Exploiting Geolocation, User and Temporal Information for Monitoring Natural Hazards on Twitter

Uso de Información de Geolocalización, Usuario y Temporal para la Monitorización de Desastres Naturales en Twitter

Víctor Fresno
UNED
E.T.S.I. Informática
vfresno@lsi.uned.es

Arkaitz Zubiaga
Uni. of Warwick
Coventry, UK
arkaitz@zubiaga.org

Heng Ji
Rensselaer Pol. Institute
Troy, NY, USA
jih@rpi.edu

Raquel Martínez
UNED
E.T.S.I. Informática
raquel@lsi.uned.es

Resumen: Cuando se producen eventos relacionados con situaciones de emergencia, es importante acceder a tanta información como sea posible relacionada con dicho evento. En este contexto algunas redes sociales como Twitter suponen un importante recurso de información en tiempo real. La técnicas clásicas de filtrado de información suelen centrarse en el análisis de coocurrencia de términos con el conjunto de palabras clave inicialmente consideradas. Sin embargo, estas aproximaciones pueden perder información, ya que no son capaces de recuperar información relevante que venga expresada con palabras que no coocurran con las palabras clave inicialmente usadas, y que expresan nuestra necesidad de información. Considerar información de geolocalización, usuario o temporal dentro de un enfoque de pseudo-relevance feedback, nos permite encontrar terminología relacionada con el evento, pero no coocurrente con las palabras clave inicialmente consideradas. Por otro lado, considerando el aspecto temporal se puede modificar una función de expansión de consultas como la divergencia de Kullback-Leibler con el fin de mejorar el filtrado de información en estas situaciones de emergencia. Nuestras propuestas se han evaluado en dos colecciones de eventos del mundo real obteniéndose resultados alentadores.

Palabras clave: Recuperación de Información, Realimentación por relevancia, Análisis de Redes Sociales en Tiempo Real, Twitter, Seguimiento de Desastres Naturales

Abstract: During emergency situation events it is important to acquire as much information about the event as possible, and social media sites like Twitter offer important real-time user contributed data. Typical Information Filtering techniques are keyword-based approaches or focused on co-occurrence with keywords. However, these approaches can miss relevant local information if messages do not contain an initially considered event-related keyword. Considering geolocation, user and temporal information within a pseudo-relevance feedback approach we can find event-related terminology but not co-occurring with initially considered keywords. Thus, taking into account the temporal aspect we can modify a query expansion function like Kullback-Leibler divergence in order to improve the Information Filtering process. Our proposed approaches have been evaluated in two Twitter datasets associated with real-world events, obtaining encouraging results.

Keywords: Information Retrieval, Pseudo-Relevance Feedback, Real-Time Social Media Analysis, Twitter, Natural Hazards Monitoring

1 Introduction

Recent years have seen the explosive growth of the social volume of information. Social media sites like Twitter aggregate a large volume of real-time user contributed data for a wide variety of events (Zubiaga et al., 2011). These events range from popular and widely known pre-scheduled events, to unexpected natural hazards, e.g., earthquakes, hurricanes, etc. In the case of emergency situations, it is important to retrieve as much information as possible about the event to make sure that no relevant information is missed. This can be helpful for humanitarian aid workers to assist citizens effectively, for the relief ac-

tivity management during such events, and for the people to stay abreast of the latest details. Therefore, the fact that social media plays an important role in monitoring the information shared by users in these situations, and the importance of avoiding to miss out relevant information, emphasize the need for developing effective Information Filtering techniques to carry out this monitoring process in the best possible way.

The most common and widely used approach to deal with this kind of Information Filtering tasks is the use of keyword-filtering techniques, which tend to obtain high precision but low recall values, i.e., tweets containing specific keywords that have been manually crafted by a user will most probably be relevant, but it is very likely that other relevant tweets will not contain these keywords. To facilitate tracking event-related contents, users on Twitter usually come to an agreement during emergency situations to use a common hashtag, which can help others follow relevant content related to these events; still, some local information will often be missed. Additional approaches that can help improve this computationally include Latent Semantic Indexing (Deerwester et al., 1990) or Topic Modeling (Steyvers and Griffiths, 2007), used as content-based filtering techniques that can help improve recall values; however, in these cases, the discovery of new relevant keywords is restricted to keywords that co-occur in the tweets that contain the keywords in the user-defined query. Other tweets which contain neither the initial keywords nor the co-occurring ones will never be retrieved using these techniques.

To the end of enhancing this process of monitoring natural hazards on Twitter, we delve into the use of three additional features of tweets, namely "user", "geolocation" and "temporal" information, which we rely on to discover new keywords which are related to the natural hazard. The contribution of our paper revolves around this idea, for which we set forth the following hypotheses:

- If a user posted a tweet message in Twitter about a natural hazard (e.g., a hurricane) in an affected area, we can expect that their immediately previous or later messages will be related with the event, irrespective of these tweets containing or not any of the initially considered hurricane-related keywords.
- If we find a tweet about a natural hazard in a specific geolocation and time, we can expect that tweets within a nearby geolocation and posted at the same time will be also related to the event, the nearest messages around this geolocation and at that time will be also related with it, irrespective of these tweets containing or not any of the initially considered hurricane-related keywords.

We introduce a new preliminary approach for harnessing social information in order to acquire as much information about a natural hazard as possible, and beyond initially considered event-related keywords. We define the Information Filtering problem as a pseudo-relevance feedback task, and propose a query expansion method using the geolocation, user and temporal information inherent to tweets. We incorporate new messages, where the initially event-related keywords are not necessarily used, to the initial set of most relevant documents. Thus, we introduce a modified Kullback-Leibler divergence as a query expansion function that considers the temporal aspect of tweets.

These approaches have been evaluated using two Twitter datasets associated with real-world events: "Hurricane Isaac" in late-August 2012 near Baton Rouge (Louisiana, US); and "Hurricane Sandy" when it affected New York City on October 29th, 2012. The results presented in this paper prove the effectiveness of the proposed approach, motivating further study of the exploitation of this kind of social information.

2 Related Work

Since its creation Twitter has become an important source of information for coverage of crisis events (Imran et al., 2014). For example, the events unfolding during the Sichuan earthquake were first reported by Twitter users. Similarly, the first report that a plane landed in the Hudson river in New York in 2009 was posted on Twitter by an eyewitness. In (Mills et al., 2009), an early study on emergency events, the authors found that Twitter had a great impact in distributing crisis-related information. Twitter was crucial during events such as the Californian fires, New England Ice Storm, Gulf of Mexico Hurricane, Cyclone Nagris in Myanmar

and Mississippi Hurricane (Sinnappan et al., 2010). In (De Longueville et al., 2009) an analysis of tweets related to a fire near Marseille was carried out, and it was shown that Twitter updates about the event were generally well synchronized to temporal and spatial dynamics of the event itself. People access Twitter during crisis events to complement information they obtain from traditional sources ((Sorenson and Sorenson, 2006); (Shklovski et al., 2008); (Sutton et al., 2008)) and it is increasingly being considered as a primary source to learn what is happening on the ground (Palen et al., 2010).

Hashtag use is the main mechanism for accessing information in order to design solutions in emergencies (Starbird et al., 2010), and Twitter activity and the nature of its use during emergencies have been subject of both mass media attention and academic research ((Sorenson and Sorenson, 2006); (Starbird et al., 2010); (Hughes, A.L. and Palen, L., 2009); (Vieweg et al., 2010)). In order to sense and analyze disaster information from social media, microblogs as a source of social data have recently attracted attention; combining messages is helpful for understanding the impact of an event. From the point of view of event management, and considering geolocation information, Twitter as a social sensor has attracted much attention [(Takahashi et al., 2010), (Vieweg et al., 2010), (Sinnappan et al., 2010)]. All these approaches consider geolocation information. In (Sasaki et al., 2012) the authors consider each Twitter user as a sensor, and tackle an event detection task based on sensory observations. They use Kalman filtering and particle filtering to estimate the locations of an event and they developed an earthquake reporting system that shows the tweets relating to the earthquake on a map. Their system is based on the difference between the number of tweets posted while the event is occurring, and while not. Several map-based systems have been developed on the web to share local knowledge. In these systems, users can enter local safety/hazard incident information on a map (Shinohara et al., 2011). Recently, in (Schnebele and Cervone, 2013) a methodology for the generation of flood hazard maps is presented fusing remote sensing and volunteered geographical data.

To the best of our knowledge, there is no research exploiting geolocation, user and temporal information to the end of detecting new related terminology, which do not necessarily co-occur with event-related keywords. This is the main contribution of this work.

3 Query Expansion for Information Filtering using Social Information

In this paper, we propose an Information Filtering approach based on a boolean IR model (Hiemstra, 2009). The reason for not ranking tweets is that we assume all tweets containing event-related keywords will be surely relevant. We apply a query expansion approach for selecting new event-related terms by means of social information.

Pseudo-Relevance Feedback (PRF) via query expansion (Xu and Croft, 1996) has been proven to be effective in many Information Retrieval (IR) tasks. (Carpineto et al., 2001) presented a method for term scoring for PRF, where the authors tried to maximize the divergence between the probability distributions of the terms estimated in the pseudo relevance set (p_{PR}) and the distribution estimated over the whole collection (p_C). They used the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951) calculated as in Equation 3.1 given that it captures the relative entropy between both distributions. To build the expanded query they selected the terms (w) that contribute most to the divergence of both distributions (i.e., higher KLD score).

$$KLD(w) = p_{PR}(w).log\frac{p_{PR}(w)}{p_C(w)} \quad (1)$$

3.1 Increasing the Pseudo Relevant Set

We re-examine the PRF assumption considering the social information stated in the initial hypotheses. Then, we consider the following tweets in order to extend the Pseudo Relevance set (PR) and extended Pseudo Relevance set ($PR*$), and then we apply the PRF process to PR^*:

- The immediately previous ($user_{pre}$) or later ($user_{post}$) tweets from a user that posted a message containing an initial keyword. Both approaches were also combined and tested.

- The messages posted from the nearest geolocations where a tweet containing a initial keyword was found, and considering a radius of 0.1 degrees in latitude and longitude ($geo_{0,1}$). These tweets should be constrained in a reasonably short timeframe (depending on collection). We tried different radius values, but we empirically found that the best results were achieved by using 0.1.
- Tweets containing a hashtag caught within a message where a initial keyword was found.

Hence, our proposal consists in extending the initial PR set to other PR^* set adding tweets from an user, geolocation and time information.

$$KLD^*(w) = \sum_{w \in V} p_{PR^*}(w).log\frac{p_{PR^*}(w)}{p_C(w)} \quad (2)$$

3.2 Temporal_KLD: Modifying KLD for Adding Twitter Temporal Aspect

In this paper we introduce Temporal-KLD (TKLD), a modification of Kullback-Leibler divergence for considering the Twitter temporal aspect within the PRF process. The main idea is to combine the query expansion process applied on PR and PR^* sets with tweets posted in a short timeframe, corresponding to the time when the natural hazard is happening, in order to find wider event-related terminology. It is important to remark that evaluation datasets were collected in a specific geolocation and time (where and when the natural hazard was hitting), as it will be seen in Section 4. We expect to find relevant information within the timeframe when event is happening. We look for into phrases not sharing event-related previously found terminology.

The inherent idea is to combine two PRF processes, one considering the PR or PR^* sets, and the other one considering a tweet set corresponding to the short period of time when hazard is hitting ($p_{Time}(w)$). Then, we propose:

$$TKLD(w) = p_{PR}(w).log\frac{p_{PR}(w)}{p_C(w)}.log\frac{p_{Time}(w)}{p_C(w)} \quad (3)$$

4 Datasets

We collected two Twitter datasets associated with real-world events: "hurricane Isaac" and "hurricane Sandy". These were two hurricanes that hit different parts of the United States in 2012.

For the "**hurricane Isaac**" dataset, we sampled 1,000 geolocated tweets posted from the Lousiana area in August 29^{th}. The selected timeframe was from 8 a.m to 9 a.m, and the specific geolocation was $\pm 2^{\circ}$ latitude and longitude degrees from the hurricane eye according to information of the US National Hurricane Center[1]. We also downloaded tweets from August 19^{th} to August 28^{th} (in the same timeframe and area) to consider background for the estimation of divergences, trying to capture the everyday vocabulary.

For the "**hurricane Sandy**" dataset, we sampled 1,000 geolocated tweets sent from NYC area in October 29^{th}. The selected timeframe was from 8:30 p.m to 8:35 p.m., and the specific geolocation was +-2^{o} from the hurricane eye according to information of the US National Hurricane Center. We also downloaded tweets for background from June 15^{th} to July 15^{th}.

In both cases, the preprocessing was the same: Porter stemming and removing stopword and tokens beginning with numbers or not-alphanumeric characters. With regard to gold-standard creation, tweets from both of the datasets were annotated as being either "related" or "non-related" to the event in question, based on the following guidelines. A tweet is "related" if:

- It explicitly refers to the hurricane.
- It refers to consequences of the hurricane (e.g., power outage).
- It is aware of the hurricane, and would not have been posted otherwise (e.g., concerned about safety of friends)

If it does not provide evidence to be considered as related, it should be categorized as non-related.

Each tweet was annotated via Amazon Mechanical Turk[2] by five US-based users. After the annotation process, 321 tweets were

[1] http://www.nhc.noaa.gov

[2] https://www.mturk.com/mturk/

annotated as related and 679 as non-related in the "hurricane Isaac" dataset, and 318 and 682 as related and non-related tweets in the "hurricane Sandy"dataset.

A Recall-Precision graph is used as a combined evaluation measure.The area under the curve is used as a simple metric to define how an algorithm performs over the whole space. Such a graph, given an arbitrary recall point, tells us the corresponding precision value. At this point it is important to remark that we are interested in increasing recall values, maintaining reasonable precision values. For this reason, we must pay attention to the right upper regions of the Figures.

5 Results

Figures 1 and 2 show the performance that can be achieved following the different query expansion approaches in a Recall-Precision space, and for the first 50 expansion terms. The initial queries are "hurricane OR isaac" and "hurricane OR sandy" and afterwards, a boolean IR model is applied to obtain all tweets containing the query terms and for creating the PR set, that will be used in the later PRF process.

Both Figures 1 and 2 show that, in the case of the baseline, when the PR set only contains tweets with original queries, precision values decrease (increasing recall values) in higher levels than when any social information approach is taken into account (approaches using PR^*), and thus the baseline area under the curve is smaller than when social information is considered. This shows that the use of social information leads to better performance than that achieved by using classic query expansion approaches.

With regard to the query expansion function, the behavior of KLD and TKLD is similar for the first expanded terms in both datasets. However, these positions represent low recall values, far from our objective of increasing the initial recall values (0.49 in "hurricane Isaac" and 0.44 in "hurricane Sandy"). In general, we can observe that TKLD expansion function obtains the best results in both datasets. Considering social information in PRF process we can achieve almost 0.75 in terms of both in precision and Recall, while not using temporal information the performance values are around 0.7 both in terms of precision and recall. While it is true that selection of the first expansion terms by TKLD function in "hurricane Sandy" dataset is worse than using KLD, these positions represent low recall values (around 0.55-0.63). The subsequent expansion term selection is able to maintain precision values in a slight decrease while recall values are increasing (up to 0.8-0.85), as it can be seen in Figure 2.

Analyzing the different social information in both datasets, we do not observe a similar performance. While for the "hurricane Isaac" dataset the geolocation information obtains encouraging results, the user information only works with high values of expanded terms (around $40 - 50$). In addition, in this dataset the hashtag approach does not work, since its results are similar to TKLD using PR. Similar conclusions cannot be drawn for the hurricane Sandy dataset, which occurred in New York City. In this case, the geolocation information decreases the recall values achieved with TKLD using PR; and the user and hashtag information do not contribute since their recall values are similar to the TKLD ones using PR. Their contrasting population densities and different social composition of these populations could be one reason. We think that initial hypotheses, related to geolocation and user, fit better in a small city like Baton Rouge and not in a high density urban environment like NYC.

In summary, using temporal information, natural hazard-related terminology can be found beyond considering co-occurrences with event-related keywords. Nevertheless, for considering user, geolocation and hashtag information, a deeper study must be carried out. There is another important aspect as well. When we expand with user, geolocation or hashtag information, and do not obtain new related terminology, in most of the cases (with the exception of geolocation information in NYC) it does not decrease the Recall-Precision values because the relative entropy between the added tweets and the background is low. The reason is that the terms contained in those added tweets are terms with similar relative frequency in the background, and then the terms are not selected in the first positions by query expansion function.

6 Conclusions

In this paper we have introduced a new approach for information filtering in the context of Twitter coverage of emergency events. We

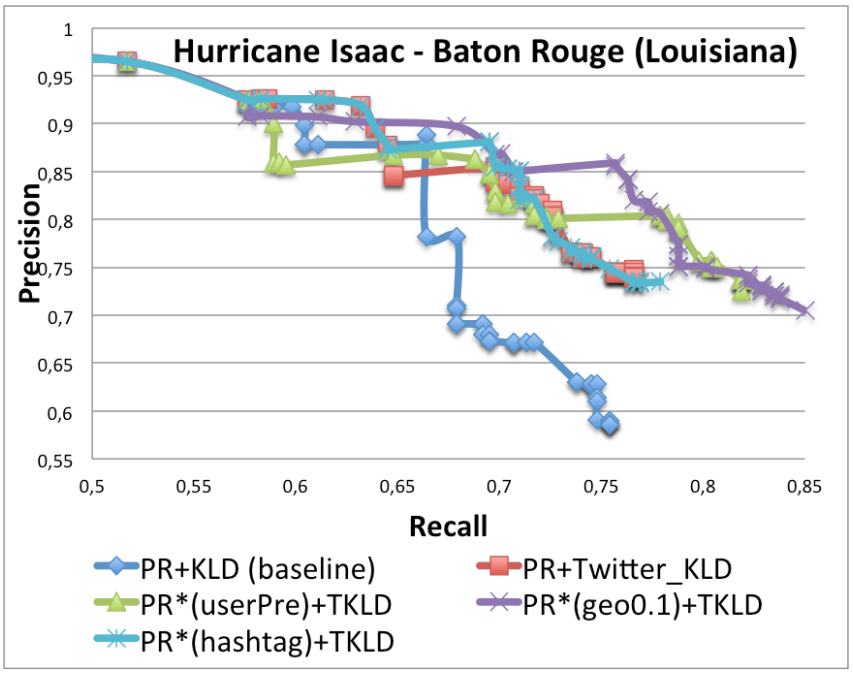

Figura 1: Recall-Precision curve for "Hurricane Isaac" dataset.

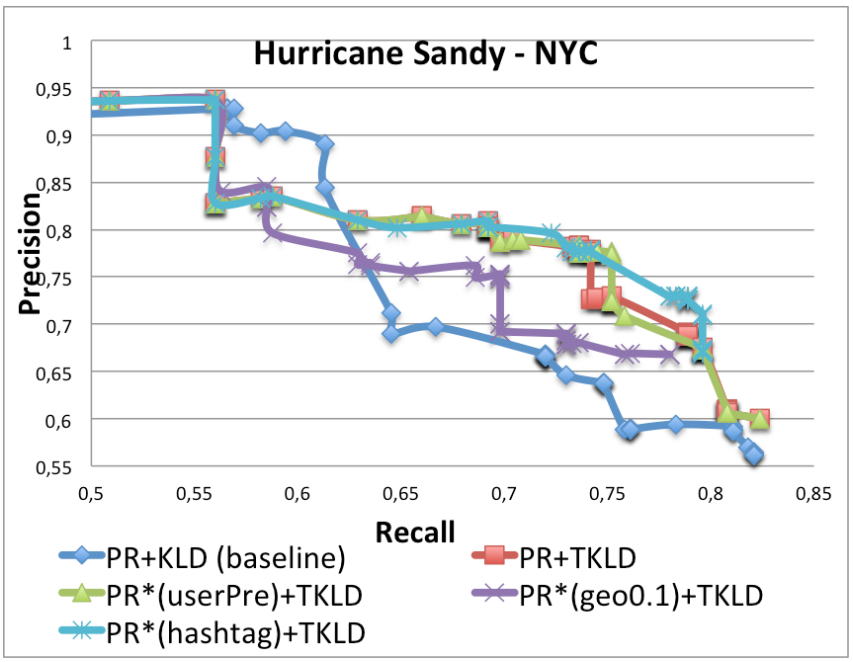

Figura 2: Recall-Precision curve for "Hurricane Sandy" dataset.

have proposed a set of novel approaches that rely on non-textual social features to capture new keywords that are related to an event. The approaches we introduce and experiment in this paper rely on geolocation, user and hashtag information, as well as temporal information in a Pseudo-Relevance Feedback via query expansion approach. Through experiments on two hand-labeled datasets associated with two natural hazards, our preliminary research shows that especially the use of temporal information can have a significant impact in the performance, improving recall values. Moreover, our results suggest that the study that the use of social information for query expansion so as to discover

new keywords related to an event can help boost the performance of the tweet retrieval.

Our plans for future work include a further exploration of the social features inherent in tweets to improve the tweet retrieval. Also, and especially motivated by the fact that previous studies found that the use of Twitter during emergencies is different than its use in other contexts, we would like to explore the applicability of our approaches to other types of events. This study, accompanied by an iterative refinement of the filtering techniques, will allow us to come up with a more generalizable approach.

Acknowledgments

This work has been part-funded by the Spanish Ministry of Science and Innovation (MED-RECORD Project, TIN2013-46616-C2-2-R) and by UNED Project (2012V/PUNED/0004). This research was also partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Bibliografía

Carpineto, C.,de Mori, R., Romano, G. and Bigi, B. 2001. *An information-theoretic approach to automatic query expansion.* Journal ACM Transactions on Information Systems (TOIS), Vol. 19 (1), pp. 1–27.

Scott Deerwester and Susan T. Dumais and George W. Furnas and Thomas K. Landauer and Richard Harshman 1990. *Indexing by latent semantic analysis.* Journal of the American Society for Information Science. Vol 41(6): 391–407

Bertrand De Longueville, B., Smith, R.S. and Luraschi, G. 2009. *OMG, from here, I can see the flames!.* Proceedings of the 2009 International Workshop on Location Based Social Networks, p. 73-80.

Djoerd Hiemstra. 2009. *Information Retrieval Models.* Ayse Goker and John Davies (eds.), Information Retrieval: Searching in the 21st Century, Wiley.

Hughes, A.L. and Palen, L. 2009. *Twitter Adoption and Use in Mass Convergence and Emergency Events.* Proceedings of the Information Systems for Crisis Response and Management.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. 2014. *Processing Social Media Messages in Mass Emergency: A Survey.* arXiv preprint arXiv:1407.7071

Kullback, S., and Leibler, R. A. 1951 *On information and sufficiency.* The Annals of Mathematical Statistics.

A. Mills,A., Chen, R., Lee, J. and Rao, H. R. 2009. *Web 2.0 Emergency Applications: How Useful can Twitter be for Emergency Response.* Journal of Information Privacy and Security, vol. 5, pp. 3-26, 2009.

Palen, L., K. Anderson, G. Mark, J. Martin, D. Sicker, and D. Grunwald. 2010. *A Vision for Technology-Mediated Public Participation and Assistance in Mass Emergencies and Disasters.* International Academic Research Conference, 14 - 16 April 2010, The University of Edinburgh.

Sasaki, K., Nagano, S., Ueno, K. and Cho, K. 2012. *Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor.* Sixth International AAAI Conference on Weblogs and Social Media. Workshop on When the City Meets the Citizen, AAAI Technical Report WS-12-04

Schnebele, E. and Cervone, G. 2013. *Improving remote sensing flood assessment using volunteered geographical data.* Natural Hazards Earth Systems Science.

S. Sinnappan, S., Farrell, C. and Stewart, E. 2010. *Priceless Tweets! A Study on Twitter Messages Posted During Crisis: Black Saturday.* Proceedings of Information Systems: Defining and Establishing a High Impact Discipline, 21st International Conference on Information Systems (ACIS 2010), Brisbane, Australia, 01-03 December 2010, paper no. 39.

Shinohara, M., Hattori, A., Ioroi, S., Tanaka, H., Hayami, H., Fujioka H. and Harada Y. 2011. *Design and Trial of a Cellphone-based Hazard Information Sharing*

System for Residents Living Close to an Incident. 2011 Fifth International Conference on Next Generation Mobile Applications, Services and Technologies.

Shklovski, I., Palen, L., and Sutton, J. 2008. *Finding Community Through Information and Communication Technology in Disaster Events.* Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work (CSCW 2008), November, San Diego, pp. 127-136.

Sorenson, J.H. and Sorenson, B.V. 2006. *Community Processes: Warning and Evacuation.* In: H. Rodriguez, E. L., Quarantelli and R. R. Dynes (Eds.). Handbook of Disaster Research, pp. 183–199. New York: Springer.

Kate Starbird , Leysia Palen, Amanda L. Hughes and Sarah Vieweg. 2010. *Chatter on the red: what hazards threat reveals about the social life of microblogged information.* Proceedings of the 2010 ACM conference on Computer supported cooperative work, pp. 241–250.

Steyvers, Mark and Griffiths, Tom 2007. *Probabilistic Topic Models.* In McNamara, D; Dennis, S.; Kintsch, W. Handbook of Latent Semantic Analysis. Psychology Press.

Sutton, J., Palen, L., Shklovki, I. 2008. *Backchannels on the Front Lines: Emergent Use of Social Media in the 2007 Southern California Fires.* Proceedings of the 2008 Information Systems for Crisis Response and Management Conference (ISCRAM 2008), Washington, D.C., pp. 624-631.

Tetsuro Takahashi, Makoto Okazaki and Yutaka Matsuo. 2010. *Earthquake shakes Twitter users: real-time event detection by social sensors.* WWW '10 Proceedings of the 19th international conference on World wide web, pp. 851–860.

Tetsuro Takahashi, Shuya Abe and Nobuyuki Igata. 2011. *Can twitter be an alternative of real-world sensors?.* HCII'11 Proceedings of the 14th international conference on Human-computer interaction: towards mobile and intelligent interaction environments, Vol 3: 240–249.

Sarah Vieweg, Amanda L. Hughes, Kate Starbird and Leysia Palen. 2010. *Microblogging during two natural hazards events: what twitter may contribute to situational awareness.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1079–1088.

Xu, J. and Croft, W. B. 1996. *Query expansion using local and global document analysis.* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR).

Zubiaga, A., Spina, D., Fresno, V., and Martínez, R. 2011. *Classifying trending topics: a typology of conversation triggers on twitter* Proceedings of the 20th ACM international conference on Information and knowledge management.