6th. The relevant tweets along with their most important keywords are then fed into the tool and used as filter in order to get a relevant subset of the news data. Topics are trained from the joined set of event tweets and the event related news. We experimented using different numbers of topics between 2 and 10, and found out that 5 topics yielded the best results, see also Figures 3, 4, and 5. Two of the 5 topics deal with an earthquake on the Philippines, one topic is located in China, one in Pakistan, and one in Syria. While going through the event keywords, articles, and tweets, we found out, that two out of the 5 topics can only be found in the news data, namely the orange and green topic, while all tweets are assigned to the remaining three topics.

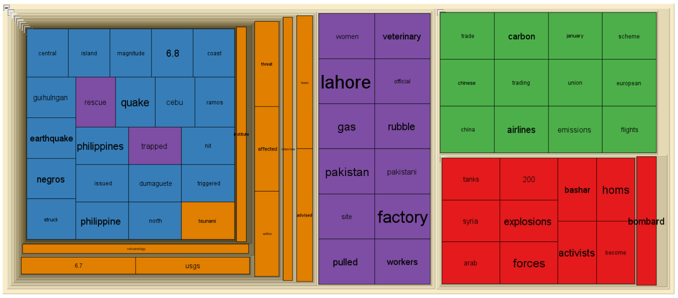

Figure 4. Treemap generated with VISONE representing the clustered keyword graph with label size mapped to keyword weight.

Relations among topics

Figure 3 reveals that the Homs bombardment topic and the carbon emission charge topic are connected through the keywords *china* and *chinese*. This points to the fact that China plays a role for both topics, i.e. it disputes with the European Union, because of the carbon emission charge for airlines, and vetoes against a United Nations Syria resolution. Actually, keywords like *china* appearing in event tweets about the Homs bombardment were the reason that the carbon emission

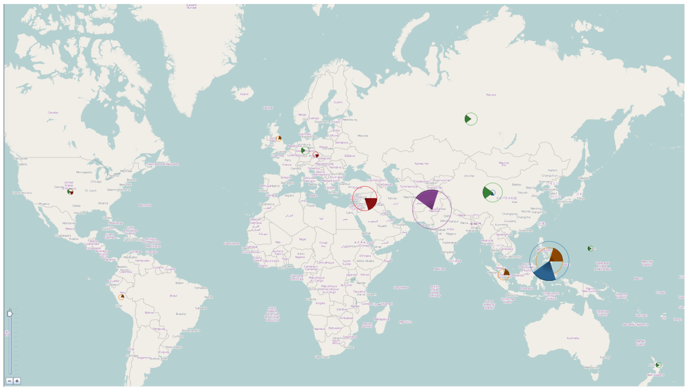

Figure 5. Map showing the geo-spatial distribution of the topics learned from the event tweets and related news published on February 6th, 2012. The high-resolution screenshot is zoomable in the pdf version of the paper. The topics have the same color coding as in Figure 3 and 4.

charge topic was taken into account in the first place. While in this case an actor, China, brings two topics into relation, the factory gas blast is connected to the Phillipine earthquake and tsunami alert through content. The keywords *trapped*, *rescue*, and *official* indicate that both incidents caused people to be buried alive. While these uncovered connections are interesting, they are less important for situational awareness. For the Phillipine earthquake this is different.

The Philippine Earthquake

Figure 3 suggests that the earthquake triggers two subtopics, the first is about the actual event and partly also about the false alarm for a tsunami caused by the quake. The second topic joins all news files dealing with the analysis of the cause of the quake and its consequences on the people and the economic situation of the country, including the tsunami alert. As mentioned before, the tweets were only assigned to the first topic. In other words, the Phillipine earthquake event gets complemented with additional information from news. The hierarchical community structure provided in Figure 4 additionally points to the fact that the orange topic consists of several sub-communities. It shows that news discuss diverse further aspects and give detailed analyses of situations.

CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the initial results of our approach to combine different sources of textual information. We use Twitter as a starting point for our analysis, where we assume that emergency events can be detected sooner on Twitter than in traditional news media. The additional news metadata, such as the extracted geographical location of the event, is used together with topic modeling results to focus the attention of the analyst on the world regions of high interest. The topics can be further analyzed and disseminated using different visualization techniques. We have presented a case study with a graph-based visualization, which shows that news articles can complement information from tweets, as in the case of the Phillipine earthquake, and that in other cases interesting connections among hardly related events can show up. Apart from describing smaller technical contributions the purpose of this paper is to provide answers for two open issues regarding the task of emergency response management:

1. *Why use Twitter*? Real-world emergency events lead to a fast and strong signal on social media such as Twitter and provide focused information for quick reaction.

2. *Why not use only Twitter*? Due to their brevity tweets provide only a very limited amount of information and often do not come with sufficient accurate geo-information. Here, online news data can be very useful to provide detailed additional information about the situational development, diverse potential impacts of the emergency event, and geo-spatial implications. This may be especially useful if an incident triggers follow-up events or its impacts last for hours, days, or weeks as for example in cases like the Fukushima catastrophe in Japan.

In our future work we aim to build on this first conclusion and continue the development of our emergency response tool integrating further data sources and additional functionality.

ACKNOWLEDGMENTS

This work was partly funded by the German Research Foundation (DFG) within the project “Visual Spatiotemporal Pattern Analysis of Movement and Event Data” (ViaMod).

REFERENCES

1. Atkinson, M., and der Goot, E. V. Near real time information mining in multilingual news. In *Proc. WWW 2009*, ACM (2009), 1153–1154.
2. Bontcheva, K., and Rout, D. Making Sense of Social Media Streams through Semantics: a Survey. *Semantic Web* (2012).
3. Cheng, Z., Caverlee, J., and Lee, K. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. CIKM 2010*, ACM (2010), 759–768.
4. Girvan, M., and Newman, M. E. J. Community structure in social and biological networks. *Proc. NAS 99*, 12 (2002), 7821–7826.
5. Manning, C. D., and Schütze, H. *Foundations of statistical natural language processing*. MIT Press, 2001.
6. Naaman, M., Boase, J., and Lai, C.-H. Is it really about me?: message content in social awareness streams. In *Proc. CSCW 2010*, ACM (2010), 189–192.
7. Rohrdantz, C., Hao, M. C., Dayal, U., Haug, L.-E., and Keim, D. A. Feature-based visual sentiment analysis of text document streams. *ACM TIST 3*, 2 (2012), 26.
8. Rohrdantz, C., Oelke, D., Krstajic, M., and Fischer, F. Real-Time Visualization of Streaming Text Data: Tasks and Challenges. In *Interactive Visual Text Analytics Workshop at IEEE VisWeek 2011* (2011).
9. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. WWW 2010*, ACM (2010), 851–860.
10. Weiler, A., Mansmann, S., and Scholl, M. H. Towards an advanced system for real-time event detection in high-volume data streams. In *Proc. 5th Workshop for Ph.D. students PIKM 2012*, ACM (2012).