Extracting Geospatial Information from Social Media Data for Hazard Mitigation, Typhoon Hato as Case Study

Jibo Xie
Institute of Remote Sensing and Digital Earth, No.9 Dengzhuang South Rd, Haidian District, Beijing 100094, China
xiejb@radi.ac.cn

Tengfei Yang
Institute of Remote Sensing and Digital Earth;University of Chinese Academy of Sciences, No.9 Dengzhuang South Rd, Haidian District,Beijing 100094;Beijing 100049, China
yangtf@radi.ac.cn

Guoqing Li
Institute of Remote Sensing and Digital Earth, No.9 Dengzhuang South Rd, Haidian District, Beijing 100094, China
ligq@radi.ac.cn

Abstract

With social media widely used for interpersonal communication, it has served as one important channel for information creation and propagation especially during hazard events. Users of social media in hazard-affected area can capture and upload hazard information more timely by portable and internet-connected electric devices such as smart phones or tablet computers equipped with (Global Positioning System) GPS devices and cameras. The information from social media(e.g. Twitter, facebook, sina-weibo, WebChat, etc.) contains a lot of hazard related information including texts, pictures, and videos. Most important thing is that a fair proportion of these crowd-sourcing information is valuable for the geospatial analysis in Geographic information system (GIS) during the hazard mitigation process. The geospatial information (position of observer, hazard-affected region, status of damages, etc) can be acquired and extracted from social media data. And hazard related information could also be used as the GIS attributes. But social media data obtained from crowd-sourcing is quite complex and fragmented on format or semantics. In this paper, we introduced the method how to acquire and extract fine-grained hazard damage geospatial information. According to the need of hazard relief, we classified the extracted information into eleven hazard loss categories and we also analyzed the public's sentiment to the hazard. The 2017 typhoon "Hato" was selected as the case study to test the method introduced.

2012 ACM Subject Classification Human-centered computing → Social media, Information systems → Geographic information systems, Computing methodologies → Information extraction, Human-centered computing → Geographic visualization

Keywords and phrases Social media, hazard mitigation, GIS, information extraction, typhoon

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.65

Category Short Paper

Funding The national key research and development program of China (2016YFE0122600).

Acknowledgements We want to thank Edward T.-H. Chu from National Yunlin University of Science and Technology in the collaboration project.

1 Introduction

Social media has been widely used in our daily information creation and propagation especially in the hazard scenario. Users of social media in hazard effaced area can acquire real-time firsthand in-situ observation data and share these data by messages, short texts, pictures, or videos. And a fair proportion of these crowd-sourcing data includes geospatial information which is valuable for geospatial analysis of Geographic information system (GIS) during the hazard mitigation process. The geo-location information of social media data plays an important role in emergency detection and quick response [3]. The useful geospatial information including position, geospatial distribution, location clustering, and status of damages related with hazard, is hidden in the large number of social media data. Unlike conventional spatiotemporal data, social media data is dynamic, massive, unevenly distributed in space and time, noisy, incomplete, biased in terms of population, and represented in stream of unstructured media (e.g. texts and photos), which pose fundamental challenges for representation and computation to conventional spatio-temporal analysis [1]. Many researchers in GIS study area have noticed the importance of social media as an important source of geospatial information. In the past few years, geospatial information created by volunteers and facilitated by social networks has become a promising data source in time-critical situations [5]. And the concept of volunteering of geographic information(VGI) [4] has been introduced. While the quantity and real-time availability of VGI make it a valuable resource for disaster management applications, data volume, as well as its unstructured, heterogeneous nature, make the effective use of VGI challenging [2]. user-generated data can provide unique and highly useful information in several contexts (e.g. brand communication, market research, political communication as well as in extreme events) [6]. The social media GIS enables disaster information provided by local residents and governments to be mashed up on a GIS base map, and for the information to be classified and provided to support the utilization of the information by local residents [8]. In our study, we introduced the method to extract geospatial information and use these information to get the hazard loss categories and map the hazard-affected area. The Typhoon, a yearly happened hazard events in northwest Pacific, was selected as case study.

2 Methods

The key step of extracting hazard related geospatial information from social media data is how to understand the meanings of messages and texts. And the Natural Language Processing (NLP) is a common method for social media information extraction. In our study, we proposed a Social Media based Hazard information Recognition and Classification (SHRC) model for hazard related geospatial information extraction and analysis based on the NLP method. The workflow of the model (As shown in Fig. 1). The key steps of the SHRC model are as followed.

1. Event-driven hazard information acquisition from social media: Social media platform usually provides the interface or API for developers to retrieve and get social media data by using time-span, location and event related key words.
2. Data cleaning and store: Many messages from social media are repeated or not related, so we need to clean and filter the redundancy. After that the data are stored in the database for further analysis.
3. Definition of hazard loss categories: To evaluate the hazard loss, we proposed a hazard loss classification method of eleven categories including loss of life, interruption of water supply, building damage, business influence, forestry loss, traffic congestion, vehicle

Figure 1 Flowchar of extracting hazard geospatial information from social media data.

damage, power supply outage, electricity equipment broken, Communication Interrupt, infrastructure damage.

4. Creation of classification knowledge base based on feature words and lexicon: The first step is to extract some feature words from the sample micro- blog text of different disaster loss categories based on Chinese grammar rules and constructed the pairs of feature words collocation. The word vector model and existing lexicon is used to supplement and expand these pairs of feature words collocation. And the external natural language corpus is used to optimize the semantic collocation relationship between feature words.
5. Hazard information interpretation and extraction: The topics of social media messages are usually random and we use the hazard classification knowledge base in the step 4) for different types of hazard damage information extraction. A Chinese language processing and information retrieval toolkit, NLPIR (`http://ictclas.nlpir.org/downloads`), is deployed for word segmentation and part-of-speech tagging(POS). Then corresponding lexicon is used to match feature words for disaster loss information classification and sentiment analysis based on the knowledge base.
6. Sentiment analysis: The model uses sentiment words for sentiment analysis. The basic sentiment words from the text base on Chinese sentiment word table from "HowNet" (`http://www.keenage.com/html/c_index.html`). And the model extends the basic sentiment words by using the feature words from social media. There are three kinds of emotions, positive, neutral, and negative.
7. Spatio-temporal visualization: The hazard information from social media is geo-located by GPS position, address match by user's position, and Identification of place names from text. Then we can use the geospatial information and hazard loss attributions for visualization and mapping of the hazard-affected area.
8. Evaluation and validation: Three parameters, precision, recall and F-Measure (F1), serve as the evaluation indexes to evaluate the experimental results.

Figure 2 Typhoon "Hato" landed on the coast of Zhuhai city at 12:50, 23rd August, 2017. Map from `http://typhoon.zjwater.gov.cn`.

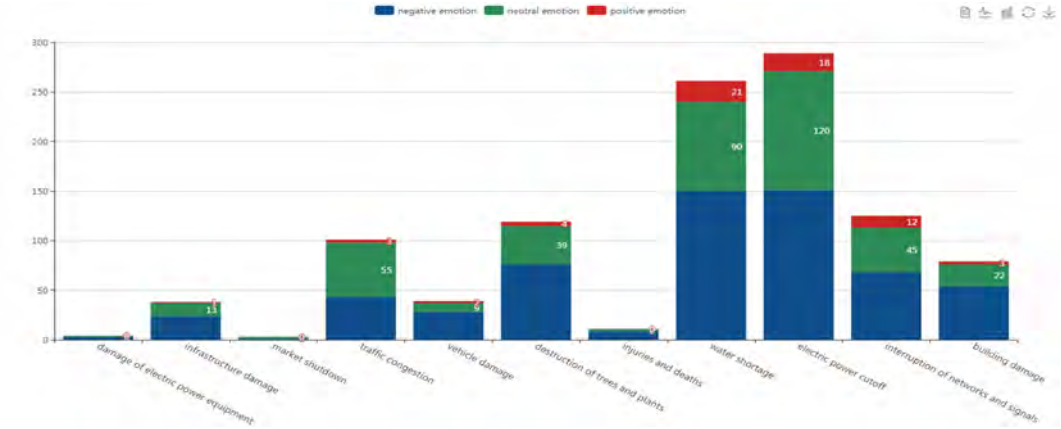

Figure 3 Sentiment analysis of hazard damage information extracted from social media.The blue, green, and red colors refer to negative, neutral, and positive emotion, respectively.

3 Case study

We used the dataset of 2017 typhoon events (about 20,000 records) [7] to train the model proposed in the paper. And typhoon "Hato" event landed on the coast of Zhuhai city at 12:50 23rd August, 2017 was selected as the case study to test the effectiveness of the model. The moving track of this typhoon is shown in Fig. 2. We selected 1600 records of the hazard related information from "sina-weibo" (`https://weibo.com/`) after cleaning and filtering redundant and irrelevant information with time span from 0:00 to 23:00 of the typhoon landfall day.

The statistics of the hazard information extraction and classification from "sina-weibo" are shown in Fig. 3. We can see that the numbers of power outage and interruption of water supply were the biggest, 289 and 261 respectively. According to the statistics, we can conclude that outage of power and water supply were the most affected hazard damages or people paid much more attention on that during the typhoon hazard. And we did further analysis on sentiment and classified the human emotion in three categories, negative, positive and neutral sentiment. Fig.3 gives us a direct illustration of people' sentiment to different categories of hazard loss types. The blue, green, red colors refer to negative, neutral, and positive emotion respectively. For example, there is a micro-blog message saying that "The typhoon is really terrible", we can identify the Chinese sentiment word "terrible" to put this short sentence in category of negative sentiment. But,there is a micro-blog message saying

Figure 4 Distribution map of different types of hzard damage information from social media.

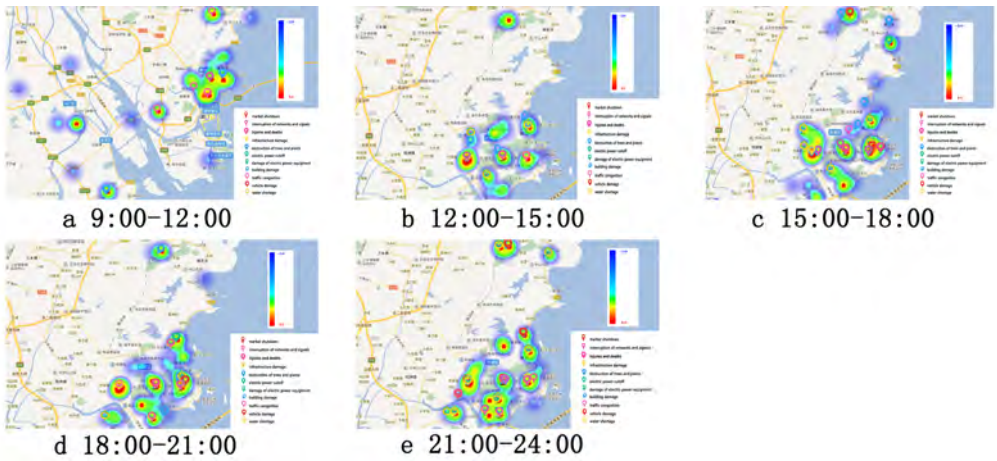

Figure 5 Social meida based spatio-temporal sentiment analysis.

that “The typhoon has great destructive power, it has no electricity until now, but thanks to the power workers who are working hard to repair it. Give them a thumbs up”. This text contains “Power supply outage”, but the emotion is positive, which shows that people were satisfied with the disaster reduction response. We can get people’s reaction to the typhoon event or know how severe of the hazard influence on the people’ life there. And we also got the geospatial information of different damage types and visualized in the map as shown in Fig. 4. This map shows the hazard damage distribution of hazard affected area. And this map can be a useful supplemental geospatial data to the offical hazard mitigation. Most important thing is that, the geospatial data can be obtained in a near real-time mannner which is just the insufficiency of the common geospatial data acquisition method. As Fig. 5 shown, we did a spatio-temporal sentiment analysis of typhoon “hato” during its landfall on coast of Zhuhai City. Before the landfall of the typoon as shown in Fig. 5a, the major hazard influence was traffic congestion. And people of the hazard-affected area were in a hurry to return home. With the typhoon landfalled and moved to the northwest, more categories of hazard damages emerged along with the landfall route as illustrated in Fig. 5b,c,d,e. And outage of power and water supply was the prominent influence types of the hazard by analysis of the number and sentiment of social media records. And as the typhoon passed by, the negative emotion decreased and positive emotion increased.

We evaluated the experimental results with precision, recall and F-1. The comprehensive evaluation index of different disaster loss categories was greater than 0.74. And the comprehensive evaluation index of different sentiment categories was greater than 0.83.

4 Methods

Social media data contains a lot of valuable near-real time geospatial information during the hazard events. This paper introduced the method to extract hazard related geospatial information for evaluation of hazard loss. And we proposed a Social Media based Hazard information Recognition and Classification model for hazard related geospatial information extraction and analysis based on the nature language processing and sentiment analysis. Typhoon "Hato" (landed on the coast of Zhuhai city at 12:50, 23rd August, 2017) was selected as the case study. Firstly, social media data of hazard event were collected and cleaned for further hazard information extraction. Nature language processing and semantic interpretation was done to understand the content of the text of social media data. A hazard damaged evaluation standard with eleven categories was proposed for the information classification. And these hazard loss categories were geo-located and mapped to show the distribution of hazard loss. Also sentiment analysis was done to extracted people's reaction to the Typhoon hazard. The geospatial time-serial map of sentiment analysis was generated. In our recent research, We are developing a near real-time social media based hazard information acquisition and analysis system. And we will use more hazard events to train the model before it can be used in practical hazard mitigation.

References

1 Valentina Cerutti, Georg Fuchs, Gennady Andrienko, NataliaAndrienko, and Frank Ostermann. Identification of disaster-affected areas using exploratory visual analysis of georeferenced tweets: application to a flood event. In *16th Annual Symposium on Foundations of Computer Science, Berkeley, California, USA, October 13-15, 1975*, pages 1–5, 2016.

2 P. Thakuriah et al. *Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies*. Springer Geography, 2016.

3 Xu et al. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(4):1–9, 2016. doi:10.1186/s13638-016-0553-0.

4 Michael F. Goodchild. *Citizens as sensors: web 2.0 and the volunteering of geographic information*. GeoFocus, 2007.

5 Linna Li and Michael F. Goodchild. The Role of Social Networks in Emergency Management: A Research Agenda. *International Journal of Information Systems for Crisis Response and Management*, 2(4):49–59, 2010. doi:DOI:10.4018/jiscrm.2010100104.

6 Milad Mirbabaie, Stefan Stieglitz, and Stephan Volkeri. Volunteered geographic information and its implications for disaster management. In *HICSS '16 Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS),Washington, DC, USA ,January 05 - 08, 2016*, pages 207–216, 2016. doi:10.1109/HICSS.2016.33.

7 Jibo Xie Tengfei Yang and Guoqing Li. A social media-based dataset of typhoon disasters. *China Scientific Data*, 2018(3), 2018. doi:10.11922/scdata.2017.0014.en.

8 Kayoko YAMAMOTO and Shun FUJITA. Development of Social Media GIS to Support Information Utilization from Normal Times to Disaster Outbreak Times. *International Journal of Advanced Computer Science and Applications*, 6(9):1–14, 2015.