Natural Disaster Database Design and Development for Himalaya Using Social Media

Kiran Zahra
Department of Geography, UZH
Winterthurerstr. 190, 8057 Zurich, Switzerland
kiran.zahra@geo.uzh.ch

Abstract

Ever increasing population of social media and increasing trend of sharing information related to several different topics has the great potential to use this information for different purposes such as disaster management. The use of Volunteered Geographic Information (VGI) has its own challenges because of the unstructured nature of social media such as extracting information out of noise and credibility related issues. This project focusses on the core issues of using user generated content for disaster management in developing countries.
Keywords: VGI, Twitter, Credibility, toponym granularity.

1 Motivation and Project Objectives

Frequency and impact of natural disasters is on the rise worldwide (Kellenberg & Mobarak, 2008 ; Whybark, 2007). For more than fifty years, researchers have focused on a series of vital questions such as human occupancy of hazard zones, response of people and societies, and how to lessen the effects of environmental hazards (Cutter, 2012). A well-structured database about natural disasters can better answer these questions and can be an efficient source of analysing causes, effects, capacity, vulnerability and further planning of disaster management such as EM-DAT[1].

The Himalayas are the youngest mountain range in the world. This mountain range is still geologically active (Lavé & Avouac, 2000), which causes a high occurrence of natural disasters in this region. Documenting natural disasters is important in order to understand current issues and future predictions. Although this kind of information is available, it is technically incomplete and difficult to maintain.

People share a bulk of information on social media sites such as Twitter during mass emergencies (Hossmann et al., 2011 ; Terpstra et al., 2012). Information shared on social media is not only usable, but also very valuable in terms of its contents (Kavanaugh & Yang, 2011). In case of disaster in a region the most important information for disaster response organizations are location, timing and impact (Jongman et al., 2015). Social media has the potential to serve in all these aspects.

The increasing trend of using smart phones and accessibility to fast internet connections are the main causes of rapid growth of social media population (Kwak et al., 2010). Social media users such as Twitter community tend to share bulk of tweets publicly available and these tweets have great potential to extract meaningful information and use this information during emergency events. But it is often very difficult to filter only "useful information" from the Twitter stream by making keyword based queries and thus it is recommended to use machine leaning algorithms to filter information (Sakaki, Okazaki, & Matsuo, 2013).

Credibility of information shared on social media is another challenge to using this type of data during disasters. Mendoza, Poblete, & Castillo (2010) state in their research that rumours tend to be questioned more than authentic news and propose a framework to deal with credibility issues.

[1] D. Guha-Sapir, R. Below, Ph. Hoyois - EM-DAT: International Disaster Database – www.emdat.be – Université Catholique de Louvain – Brussels – Belgium

Gelernter & Balaji (2013) discuss the increased use of geographical content in tweets during disasters. This geographic location information can be used to manage rescue operations during and after disasters and also for generating maps. Although disaster relief is very important issue, but no research has done on utilizing VGI to design and create natural disaster database.

Overall aim of this study is to analyse different aspects of disaster related information shared on Twitter about any disaster happening in Himalayan region and its surrounding countries. This social media data has the potential to contain helpful geographic information and situational updates during mass emergency events. These can be used by disaster response agencies for efficient management. This research will help to investigate the issues brought about by natural hazard phenomena.

The objectives of this research are to:

- Apply machine learning algorithms to resolve keyword ambiguities and optimize tweet classification
- Assess credibility of tweets based on data quality standards of tweets shared during disasters
- Analyse the granularity of geographic features reported in tweets during disasters, geoparsing and geocoding those geographic features to display on the map for visual representations
- Conceptualize and implement a VGI based natural disaster database

2 State of the art

Sakaki, Okazaki, & Matsuo (2013) discussed 'semantic analysis of tweets' in their research. They did this analysis to detect target events from tweets. They also mentioned that identifying keywords alone from tweets about disasters is not enough to gather valuable information about the event. Verma et al. (2011) studied tweets 'contributing to situational awareness' and tweets 'not contributing to situational awareness'. They used examples of tweets discussing the disaster but not contributing to situational awareness, as they do not give any actionable information.

Acar and Muraki (2011) argue that it is often very difficult to filter correct information out of the Twitter data stream. Uncontrollable informal retweets question the certainty and reliability of Twitter in disaster situations because during a disaster credibility of information is very important. Semantic analysis of tweets and keyword ambiguities have very little or no research about tweets originating from Himalayan region and its surrounding countries, moreover the role of geographic features has never been studied in Naïve Bayes classification efficiency.

Twitter streaming Application Program Interface (API) downloads many fields called metadata of tweet with tweet text field. This metadata can be used to effectively assess tweet credibility (Canini, Suh, & Pirolli, 2011). Credibility assessment issues are expected to be more complex due to cultural and language diversity in the Himalayan region.

When there is a disaster, precise location identification is a very important element in many situations. Disaster/emergency response organizations can only respond if they know precise locations. People tend to add location information when they share something about disasters on social media. Lingad, Karimi, & Yin (2013) discuss a tool "Named Entity Recognizers". They state that location extraction from microblogs is becoming more accurate and that such tools have great potential to find out locational information more correctly.

Social media's application to disaster management is getting more and more important because of its real-time nature such as TweetTracker (Kumar et al., 2011), Artificial Intelligence for Disaster Response (AIDR) (Imran et al., 2014), Twitcident (Abel et al., 2012), ScatterBlogs for situational awareness (Thom et al., 2015).

Since forecasting disaster is difficult, efficient disaster management is very important to improve disaster response. There are many national and international organizations maintaining records of natural disasters such as EM-DAT, a famous global natural and technological disaster database. Such databases are created and maintained by data collected from different primary and secondary data sources. Witham (2005) discussed in his research various issues about completeness and accuracy of maintaining databases about disasters. He also presented a new database about volcanic disasters and analyse this information for risk management.

The role of social media as a great potential source of data during disasters has already been discussed, this research focusses on using information shared on Twitter to design a VGI based database. Designing and developing VGI based databases is a novel concept for the Himalayan region. This research gap formulated the following research questions:

RQ 1: Which data mining techniques are successful in identifying relevant tweets originating from developing countries?

RQ 2: Does information shared on social media meet quality indicators for user generated content (UGC) to use in case of disasters in the Himalaya region?

RQ 3: What is geographic feature granularity in UGC during disasters?

RQ 4: What are the potential contributions of social media data in designing and developing VGI based disaster database for the Himalaya region?

3 Methods

Twitter offers free, real-time data (tweets) downloads through its streaming API. This API requires certain parameters to capture tweets such as particular keywords, tweets sent from particular users, or tweets originating from a particular region. For this research, an application has been designed in R to capture real-time tweets without any pause, based on disaster-related keywords such as flood, earthquake, and landslide. Starting from 25.4.2015 more than 25 Million tweets based on disaster-related keywords have been downloaded so far. The supervised Naïve Bayes topic model is used to classify tweets in two classes: "information" and "not information".

The Twitter streaming API downloads tweets with their metadata. In addition to the tweet text field, forty-one associated fields are also downloaded. Many researchers have identified different sets of features to assess the credibility of tweets. Features such as the "number of followers, retweet count, number of URL's, verified, time, user description, status count, [and] length of a tweet" (Canini, Suh, & Pirolli, 2011; Morris et al., 2012) will be used to assess the credibility of tweets in the Himalayan region.

A natural disaster event will be selected as a case study to analyse the granularity of geographic features reported during the event. These geographic features then later be geoparsed and geocoded by available online platforms such as address-parser and gisgraphy. Geoparsing is the process of automatically identifying geographical information (names of places) from text (Gelernter & Balaji, 2013). Geocoding is the process of converting text into geographical coordinates (Ratcliffe, 2004). The spatial information identified from the tweets will be displayed on the map. This map is expected to reveal complex locations patterns about social media user's way of reporting locations during disasters.

Tweet text with its metadata will be analysed to identify potential attributes and entities to design an Entity Relationship Diagram (ERD). This analysis will be performed on adequate subset of acquired data. Designed ERD will be used for development of the database.

4 Conclusion

This research project has great potential to be implemented by disaster response agencies within the Himalayan region and also by international organizations by adapting during and post-disaster framework. The Himalayan region is important in determining world climate, prone to frequent hazards, and is generally less accessible due to a lack of facilities and resources. This research can be a good platform to minimize the impact of disasters in this region. This research project enriches the Geographic Information Science methods and concepts by focusing on unique spatio-temporal phenomena of natural disasters.

5 References

- Acar, Adam and Muraki, Y. (2011). Twitter for crisis communication : lessons learned from Japan' s tsunami disaster. 7(3), 392–402.
- Adam, N. R., Shafiq, B., & Staffin, R. (2012). Spatial computing and social media in the context of disaster management. IEEE Intelligent Systems, 27(6), 90-96.
- Bilham, R., Gaur, V., & Molnar, P. (2001). Pollen Tubes Labeled With Green Fluorescent Protein Show That the Pollen

Tube Also Pen- Etrates an, (August), 1442–1444.

- Canini, K. R., Suh, B., & Pirolli, P. L. (2011). Finding Credible Information Sources in Social Networks Based on Content and Social Structure. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 1–8. http://doi.org/10.1109/PASSAT/SocialCom.2011.91
- Clay Whybark, D. (2007). Issues in managing disaster relief inventories. *International Journal of Production Economics*, *108*(1–2), 228–235. http://doi.org/10.1016/j.ijpe.2006.12.012
- Cutter, Sussan L., (2012). *Hazards, vulnerability and environmental justice.* New York, NY: Taylor and Francis
- Earle, P. S., Bowden, D. C., & Guy, M. (2011). Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6), 708–715. http://doi.org/10.4401/ag-5364
- Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, *17*(4), 635–667. http://doi.org/10.1007/s10707-012-0173-8
- Hossmann, T., Carta, P., Schatzmann, D., Legendre, F., Gunningberg, P., & Rohner, C. (2011). Twitter in Disaster Mode: http://doi.org/10.1145/2079360.2079367
- Jongman, B., Wagemaker, J., Romero, B., & de Perez, E. (2015). Early Flood Detection for Rapid Humanitarian Response: Harnessing Near Real-Time Satellite and Twitter Signals. *ISPRS International Journal of Geo-Information*, *4*(4), 2246–2266. http://doi.org/10.3390/ijgi4042246
- Kattelmann, R. (2003). Glacial Lake Outburst Floods in the Nepal Himalaya: A Manageable Hazard? Retrieved December 21, 2015, from http://download.springer.com/static/pdf/938/art%3A10.1023%2FA%3A1021130101283.pdf?originUrl=http://link.springer.com/article/10.1023/A:1021130101283&token2=exp=1450710999~acl=/static/pdf/938/art%253A10.1023%252FA%253A1021
- Kavanaugh, A., & Yang, S. (2011). Microblogging in crisis situations :, 1–6.
- Kellenberg, D. K., & Mobarak, A. M. (2008). Does rising income increase or decrease damage risk from natural disasters? *Journal of Urban Economics*, *63*(3), 788–802. http://doi.org/10.1016/j.jue.2007.05.003
- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter , a Social Network or a News Media? *The International World Wide Web Conference Committee (IW3C2)*, 1–10. http://doi.org/10.1145/1772690.1772751
- Lavé, J., & Avouac, J. P. (2000). Active folding of fluvial terraces across the Siwaliks Hills, Himalayas of central Nepal. *Journal of Geophysical Research*, *105*(B3), 5735. http://doi.org/10.1029/1999JB900292
- Lingad, J., Karimi, S., & Yin, J. (2013). Location extraction from disaster-related microblogs. *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, 1017–1020. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84892685783&partnerID=tZOtx3y1
- Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter Under Crisis : Can we trust what we RT ?, 71–79.
- Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17. http://doi.org/10.1109/MIS.2013.126
- Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012). Tweeting is believing? *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12*, 441–450. http://doi.org/10.1145/2145204.2145274
- Murthy, D. (2013). NEW MEDIA AND NATURAL DISASTERS: Blogs and the 2004 Indian Ocean tsunami. *Information Communication and Society*, *16*(7), 1176–

1192. http://doi.org/10.1080/1369118X.2011.611815

- Murthy, D., & Longwell, S. A. (2013). Twitter and Disasters. *Information, Communication & Society*, *16*(6), 837–855. http://doi.org/10.1080/1369118X.2012.696123
- Nasikhin, & Adriani, M. (2007). Location Identification for the Geographic information Retrieval.
- Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June), 181–189. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-80053272732&partnerID=tZOtx3y1
- Poblete, B., Castillo, C., & Mendoza, M. (2011). Information Credibility on Twitter, 45–59.
- Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, *18*(January), 61–72. http://doi.org/10.1080/13658810310001596076
- Richardson, S. D., & Reynolds, J. M. (2000). An overview of glacial hazards in the Himalayas. *Quaternary International*, *65–66*, 31–47. http://doi.org/10.1016/S1040-6182(99)00035-X
- Sadler, J. C., & Ramage, C. S. (1976). The Role of Mountains in the South Asian Monsoon Circulation. *Journal of the Atmospheric Sciences*. http://doi.org/10.1175/1520-0469(1976)033<2255:COROMI>2.0.CO;2
- Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 919–931. http://doi.org/10.1109/TKDE.2012.29
- Singh, A. (2010). Climate Change and Disasters in the Hindu Kush Himalayan Region. Article, New Delhi, India: Solution Exchange. Retrieved from http://amritapeoplesvoice.blogspot.com
- Terpstra, T., Stronkman, R., Vries, a De, & Paradies, G. L. (2012). Towards a realtime Twitter analysis during crises for operational crisis management. *Proceedings of ISCRAM 2012*, (April), 1–9.
- Thayyen, R. J., & Gergan, J. T. (2010). Role of glaciers in watershed hydrology: a preliminary study of a "Himalayan catchment." *The Cryosphere*, *4*(1), 115–128. http://doi.org/10.5194/tc-4-115-2010
- Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., … Anderson, K. M. (2011). Natural Language Processing to the Rescue ?: Extracting " Situational Awareness " Tweets During Mass Emergency.
- Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 1079. http://doi.org/10.1145/1753326.1753486
- Viviroli, D., & Weingartner, R. (2004). The hydrological significance of mountains: from regional to global scale. *Hydrology and Earth System Sciences*, *8*(6), 1017–1030. http://doi.org/10.5194/hess-8-1017-2004
- Witham, C. S. (2005). Volcanic disasters and incidents: A new database. *Journal of Volcanology and Geothermal Research*, *148*(3–4), 191–233. http://doi.org/10.1016/j.jvolgeores.2005.04.017
- Yamada, T., & Sharma, C. (1993). Glacier lakes and outburst floods in the Nepal Himalaya. *IAHS Publications-Publications of …*, (218), 319–330. Retrieved from

http://ks360352.kimsufi.com/redbooks/a218/iahs_218_0319.pdf

- Imran, M., Castillo, C., Lucas, J., Meier, P., et al. (2014) AIDR: Artificial intelligence for disaster response. In: *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. [Online]. 2014 pp. 159–162. Available from: doi:10.1145/2567948.2577034.
- Kumar, S., Barbier, G., Ali Abbasi, M.A. & Liu, H. (2011) TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In: *Fifth International AAAI Conference on Weblogs and Social Media*. 2011 pp. 661–662.
- Thom, D., Kruger, R., Ertl, T., Bechstedt, U., et al. (2015) Can twitter really save your life? A case study of visual social media analytics for situation awareness. In: *2015 IEEE Pacific Visualization Symposium (PacificVis)*. [Online]. 2015 pp. 183–190. Available from: doi:10.1109/PACIFICVIS.2015.7156376.
- Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., et al. (2012) Twitcident: Fighting fire with information from social web streams. In: *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. [Online]. 2012 p. 305. Available from: doi:10.1145/2187980.2188035.