Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the River Elbe Flood in June 2013

Benjamin Herfort
GIScience Chair, Heidelberg University, Germany
Herfort@stud.uni-heidelberg.de

João Porto de Albuquerque
GIScience Chair, Heidelberg University, Germany
Dept. of Computer Systems/ICMC, University of Sao Paulo, Brazil, jporto@icmc.usp.br

Svend-Jonas Schelhorn
GIScience Chair, Heidelberg University, Germany

Alexander Zipf
GIScience Chair, Heidelberg University, Germany

ABSTRACT

In this paper we present a new approach to enhance information extraction from social media that relies upon the geographical relations between twitter data and flood phenomena. We use specific geographical features like hydrological data and digital elevation models to analyze the spatiotemporal distribution of georeferenced twitter messages. This approach is applied to examine the River Elbe Flood in Germany in June 2013. Although recent research has shown that social media platforms like Twitter can be complementary information sources for achieving situation awareness, previous work is mostly concentrated on the classification and analysis of tweets without resorting to existing data related to the disaster, e.g. catchment borders or sensor data about river levels. Our results show that our approach based on geographical relations can help to manage the high volume and velocity of social media messages and thus can be valuable for both crisis response and preventive flood monitoring.

Keywords

Social Media, Twitter, Flood, Water Level, Crisis Management, Situational Awareness

1. INTRODUCTION

Managing an emergency puts high demands on authorities and crisis management organizations. Collecting as much information as possible about the crisis and sense making of that information in a timely manner is critical to enhance situational awareness. Social media platforms like Twitter, Flickr or Instagram are broadly used by many crisis-affected individuals. Hence, this shared local knowledge can be vital sources for crisis relevant information. However the process of collecting and analyzing social media information has to be further evaluated to gain better insights which information contributes to situation awareness.

Scientific research on crisis management and social media has concentrated on filtering and classifying microblog posts, e.g. tweets, applying crowdsourcing or machine learning methodology [Vieweg et al., 2010, Sakaki et al., 2010, Kongthon et al., 2012, Imran et al., 2013]. For instance, Sakaki et al. (2010) were able to detect crisis related twitter messages using a support vector machine. Kongthon et al. (2012) filtered potential relevant twitter messages containing information about the flood that affected Thailand in 2011 using the flood related hashtag "#thaiflood". Imran et al. (2013) tested an automatic method for filtering crisis relevant social media messages versus a crowdsourcing approach. Graham et al. (2012) analyzed the Twitter use during the UK floods in November 2012 for the Guardian Data Blog and mapped geo-referenced tweets mentioning the words

"flood" and conclude that the digital trails of twitter messages are mostly matched to official data on floods and precipitation.

Nevertheless, a crucial problem remains unsolved. During a crisis the volume and the velocity of posted tweets is extremely high. Distinguishing messages that contain critical information from off-topic messages in an efficient and credible way is the basic requirement for any feasible approach for handling information overload. In the end, this leads to relevant and actionable information contributing to situational awareness and better decision-making. Crowdsourcing and machine learning methods can suffice for this only in part. Crowdsourcing-based approaches face scalability problems due to the sheer amount of tweets that need to be manually processed. In contrast, most machine-learning based methods are scalable but are usually defined post-hoc for a specific content and task, thus undermining their generalizability to other crisis scenarios.

Towards making a contribution in this context, we apply a geographical approach to prioritize crisis-relevant information from social media. There is initial work in this field [e.g. Triglav-Čekada and Radovan, 2013], but this research direction still needs to be further pursued. Combining existing and well-studied geographical models about natural hazards with social media therefore offers chances to enhance crisis management. Our methodology is based on specific geographical relations of flood phenomena, for example hydrological features and models of terrain and affected areas, which are generally valid for every flood scenario. In this paper, we conduct a case study for the River Elbe Flood in Germany in June 2013 to validate our approach.

This paper is organized as follows: In the next Section we present our approach. Information about the River Elbe Flood in June 2013 and about the different datasets is given in Sect. 3. In Sect. 4 and Sect. 5 we describe our methodology and present first results. Finally we will discuss our findings and future research directions.

2. RESEARCH APPROACH AND CASE STUDY

Our approach adds a new geographical component to the existing models of information extraction presented in the previous chapter. Taking up the first law of geography [Tobler, 1970] we assume that near things are more related than distant things. Regarding crisis events this implies that the spatial-temporal characteristics of the catastrophe affect the spatiotemporal characteristics of social media messages. Better understanding of the geographical relations between social media and crisis phenomena therefore offers the chance to enhance crisis management and contributes to situational awareness.

We provide an approach that takes these geographical relations into account by combining analysis techniques from both social media research and research on flood phenomena. Combining information from tweets, water level measurements and digital elevation models we examine the River Elbe Flood in Germany in June 2013 and apply our approach to investigate the following research question: *Does the tempo-spatial distribution of flood related tweets refer to the tempo-spatial distribution of the flood phenomenon*?

In the period from 30th May to 3rd June 2013 extreme heavy rain affected large parts of eastern and central Europe. The distribution of precipitation in the basin of the rivers Elbe, Moldau and Saale reached values two to three times higher than that for an average June. This is equivalent to a centennial probability of occurrence. The soil was already highly saturated at this time due to a wet climate in May 2013. Therefore, the heavy rain rapidly resulted in surface runoff causing the severe flood situation. The monthly average flow was three to four times higher than the longstanding average and in some places even higher than the ever recorded value. The same finding follows from the examination of the water level data. Some gauging stations measured values that were never recorded before.

The Twitter dataset contains of 60.524 geo-referenced tweets within the territory of Germany. We queried the Twitter API using the 1% garden hose access from 08th June 2013, 1.30 pm to 10th June 2013, midnight and collected every geo-referenced tweet within a bounding box covering Germany. Afterwards we filtered tweets by their location and excluded those outside the territory of Germany. Each tweet in the sample can be identified clearly by its ID and timestamp.

We analyzed official water level data from 54 water level measurements stations along the rivers Elbe and Saale provided by the German Federal Waterways and Shipping Administration and the German Federal Institute for Hydrology. The Dataset includes information about the location of each measurement station, the current water level, the average flood water level over a time period from 1st November 2000 to 31st October 2010 and the highest water level ever recorded. The current water level measurements were provided in a 15 minutes resolution for the whole examination period.We used HydroSHEDS information derived from elevation data of the Shuttle Radar Topography Mission (SRTM) at 3 arc-second resolution to compute hydrographical features of the river Elbe basin including information about flow accumulation, stream network and catchment boundaries [Lehner et al., 2008].

3. METHODOLOGY

Our methodology is divided into two steps. At first we assess flood-affected regions. Step 2 contains classification and analysis of geo-referenced twitter messages.

Starting with the HydroSHEDS flow direction raster, based on SRTM elevation data, we computed catchment polygon features for each location where two streams flow together using the ArcHydro Toolset for ArcGIS. Next we analyzed the water level data collected from 54 water level measurement stations along the rivers Elbe and Saale. To assess the severity of the flood at the gage station we computed the difference between the daily maximum water level and the average flood water level for the time period from 1st November 2000 to 31st October 2010. At last we combined both information on catchments and water level based on the location of the water level measurement stations. The normalized water level values were then matched to the corresponding catchment regions.

In the second step we grouped twitter messages into the categories “flood-related” and “non flood-related”. This was accomplished using keyword filtering as common practice in the analysis of twitter messages (e.g. Graham et al., 2012, Kongthon et al., 2012, Vieweg et al., 2010). Tweets containing the keywords in German “Hochwasser”, “Flut”, “Überschwemmung” (“Hochwasser”, “Flut” and “Überschwemmung” are the German words meaning “flood”) and the English word “flood”, regardless of case-sensitivity, were considered “flood-related”. The selection of these keywords was based on the definition of the German dictionary “Duden” for the word “Hochwasser”. Furthermore, we included the addtional words “Deich” (dike) and “Sandsack” (sandbag), which were found to be common in reports in the media.

4. PRELIMINARY RESULTS

Figure 1. shows flood-affected catchments and the severity of the flooding calculated from digital elevation data and water level data for the time period from 8th to 10th June 2013. The maps visualize the shift of the flood peak from the upper reaches to the lower reaches. On 8th June 2013 the catchments along the river Elbe in the federal state of Saxony were most affected, whereas the lower reaches of the river Elbe were affected not until 10th June 2013.

The results of the first classification of twitter messages based on keywords are listed in Table 1. Overall we examined 60,524 tweets within the territory of Ger-many. The majority (99.34%) of them do not contain the query words. These tweets were marked as “non flood-related”. For the period from 8th to 10th June 2013 we selected 398 tweets containing the query words and marked these tweets as “flood-related”.

period	8th-10th June 2013	8th June 2013	9th June 2013	10th June 2013
# all tweets	60,524 (100%)	14,286 (100%)	23,093 (100%)	23,145 (100%)
# flood-related tweets	398 (0.66%)	75 (0.52%)	197 (0.85%)	126 (0.54%)
# non flood-related tweets	60,126 (99.34%)	14,211 (99.55%)	22,896 (99.15%)	23,019 (99.46%)

Table 1. Classification of twitter messages using query words

Does the tempo-spatial distribution of flood related tweets refer to the tempo-spatial distribution of the flood phenomenon?

At first we examined the spatial distribution of flood-related and non flood-related twitter messages to review whether they follow the tempo-spatial distribution of the flood phenomenon. Figure 2. shows the density of tweets depending on keyword classification. Flood related tweets (on the right side) show peaks in the regions of Magdeburg, Berlin and Halle. Overall flood-related tweets appear only in a few parts of Germany. Non flood-related tweets (on the left side) concentrate in dense populated regions, e. g. urban areas like Berlin, Hamburg, Munich and the Ruhr area. The tweets cover almost all of Germany, except for some regions in the federal states of Brandenburg and Mecklenburg-Hither Pomerania.

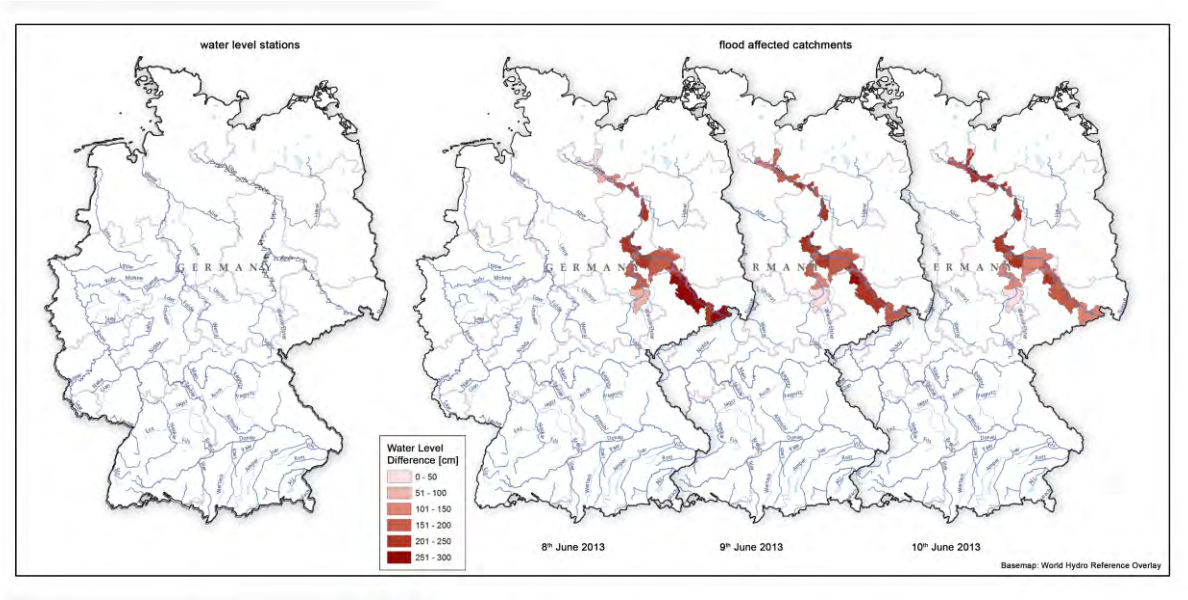

Figure 1. Spatiotemporal distribution of flood affected catchments

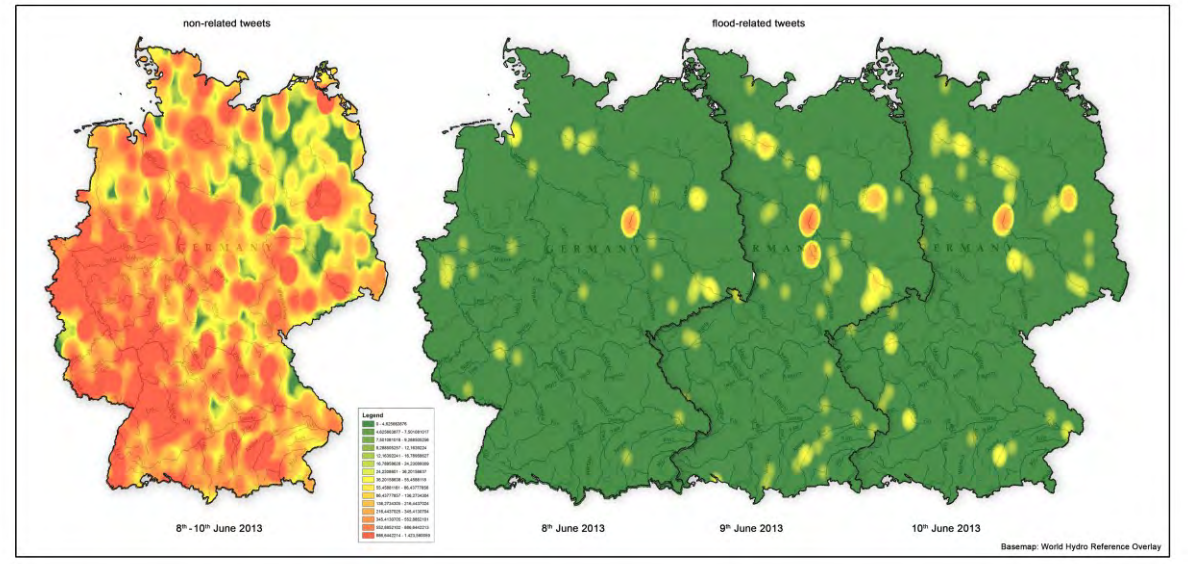

Figure 2. Spatiotemporal distribution of non-related and flood-related tweets

Comparing the distribution of flood-related tweets to the spatial distribution of flood affected catchments (see Figure 1 and Figure 2) one can notice similarities at the first look. Not the location of all flood-related tweets, but at least of a considerable amount of them does correspond to the location of flood affected catchments. In the area around the city of Magdeburg, which was severely affected by the flood for the whole examination period, the density of flood-related tweets is strikingly increased. Furthermore, the increasing water levels in the lower reaches of the river Elbe are also represented in shift of the twitter activity.

	# tweets	Average distance [km]	Standard deviation
non-related	60,126	221	125
flood-related	398	78	121

Table 4. Average distances to flood-affected catchments

To further examine the relationship between flooded areas and flood-related tweets we statistically analyzed the distance of all tweets to flood affected catchments (Table 4). We run an independent sample t-test to determine

if there are differences in distance to flood-affected catchments between flood-related and non flood-related twitter messages. In our study we found out that the distance to flood-affected catchments for flood-related twitter messages was statistically significantly lower (78± 121 km) compared to non-related twitter messages (221 ±125 km), $t(60522) = 22.674$, $p = 0.000$.

This implies that the locations of flood-related twitter messages and flood-affected catchments match to a certain extent. In particular this means that mostly people in regions affected by the flooding or people close to these regions posted twitter messages referring to the flood. That is remarkable as there are for instance far more tweets posted in greater distance to flood-affected regions compared to the number of tweets posted in the proximity to flood-affected regions and as such as that media coverage about the River Elbe Flood was enormous since it was one of the most severe floods ever recorded in Germany. Regarding these circumstances one would have expected a great amount of tweets referring to the flood posted in the urban areas like Munich, Hamburg or the Ruhr area. However, that was not the case. The majority of tweets referring to the flooding was posted by locals.

5. CONCLUSION

In this paper we present a new geographical approach to analyze crisis relevant information from social media platforms like Twitter. Our results show that the spatial distribution of twitter messages referring to the flooding of the river Elbe in Germany in June 2013 is significantly different from the spatial distribution of off-topic messages. This could lead to distance-based prioritization for enhanced filtering and classification of crisis relevant social media messages.

Unfortunately, only a small fraction (3% is the estimated average) of tweets are currently georeferenced by users, and this consists of a limitation for analysis approaches based on the location like the current study. Furthermore, filtering tweets using query words can only be adequate using the "right" keywords. While we manually verified each twitter message included after our filtering to ensure there are no false positives, we cannot rule out the threat that relevant messages were filtered out. Future work will concentrate on refining the approach including additional information from other social media platforms like Instagram or Flickr and on testing our findings using larger datasets and longer time series. In this regard, applying more sophisticated algorithms for filtering, clustering and classification of messages is a major issue for improvement. Furthermore, the integration of other official datasets, e.g. precipitation data, is one additional avenue for better understanding the relations between social media and crisis phenomena from a geographical perspective. Implementing more detailed hydrological models will additionally extend the validity of our method regarding flood phenomena.

REFERENCES

1. Graham M, Poorthuis A, Zook M (2012) Digital trails of the UK floods - how well do tweets match observations? The Guardian Datablog ; http://www.guardian.co.uk/news/datablog/2012/nov/28/data-shadows-twitter-uk-floods-mapped. Accessed 20 June 2013
2. Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. ISCRAM, Baden-Baden, Germany.
3. Kongthon, A., Haruechaiyasak, C., Pailai, J., & Kongyoung, S. (2012). The role of Twitter during a natural disaster: Case study of 2011 Thai Flood. In Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET'12: (pp. 2227-2232). IEEE.
4. Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data. EOS, Transactions American Geophysical Union, 89(10), 93-94.
5. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.
6. Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 46, 234-240.
7. Triglav-Čekada, M., & Radovan, D. (2013). Using volunteered geographical information to map the November 2012 floods in Slovenia. Natural Hazards and Earth System Sciences, 1(3), 2859-2881.
8. Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1079-1088). ACM.