Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods

Luiz Fernando F. G. de Assis[1], **Benjamin Herfort**[2], **Enrico Steiger**[2], **Flávio E. A. Horita**[1], **João Porto de Albuquerque**[1,2]

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

[2]GIScience Research Group
Heidelberg University – Heidelberg, Germany

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

Abstract. *Social networks have been used to overcome the problem of incomplete official data, and provide a more detailed description of a disaster. However, the filtering of relevant messages on-the-fly remains challenging due to the large amount of misleading, outdated or inaccurate information. This paper presents an approach for the automated geographic prioritization of social networks messages for flood risk management based on sensor data streams. It was evaluated using data from Twitter and monitoring agencies of different countries. The results revealed that the proposed approach has a potential to identify valuable flood-related messages in near real-time.*

1. Introduction

The growing number of natural disasters, such as floods, has been leading for better preparation from vulnerable communities. In this sense, in-situ and mobile sensors are providing historical and updated information through the monitoring of environmental variables (e.g. temperature of the water or the volume of rainfall). Although these data are useful for supporting decision-making, further information is required for estimating the overall situation at an affected area [Horita et al. 2015]. Social networks like Twitter, Facebook, and Instagram, can overcome this issue either by providing information from areas which are not covered by sensors or complementing semantically the data provided by them [Starbird and Stamberger 2010, Vieweg et al. 2010, Zielinski et al. 2013, Horita et al. 2015].

Despite the fact that the combination of sensor data streams and social network messages might provide better information for supporting decision-making in critical situations like floods [Mooney and Corcoran 2011], it raises some challenges. On the one hand, the huge volume of messages shared

through social network makes difficult the identification of relevant messages, i.e. decision-makers do not want to analyse thousands of messages, they need the most valuable in order to make faster their decision-making [Vieweg et al. 2014]. On the other hand, the near real-time integration of sensor data and social network messages raises issues regarding the combination of distinct data streams (e.g. per second or per minute) and different data formats (e.g. numbers and texts) [Dolif et al. 2013].

In this context, this paper aims to present an approach for supporting flood risk management by means of the near real-time combination of social network messages and sensor data streams. It extends our previous works [Assis et al. 2015, Albuquerque et al. 2015] by adopting a workflow analysis which structures and defines an automated near real-time prioritization of social network messages based on the sensor data stream. Furthermore, it describes the formal representation of the problem statement, and makes an evaluation of the approach through case studies. In summary, the main contributions of this work are described below:

1. To define an approach to combine a sensor data stream with social network messages, aiming at identifying high value messages in near real-time for flood risk management;
2. To learn lessons from the application of the proposed approach in a real-world flood scenario in our application case study.

The remainder of this paper is structured as follows. Section 2 examines the related works. Section 3 sketches the background of the basic concepts and introduces the approach and methodology employed in this work. Section 4 describes the evaluation of the approach and their results are analyzed in Section 5. Finally, 6 draws conclusions and recommends some future works.

2. Related Works

Several applications have attempted to combine authoritative and non-authoritative data to improve the limitations of each other. Existing approaches integrate authoritative and non-authoritative data to provide location-based eventful visualization, statistical analysis and graphing capabilities in near real-time [Wan et al. 2014, Schnebele et al. 2014]. The combination of information provided by a collaborative platform (esp. Ushahidi) and sensor data collected via a wireless sensor network have been built for decision-making in flood risk management [Horita et al. 2015].

Several other studies aim at analyzing the large amount of information provided by social networks [Ediger et al. 2010, Gao et al. 2011], exploring e.g., relations between spatial information from both social network messages and knowledge about flood phenomena [Albuquerque et al. 2015]. An algorithm for monitoring social network messages (esp. tweets) and detecting

upcoming events is another eminent approach [Sakaki et al. 2010]. This algorithm classifies tweets using their keywords, number of words, and context. There are also systems for processing and analyzing social network messages in near real-time [Song and Kim 2013]. The results of its application to monitor Korean presidential elections showed that social network can support the detection and prediction in the changing of communities' behaviour. Finally, examining earliest social network messages that have produced a trend with the aim at identifying and creating a classification schema allows a categorization of messages, and thus a discovery of potential trends in near real-time [Zubiaga et al. 2015].

Although some studies have been done in the field, none of them tackles the combined use of social network messages and data collected from sensor streams in near real-time. In this manner, the processing of different data flows and data formats still pose challenges for the the use of sensor data as an alternative to support the filtering and extraction of high value social network messages in near real-time.

3. Problem Statement and Approach

3.1. Prioritization of Location-based Social Network Messages

Problem Statement. Due to the high volume of social network messages, the process of extracting relevant messages has been becoming more complex. This is because most of these messages are shared from several platforms (e.g. Flickr and Twitter) in distinct formats (e.g. photos and texts) with different data flows (e.g. per hour or per second).

Research Question. Is a sensor data stream able to support the near real-time identification of relevant social network messages in flood risk management?

Hypothesis. Given a set of catchments C, sensor data stream S and location-based social network messages M, we assume that the n-messages $M = \{m_1,...,m_n\}$ nearest to the m-flooded areas $F_{t_r} = \{f_1,...,f_m\}$ available at a time t_r tend to be more flood-related than the more distant (p-n)-messages $M = \{m_{n+1},...,m_p\}$, where n, m, p and $r \in \mathbb{N}$, and t is a timestamp. F is defined here as a time series of flooded areas. $F = \{F_{t_1}, ..., F_{t_r}\}$.

Definition 1. This paper uses a set of georeferenced catchments $C = \{c_1,...,c_n\} \subseteq R^2$. Each c_j denotes a two-dimensional Euclidean space that can either contain sensors or not. A sensor data stream $S = \{s_1,...,s_m\}$ contains a set of continuous sensor data s_k = [i, v, t, p, c]. Each sensor data has an id s_k.i, a water level value s_k.v at a timestamp s_k.t, a geographic position s_k.p = (x; y) and a s.c catchment. In case a sensor data s_k contains s_k.v equal to "high" at a timestamp s_k.t, and a s_k.p within a catchment c_j, this catchment c_j become a flooded area $f_p \in F_{s_k.t} \subseteq C$. The f_p is available until that s_k.v and any other sensor value contained in the catchment c_j are not "high" anymore at a subsequent timestamp to s_k.t.

$$F = \{F_{t_1}, ..., F_{t_r}\} \tag{1}$$

$$\begin{gathered} F_{t_r} = \{f_0, ..., f_p\} \mid \exists\ s_k\ \in\ S\ and\ f_p \in F_{t_r}, \\ where\ s_k.v = \text{“}high\text{”}\ and\ s_k.c = f_p\ and\ s_k.t = t_r. \end{gathered} \tag{2}$$

Location-based social network users U = {u_1,...,u_q} produce georeferenced messages represented by M = {m_1,...,m_n}. Each message m_i = [u, t, v, g] contains a value text m_i.v, as well as is produced by an user m_i.u at a timestamp m_i.t in a geographic location m_i.g. If there is a flooded area f_p, for each new incoming message m_i, a distance d is computed by the nearest neighbour (Euclidean) distance of the m_i.g position to every element of the set F_{t_r}) = $\{f_p\}_{p=0}^{n}$ at a timestamp t_r.

Definition 2. In general, the cartesian minimum distance between two points p (message location) and p’ (the nearest point contained in a flooded area) in a Euclidean space R^n is given by:

$$d(p, p') = \sqrt{\sum_{i,j=1}^{n} |p_i - p'_j|^2},\ where\ i,\ j \in \mathbb{N}. \tag{3}$$

3.2. Sensor Data Stream and Social Network Message Workflows

Given the extent of the flood phenomena, spatiotemporal characteristics of both sensor data stream and social network can be combined and more explored. Social networks messages can be used to complement sensor data stream with semantic information, while sensor data stream can add reliability to the social network messages. For this reason, this approach is designed to suit an on-the-fly prioritization for different levels of data availability that are require to ensure an effective flood risk management.

The proposed workflow is performed in a pipeline way that contains both a sensor data stream S and a social network messages stream M. In the sensor data stream part (see Figure 1), there are three possible kinds of data availability. The first alternative is to have highly qualitative information about the extent of the flood phenomena, which can be provided by Unmanned Aerial Vehicles (UAVs). Although they provide the best degree of accuracy for detecting events in near real-time, they are rarely gathered.

If no direct sensor data is able to show the existence of a flooded area f_p, thus it is analyzed if they can either provide a flood hazard area or not. Local data about the affected areas e.g., maps of risks can potentially provide this kind of information. In the last stage of the verification, if neither the flooded area f_p nor the flood hazard area are initially available, a digital elevation model (provided by an user) is used to calculate a flood hazard area. After calculating this flood hazard area, they are used as an input (along with

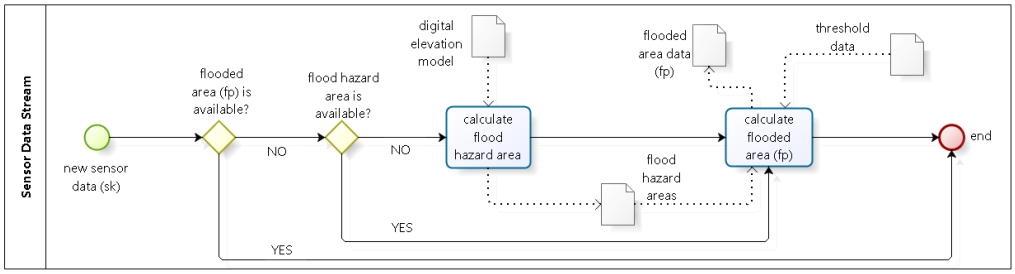

Figure 1. Sensor Data Stream Workflow.

threshold data) to calculate the flooded area f_p. The flooded area f_p in this part, is used to calculate the prioritization-based distance $P(m_j)$ when a new social network message m_j arrives (see Equation 4). In case more than one flooded area is available, the nearest flooded area distance is assign to the message prioritization.

$$p(m_j) = min(d(m_j, f_p)),\ where\ m_j \in M,\ f_p \in F\ and\ j,\ p \in \mathbb{N}. \tag{4}$$

Social network messages m_j and sensor data s_k should gathered simultaneously so that even delayed flood-related messages can be acquired. In the social network message stream part (see Figure 2), for each new incoming social network message m_j, prioritization-based distance is computed according to the nearest existing flooded area available produced by the sensor data stream part of the workflow. After calculating this prioritization-based distance, the messages are filtered to find messages that are likely to refer to a flood event. At first, our aim is to store all the messages since the specific keywords for floods might change during a flood. In this way, the filtering process can be easily adjusted without losing any flood-related messages.

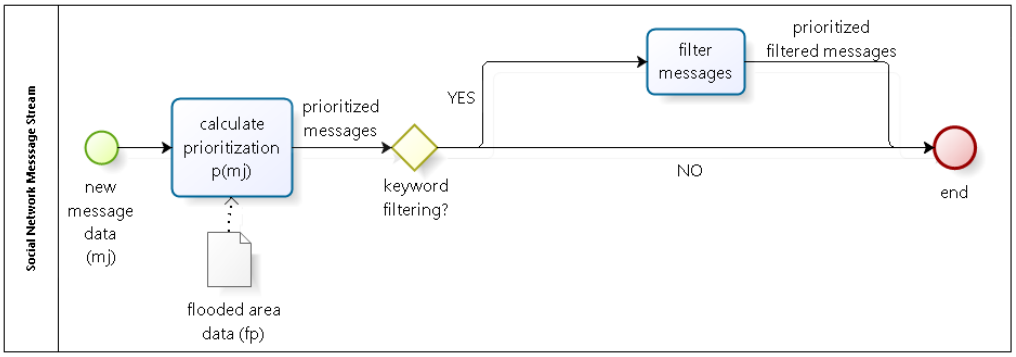

Figure 2. Social Network Message Workflow.

4. Case Studies and Experimental Setup

The approach was evaluated by means of an application case study of floods in Brazil, since it is currently taking measures to cope with and alleviate flood

situations. This set of measurements includes activities that involves pre-flood planning, managing emergency situations and post-flood recovery [Ahmad and Simonovic 2006].

Updated knowledge of river conditions plays an important role at supporting decision-making, since several technical factors can prevent this kind of information from being obtained. Countries such as Brazil where there are flash floods caused by heavy rain or the overflow of streams and narrow gullies needs this kind of management to mitigate the damage to the local infrastructure. This means that emergency agencies have to cope with the risk of human casualties and the extent of flood damage in their decision-making.

In this application case study, we considered as data source, authoritative static data (shapefiles), authoritative dynamic data (sensor data streams), and social network messages (Twitter).

4.1. Case Study: Flash Floods in Brazil

The analysis is confined to São Paulo as a single region within Brazil so that it is easier to compare and link the results of the case study. The shapefile of the State of São Paulo was produced using *geotools*[1] operations and geo-referenced data sets provided by HydroSHEDS[2]. It contains 315 small catchments. The sensor data streams was obtained from the national center for monitoring disasters and issuing warnings in Brazil (Cemaden). This agency operates by continuously installing new stations and providing their data through a Rest API.

In the State of São Paulo, Cemaden provided data from 465 stations. Each station and measurement of Cemaden are combined at the same file. The provided measurements from all the stations regarded the last four hours, considering an offset. This is the difference between the distance of the installation position and the real water level. As soon as it starts raining, the stations measure the rainfall every 10 minutes, otherwise they measure every 60 minutes. The floods in Brazil are represented at their most extreme by flash floods.

The catchments, stations and flooded areas are depicted in Figure 3, while a density map of the prioritized tweets is shown in Figure 4.

4.2. Experimental Setup

Runtime Environment: The implementation to retrieve Cemaden data was based on a simple Java toolkit for JSON[3], while tweets were retrieved using a Java library for Twitter API called Twitter4j [4]. Both of them were implemented in a pipelined fashion as a data stream. The experiments were run on a server

[1] http://www.geotools.org
[2] http://hydrosheds.cr.usgs.gov/index.php
[3] https://code.google.com/p/json-simple/
[4] http://twitter4j.org/en/index.html

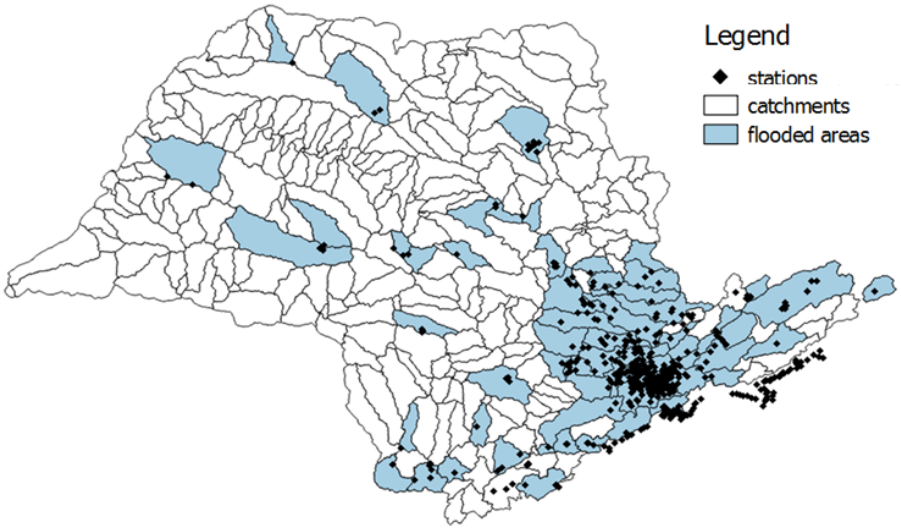

Figure 3. São Paulo - An analysis of Brazilian Stations and Catchments.

with 2GHz AMD Opteron(tm) Processor 4171 HE and 3.4 GB RAM memory running Ubuntu 12.04.5 LTS (64 bit).

Dataset: Twitter Streaming API and Cemaden Rest API were used to retrieve data in the period from April to May 2015.

5. Results and Analysis

In our case study, only 6% (68,195) of the tweets were prioritized mainly due to the flash floods. Such application case study helped to represent the scenarios when flood occur, since a large number of tweets were posted at critical moments. At first, all the flood-related tweets (403) were filtered by making a selection of the prioritized tweets (1,136,583) based on specific keywords and their synonyms. This "keyword selection" was based on the Brazilian words in the dictionary for "flood", taking into account differences of case sensitive letters without spelling mistakes. The Brazilian keywords were "enchente", "inundação" and "alagamento".

Table 1. Keyword-based filtering description of the location-based social network messages

# all the tweets	# prioritized tweets	# prioritized flood-related tweets	# prioritized non flood-related tweets
1,136,583 (100%)	68,195 (6%)	403 (0,04%)	67,792 (5.96%)

All the stations provided 284,663 measurements, which included 311 distinct stations that measured 1,030 high values. These values set for up to 59 distinct catchments areas that are considered to be flooded. A detailed description of the data provided by stations and their measurements is depicted in Table 2.

We also decided to calculate the time that our approach takes to prioritize all the tweets. The latency of the tweets was considered to be acceptable

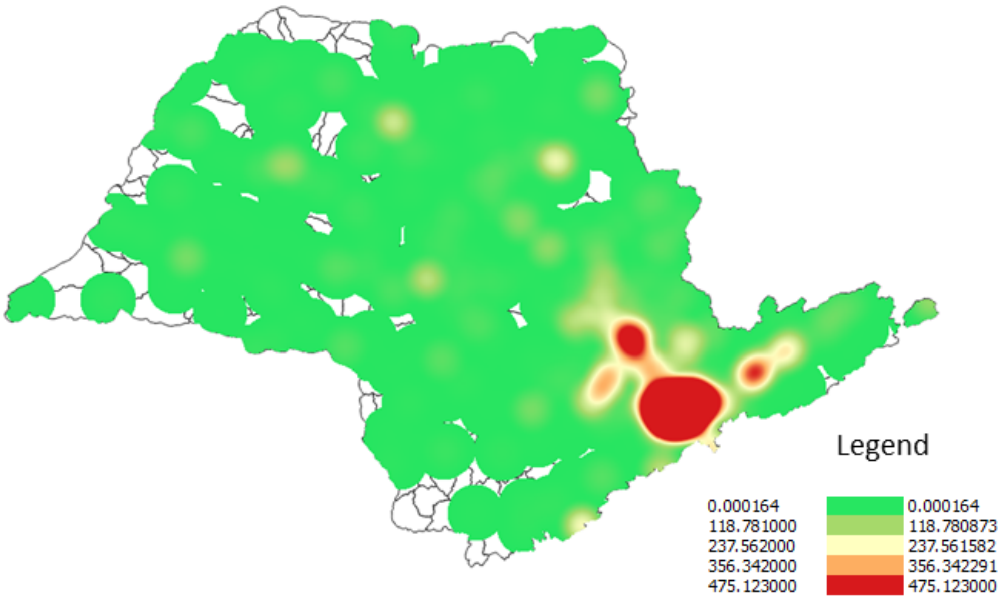

Figure 4. São Paulo - An analysis of Prioritized Tweets during periods of floods.

Table 2. Sensor measurement description of Cemaden stations.

# catchments (flooded areas)	# stations (high values)	# measurements	# all the measurements (high values)
59 (18.7%)	311 (66.9%)	284,663 (100%)	1,030 (0.0037%)

for the prioritization of social network messages for floods. This latency was the difference in time between the arrival from the Streaming API and the storage at the database. The latency was only calculated for the prioritized tweets. The processing average time per minute of all the prioritized tweets was less than one second.

Tweets containing relevant information presented not only flood-related text, but also georreferenced images that can help in flood risk management. As can be seen, exemplary prioritized tweets containing relevant messages are shown in Table 3 and Figure 5.

In addition, a statistical hypothesis test was conducted to check whether flood-related tweets are closer to hazard areas than those that are non- flood-related during floods. The Mann-Whitney U-test was employed for the samples of prioritized flood-related and non-flood related tweets. The test returns the p-value of a two-sided Wilcoxon rank sum test, which tests the null hypothesis that the distance of independent samples with different lengths from continuous distributions of flood-related and non-flood related tweets are with equal medians, against the alternative that they are not.

The p-value of 7.2940e-016 indicates a rejection of the null hypothesis of equal medians at a 5% significance level. That means, the sample of tweets which contain flood-related keywords are not equally nearer to the hazard ar-

Table 3. Examples of prioritized Tweets containing flood-related keywords without images (On Topic, Relevant).

Flood-Related Prioritized Tweets	Translation
“tá td alagado aq na marquês”	“everything is flooded here in the Marquês” (name of a place or avenue name)
“Ta alagado aqui”	“It is flooded here” (the georeference of the tweet can supplement this information)
“Nações alagada”	The “Nações (Avenue Name) is flooded”
“Alagamento na av dos Tajuras (at @AG2 Nurun in São Paulo, SP)” https://t.co/r7ECeAtiFB”	“There is a flood in Tajuras avenue” (the georeference of the tweet can suplement this information)
“@VCnoSPTV chuva de 30 minutos e alagamento na região do Brás, pra variar http://t.co/wWVqIKtz3z"	“@VCnoSPTV (Twitter account of a TV program) it has been raining for 30 minutes and the region of Bras is flooded, as always http://t.co/wWVqIKtz3z”
“5min de chuva e rua já fica alagada"	“After 5 minutes of rain, the street is already flooded” (the georeference of the tweet can supplement this information)

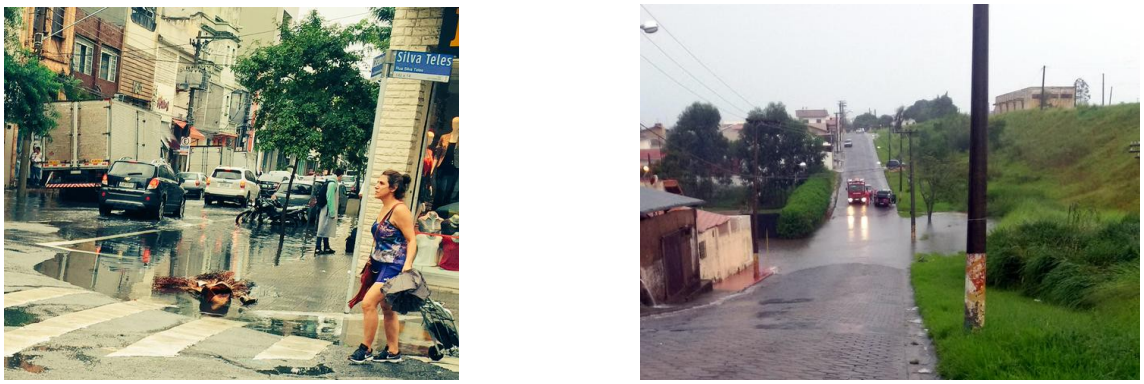

Figure 5. Examples of flood-related tweets containing images that can help in flood risk management.

eas to the ones that not contain flood-related keywords. The median distance of the sample of non- flood-related tweets was 10,905 meters away from those areas, while the sample of flood-related tweets was 3,027 meters away. Figure 6 shows the two distribution of non flood-related and flood-related tweets based on their prioritization (distance in km to flooded areas).

6. Conclusion and Discussions

This paper presents an approach for supporting flood risk management by means of a near real-time prioritization of social network messages based on sensor data streams. One case study was used for evaluating the approach. The results confirmed that the geographical relations are useful for prioritizing social network messages related to floods. They showed that there are about 3,6 times more flood-related social network messages near to flood-affected areas than non-flood-related messages. Although our approach was evaluated in a specific context of floods and using Twitter messages, it can be used to other types of disasters (e.g. droughts and landslide) and social network (e.g. Instagram and Flickr), i.e. considering images or videos instead of only texts messages.

Our approach gathered the messages per minute during a flood at an

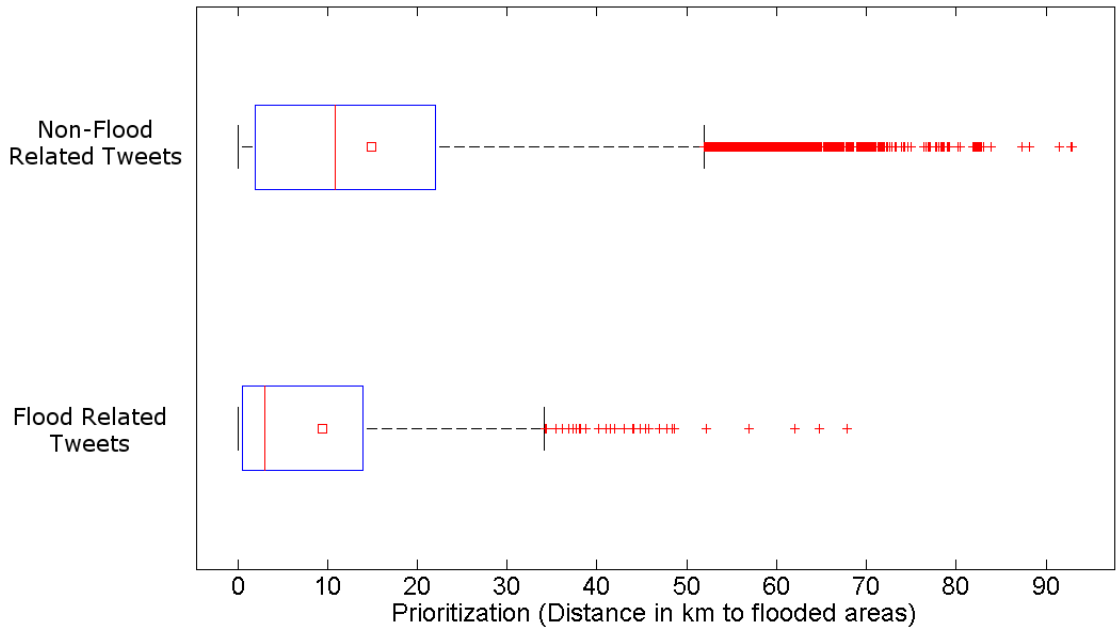

Figure 6. Median, Average and Outliers of flood-related and non flood-related tweets.

average processing time of less than one second. Given the large number of messages (time peaks should be treated as critical periods since more messages tend to be posted), such processing time to prioritize does not significantly change. This work has shown that social networks messages and sensor data streams can complement each other. Sensor data streams are accurate, dynamic, heterogeneous and continuous, although they are scarce and hard to implement and maintain. On the other hand, social network messages can enhance semantic sensor data, but their large number is not easy to handle since they can be misleading, outdated or inaccurate. Despite the lack of user experience and knowledge, social networks have been used in crisis management revealing their remarkable and positive features.

Although most of the existing approaches are still insufficient for near real-time decision-making since they fail to take note of the fact that data in disasters should be analyzed on-the-fly and automatically. Our approach searches for georeferenced social network messages using a grid 5x5 bounding box based on the catchments dimension. Although most of the messages are not flood-related (and do not contain any important keywords such as “floods” or “inundation”), they were stored at the database after first being filtered because a keyword search is arbitrary, especially for near real-time event detection.

All the social network messages located within a flooded area were prioritized with zero meters “0 m” as distance, which is the main value-based prioritization. Some of the prioritized messages have images embedded in them, which were really useful when they were geolocated because they could show

the exact situation of a particular place and sometimes helped more than simply by the words. During our analysis, a few of the total amount of available messages were both georeferenced and considered to be flood-related.

Furthermore, heavy rains might affect the connection infrastructure (e.g. cellphone services or wi-fi), which in turn may reflect on the unavailability of information sharing. Although this issue is important when dealing with social network messages, it is beyond the scope of this work. In this sense, a better time resolution and spatial distribution of the sensor measurements would improve the availability of information provided by sensors. In situations that sensors are measuring high values all the time, machine learning techniques would be an useful way to check whether the sensors are really in a flood situation or only measuring high values all the time because of its position on the river.

Future work lines should take account of using the prioritization of social network messages as one step to further filtering and classifying the quality of crowdsourcing. Besides that, it can serve as basis to improve machine learning models that consider geographical links.

References

Ahmad, S. and Simonovic, S. P. (2006). An Intelligent Decision Support System for Management of Floods. *Water Resources Management*, 20(3):391–410.

Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4):667–689.

Assis, L. F. F. G., Herfort, B., Steiger, E., Horita, F. E. A., and ao Porto Albuquerque, J. (2015). A geographic approach for on-the-fly prioritization of social-media messages towards improving flood risk management. In *Proceedings of the 4th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 1–12.

Dolif, G., Engelbrecht, A., Jatobá, A., da Silva, A. J. D., Gomes, J. O., Borges, M. R., Nobre, C. A., and de Carvalho, P. V. R. (2013). Resilience and brittleness in the alerta rio system: a field study about the decision-making of forecasters. *Natural hazards*, 65(3):1831–1847.

Ediger, D., Jiang, K., Riedy, J., Bader, D., Corley, C., Farber, R., and Reynolds, W. (2010). Massive social network analysis: Mining twitter for social good. In *Proceedings of the 39th International Conference on Parallel Processing (ICPP)*, pages 583–593.

Gao, H., Barbier, G., and Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14.

Horita, F. E., Albuquerque, J. P., Degrossi, L. C., Mendiondo, E. M., and Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. *Computers & Geosciences*, 80:84–94.

Mooney, P. and Corcoran, P. (2011). Can Volunteered Geographic Information be a participant in eEnvironment and SDI? In *Environmental Software Systems. Frameworks of eEnvironment*, pages 115–122.

Sakaki, T., Okazaki, M., and 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*, pages 851–860.

Schnebele, E., Cervone, G., Kumar, S., and Waters, N. (2014). Real time estimation of the calgary floods using limited remote sensing data. *Water*, 6(2):381–398.

Song, M. and Kim, M. C. (2013). Rtˆ2m: Real-time twitter trend mining system. In *Proceedings of the 2013 International Conference on Social Intelligence and Technology*, pages 64–71.

Starbird, K. and Stamberger, J. (2010). Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting.

Vieweg, S., Castillo, C., and Imran, M. (2014). Integrating social media communications into the rapid assessment of sudden onset disasters. *Social Informatics*, 8851:444–461.

Vieweg, S., Hughes, A. L., Starbird, K., and 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*, pages 1079–1088. ACM.

Wan, Z., Hong, Y., Khan, S., Gourley, J., Flamig, Z., Kirschbaum, D., and Tang, G. (2014). A cloud-based global flood disaster community cyberinfrastructure: development and demonstration. *Environmental Modelling & Software*, 58:86–94.

Zielinski, A., Middleton, S. E., Tokarchuk, L., and Wang, X. (2013). Social media text mining and network analysis for decision support in natural crisis management. *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 840–845.

Zubiaga, A., Spina, D., Martínez, R., and Fresno, V. (2015). Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473.