A geographic approach for on-the-fly prioritization of social-media messages towards improving flood risk management

Luiz Fernando F. G. de Assis¹, Benjamin Herfort²,
Enrico Steiger², Flávio E. A. Horita¹, João Porto de Albuquerque¹,²

¹Institute of Mathematics and Computer Science (ICMC)
University of São Paulo (USP) – São Carlos/SP – Brazil

²GIScience Research Group
Heidelberg University – Heidelberg, Germany

{luizffga,horita,jporto}@icmc.usp.br,
enrico.steiger@geog.uni-heidelberg.de, herfort@stud.uni-heidelberg.de

Abstract. Flood risk management requires updated and accurate information about the overall situation in vulnerable areas. Social media messages are considered to be as a valuable additional source of information to complement authoritative data (e.g. in situ sensor data). In some cases, these messages might also help to complement unsuitable or incomplete sensor data, and thus a more complete description of a phenomenon can be provided. Nevertheless, it remains a difficult matter to identify information that is significant and trustworthy. This is due to the huge volume of messages that are produced and which raises issues regarding their authenticity, confidentiality, trustworthiness, ownership and quality. In light of this, this paper adopts an approach for on-the-fly prioritization of social media messages that relies on sensor data (esp. water gauges). A proof-of-concept application of our approach is outlined by means of a hypothetical scenario, which uses social media messages from Twitter as well as sensor data collected through hydrological stations networks maintained by Pegelonline in Germany. The results have shown that our approach is able to prioritize social media messages and thus provide updated and accurate information for supporting tasks carried out by decision-makers in flood risk management.

1. Introduction

In the past few years, several floods have affected the communal life of a number of countries (e.g. Australia in 2010-11, Japan in 2011, and the Philippines in 2013) and have caused serious damage to all of them. Moreover, this has led to the inhabitants of these countries facing dire circumstances, resulting in a need for assistance at both a national and international level [Jha and Lamond 2012]. These facts show the need for carrying out relief activities by building forms of resistance against these disasters, i.e. enabling communities to resist, change or adapt, if a disaster occurs [Norris et al. 2008]. Flood risk management is an important adaptive measure for dealing with climate change, since it is expected that in future there will be an increase in the frequency of extreme weather conditions that can lead to flooding.
In this context, social media as a means of supplying Volunteered Geographic Information (VGI) [Goodchild 2007], have been regarded as a valuable source of information since it enables users to share texts, images and videos about vulnerable areas in real-time [Vieweg et al. 2010, Starbird et al. 2010, Herfort et al. 2014, Horita et al. 2013]. These data can increase our understanding of the overall situation in vulnerable areas [Schnebele et al. 2014a]. However, the large number of produced messages can hamper attempts to find the information that is most relevant and trustworthy [Yin et al. 2012, Middleton et al. 2014]. Problems regarding authenticity, confidentiality, ownership, reliability, and quality should also be taken into account since these factors play an important role during disasters [Agichtein et al. 2008, Kietzmann et al. 2011, Manca and Ranieri 2013].

In view of this, the use of real-time sensor data provided by stationary sensors (e.g. water level, volume of rainfall, and images) can assist in the prioritization of social media messages [Albuquerque et al. 2015]. The reason for this is that the combination of these data sources provides accurate and useful information for flood risk management [Horita et al. 2015]. However, most of the current approaches seek to extract useful information within social media in an offline manner without combining different data sources, which can be feasible for the applications that require real-time decisions and seek to improve all data sources limitations involved.

In light of this, this paper attempts to adopt an approach for the automatic prioritizing of social media messages based on sensor data to support near real-time decision-making for flood risk management. This is based on an analysis that examines the spatio-temporal characteristics of social media message and the closest hydrological station so that the appropriate prioritization can be achieved. The main strategies employed in this paper are as follows:

- Defining a workflow to combine official and social media data for floods;
- Supporting real-time prioritization of social media messages;
- Application of a framework to support near real time decision-making during floods;

The remainder of this paper is structured as follows. Section 2 discusses related works. Section 3 describes the approach, while Section 4 details the proof-of-concept. Finally, Section 5 outlines the conclusion and makes suggestions for future work.

2. Related Work

Owing to the growing number of geo-enabled devices, social media such as Facebook, Twitter, Instagram, and others have allowed users to share geo-located posts [Naaman 2011]. These posts might include references to events occurring at or affecting specific locations. As a result, social media has emerged as an alternative source to authoritative when there is a data crisis, because their users can communicate and share geographical data without any specialist knowledge [Herfort et al. 2014].

In addition, social media users share up-to-date situations, express their opinions, give emotional support and call for assistance by taking steps to coordinate and monitor the crisis [Qu et al. 2011]. In [Terpstra et al. 2012], the authors discuss the use and potential value of social media in disasters, by analyzing a storm event using monitoring and
analytical tools, focusing on aspects of the communication process, as well as identifying the questions and answers provided by the general public.

In [Zielinski et al. 2013], the authors obtain important information about public awareness of a situation during mass emergencies, by integrating a prototypical application to a decision-support component of a Tsunami early warning system. Their aim is to classify tweets during these events. In [Chae et al. 2012], the authors put forward an analytical and visualized approach involving abnormal topics and events within various social media data sources so that they can be checked and ranked probabilistically.

Other important benefits of using social media during a crisis are that spatial and temporal trends in Twitter can reach distant locations more quickly than physical events [Sakaki et al. 2010]. This is because, people located in the affected areas tend to share their dangers with each other while people in remote areas only share the secondary effects such as transport problems [Acar and Muraki 2011]. The (near) real-time aspect of social media is something that should be emphasized since this is really helpful when traditional telecommunication services have been destroyed [Yin et al. 2012].

However, when there is a crisis, the social media data raise issues regarding authenticity, privacy, trustworthiness and ownership [Tapia et al. 2011]. In addition, the reliability and quality of the information play a critical role during these events [Gupta and Kumaraguru 2012], and thus, advanced planning is really important [Morris et al. 2012]. It is also necessary to consider what means can be employed to distinguish relevant from irrelevant information, tools to extract meaning from the generally ill-structured messages and the means of distinguishing reliable from unreliable information [MacEachren et al. 2011]. Hence, misleading, outdated or inaccurate information from social media can hamper the efforts of the official agencies in these situations [Kongthon et al. 2012].

These are the reasons why authoritative data is still required. Several applications have been employed to combine authoritative and volunteered data. [Schnebele et al. 2014b] describe how volunteered data can be used to improve authoritative data and methods, based on a fusion of different layers from various data sources. These data sources can be differentiated by resolutions and levels of uncertainty, by identifying affected roads during a flood disaster and creating additional layers that taken together, can provide estimates for flood extent mapping. In [Zook et al. 2010], the authors describe a number of information technologies, such as web-based mapping services, used by volunteers to help the Haiti relief efforts by providing disaster management and response.

In [Vieweg et al. 2010], the authors identify and measure items of information regarding time- and safety-critical domains during critical situations so that they can understand better the individual and social contexts. In [Schade et al. 2013], the authors highlight the importance of Sensor Web Enablement (SWE) standards to streamline volunteered geo-referenced data which serve as stimuli, as well as a valuable and timely source for a framework of geographic data. This approach not only considers a single volunteered data source but also the means of combining them.

Although it is not yet possible to fully integrate and exploit in real-time the volunteered data within frameworks of geographic data, they should not be seen as separate
entities but as a complementary phenomenon [Mooney and Corcoran 2011]. For example, when a natural disaster strikes, it is possible to analyze the local area by filtering the keywords at the posts related to disasters, and hence, this can help to coordinate the degree of awareness of the situation with information about these local condition [Szomszor et al. 2011]. However, there is little variance of vocabulary used in crisis situations [Mendoza et al. 2010].

Another interesting point is that most of the applications do not focus on images and videos. In [Schnebele and Cervone 2013], the authors combine satellite images and topographic data with social media data for flood assessment. In [Triglav-Čekada and Radovan 2013], the authors describe how to gather images and videos from volunteers as a means of tracing the extent of the floods. They made public calls for volunteers to carry out and searches on the web for images. Despite of this, flooded areas may also be lost if official agencies only rely on these kind of measures.

In attempting to overcome the problem of a lack of important information, [Herfort et al. 2014] set out an approach that involves exploring the relations between spatial information from twitter messages and the knowledge obtained about flood phenomena both from hydrology and official sensor data. This is of value in prioritizing all the information involved but as this project was performed offline, there is still a lack of an automatic prioritizing method, which is the main goal of this project.

This real-time integration by means of automated methods is also valuable for improving communication during a crisis. In the first place, it can enable essential information to be conducted by plotting tweets over time in a crisis mapping tool to see the locations people are seeking and to show the location of the incident. In the second place, it can provide up-to-date information about the impact the incident has on those people and the people at other locations [Palen et al. 2010].

3. Geographic-based Approach

The amount of social media message is still increasing, which means that a good structural representation is necessary that should either focus on visualization techniques data or clusters of similar topics [Rogstadius et al. 2011]. During disasters, use is growing largely because of the spatial distribution of people with smart-phones in a high densification and the fact that sensor networks are hard to implement and deploy [Crooks et al. 2013].

Authoritative data can be provided by stationary and mobile sensors, which provide measurements of rainfall, water levels, weather patterns, images, and much more. Although they gather different types of data, they is have limitations to be considered. Stationary sensors only have one set position, which hamper the increase of the spatial resolution. Furthermore, mobile sensors are unable e.g. to extract information from images easily. Thus, the integration of social media and sensor data can either overcome their limitations by adding semantic information to the authoritative data, which is missing, as well as trustworthiness and quality to the social media data.

In our approach, priority is given to social media data based on sensor data in real-time. Since disasters are phenomena that are widespread, it is of crucial importance to get information about the spatio-temporal characteristics of these events. Once the extent of the event is known, it is possible to prioritize social media messages, e.g. twitter
messages, by using geographical relations, e.g. distance, by taking account of the Twitter data and the extent of the disaster [Herfort et al. 2014].

Official information about the extent of the disaster can vary in quality (e.g. temporal and spatial resolution) or it may not even exist at all. In view of this, our approach is designed in a way that is adapted to different levels of data availability. With regard to the flood phenomenon, account is taken of the workflow that corresponds to the different levels of data availability (see Figure 1).

![Figure 1. Workflow Approach](image)

The workflow can be divided into 2 parts. One related to the sensor data, and the other to the social media data. In the sensor data part, the workflow can be divided into three stages with regard to the availability of the data involved.

In the first stage, an attempt is made to determine the availability of highly qualitative information about the extent of the flood phenomenon. This could be information about flooded areas obtained from satellite images or mobile sensors. With the aid of this data, the prioritization-based distance can be calculated immediately without any further preprocessing. This calculation will offer the best degree of accuracy, even though the data might not be available in many cases.

In the second stage, when no information about flooded areas is directly available, data areas are required as an additional input from the user, or example, data containing potentially flood affected areas. This information could be derived from maps of flood...
risks, e.g. for flooding over a large period of flood.

Finally, sensor data will be used to decide whether these areas are affected by the flood, so that the calculation prioritization-based distance can be calculated. No additional information is required from the user in this case. Thus official elevation data and on-the-fly sensor data can be obtained, e.g. from gauging stations. As a result, an on-the-fly calculation is made of the flooded areas and the prioritization-based distance. This is the easiest way to prioritize social media messages, although it leads to the most inaccurate results, since further analysis of the sensor data should be analyzed to calculate existing flooded areas. The most accurate results are found when the flooded areas are initially available.

It should also be important mentioned that while the sensor data is gathered, data is acquired from the social media. In the social media part, the specific keywords of floods and other related words that can be found in the social media are used to find text messages referring to the flood event. Since keywords might change during a disaster, it was believed that the best approach was to save all the messages and filter them afterwards. Moreover, the possible temporal delay of disaster relevant information within social media should be considered. In this way, the filtering process can be adjusted to prevent the messages from being lost. Our infrastructure is generic since it is able to acquire, store and process social media data in a real-time manner. We also seek to provide a workflow for (near) real-time processing so that the collected data can be explored and visualized.

4. Proof of Concept

The implementation of the approach was divided into two phases: (1) definition of the catchments, (2) configuration of the acquisition of official sensor data and Twitter messages.

In the first phase, a shapefile containing 779 catchments1 (Figure 2) located in Germany was interpreted by means of geotools2. These catchments were then inserted into the database using a shp2pgsql function so that a shapefile could be translated into SQL operations. The PostgreSQL with a geographic extension (PostGIS) was used as a database. The identification (id) and respective area (multipolygon geometry type) of the catchments were stored in the database.

Once these catchments had been stored in the database, we were able to start the second phase. Sensor data were collected from Pegelonline3. This official agency publishes data provided by hydrological stations installed on the riverbed in Germany. They share these data by using web services, e.g. Rest API or Web Feature Service (WFS). In our case, use was made of simple Restful URLs calls via the Rest API4 for handling stations. Before including the sensor data in the database, the json file5 returned by these calls was decoded. Each station contains metadata such as identifier, name, coordinates, its respective catchment (which catchment the station is inside), and among

---

1 Catchment area is the geographic area generally founded either on formal local government boundaries or else on some other geographic basis, e.g. neighborhood or district of a city.
2 http://www.geotools.org
3 https://www.pegelonline.wsv.de/gast/start
4 http://www.pegelonline.wsv.de/webservice/guideRestapi
5 http://www.pegelonline.wsv.de/webservices/rest-api/v2/stations.json
other characteristics.

The individual measurements\(^6\) (e.g. date, value, and current water level) of each station contains timestamp, value and the current water level either with the lowest and highest average values. When a station returns a “high” water level, its catchment and the date of the measurement are stored in the database (table flooded_area) so that this information is used to define the period of a flood (date) and the affected area (catchment).

Figure 3 displays the data model of the database.

Simultaneously, tweets were collected through a real-time API streaming that requires an open persistent connection. Any parsing and filtering can be carried out before

\(^6\)http://www.pegelonline.wsv.de/webservices/rest-api/v2/stations/BONN/W/currentmeasurement.json
the tweets are stored in the database. In our case, the geofilter was based on a grid 5x5 of the catchments. Before storing each tweet in the database, the distance was calculated between the location of this tweet and every flooded area (the catchment that has a station with a “high” water level measurement). This calculation was made by using \( \text{ST\_DISTANCE}([\text{catchment area}], [\text{tweet location}]) \) operation of PostGIS, which returns the shortest cartesian distance between two geometries. Moreover, since this distance represents the prioritization of this tweet, it means that the closer a tweet is to a flooded area more the priority it has.

A hypothetical scenario, displayed in Figure 4, is used for analysis of the proposed approach. The first scenario represents the catchments, Pegelonline sensors and gathered tweets in Germany. The measurements in some stations showed a “high” water level which indicates that the catchments were flooded. After that, the distance between the tweet and all flooded areas was calculated for each arriving tweet. In this case, the shortest distance found between them was used to prioritize the tweet.

![Figure 4. Tweet Prioritization Scenarios](image)

In this diagram, the shortest distance was \( d5 \). This means that, if the shortest distance of another tweet from a flooded catchments were longer than \( d5 \), this tweet would have less priority, otherwise it would have a greater priority. It should also be taken into account that not all the tweets are useful for flood risk management because they could not be flood-related (e.g. tweets which do not contain any important keywords such as ”floods” or ”inundation”). This calculation need only be made only for flooded-related tweets although all of them should be stored in the database, since a filter within the database can have a better performance because there are some limitations to Twitter API Streaming limitations. Furthermore, some important keywords for floods might vary.

5. Conclusion

It has been argued that the use of social media to support flood risk management can be a valuable source of information. However, the large number of provided messages still

\(^7\)https://dev.twitter.com/streaming/overview
remains an obstacle to on-the-fly filtering of useful and accurate messages. This paper has thus adopted an approach that allows social media messages to be given priority by means of real-time authoritative data provided by in situ sensors. The closest distance between a message and a catchment area where the measurement of a sensor can record a high value, can be regarded as a parameter for this prioritization.

By simulating hypothetical scenarios, the implementation and analysis showed that the approach is able to prioritize social media messages automatically by using real-time sensor data. This prioritization might be an useful means of supplying accurate information, as well as complementing data provided by stationary sensors. Finally, it offers support for the estimates of the overall situation in vulnerable areas, and decision-making in flood risk management.

In the future studies, we intend to examine the quality of filtering and classify social media messages by using crowdsourcing such as Ushahidi or machine learning techniques e.g. for detecting the sensor data outliers so that floods can be predicted even before sensors have to measure a “high” water level. With regard to authoritative data, the combination of data from different agencies could improve time resolution (e.g. satellite images), and spatial distribution of station measurements. This could be achieved by using better APIs and thus being able to overcome the lack of information. Finally, an attempt will be made to assess the approach in different scenarios so that the results can be generalized to other disasters and social medias.

6. Acknowledgments
The authors would like to express thanks for the financial support provided by CAPES (Grant Pró-alertas no. 88887.091744/2014-01). Luiz F. F. G. Assis is grateful for the financial support from FAPESP (Grant no. 13/16202-1) and (14/01515-7). João Porto de Albuquerque is grateful for the financial support from CAPES (Grant no. 12065-13-7) and Heidelberg University (Excellence Initiative fund: 2300054, assignment: 7812617). Flávio E. A. Horita is grateful for the financial support from CNPq (Grant no. 202453/2014-6).

References


the 2011 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pages 320–323, Lyon, France.


