Journal of Geospatial Surveying (2021) 1:1
DOI: https://doi.org/10.4038/jgs.v1i1.25

**ORIGINAL ARTICLE**

# Crowdsourced Data Relevance Analysis for Crowd-assisted Flood Disaster Management

Kaushalye, N. A. V. O.*[a], Koswatte, S.[a,b]

[a] *Department of Remote Sensing and GIS, Faculty of Geomatics, Sabaragamuwa University of Sri Lanka, Sri Lanka*

[b] *School of Civil Engineering and Surveying, Faculty of HES, University of Southern Queensland, Australia*

ARTICLE INFO

*Article history:*
Received 12 March 2021
Received in revised form 4 May 2021
Accepted 5 May 2021
Available online 31 May 2021

*Keywords:*
Crowdsourced Data
Quality assessment
Relevance ranking
Disaster Management
Natural Langue Processing

ABSTRACT

The recent climate changes have significantly increased the number and intensity of natural disasters around the world. This includes floods that cause a great deal of damage to properties, and more importantly, to the lives of the people. The reporting of current disasters has changed from official media to public reporters through social media and crowdsourcing technologies which have guaranteed the availability and up-to-date nature of the reported data. However, crowdsourced data (CSD) are often questioned due to issues in reliability and relevancy, heterogeneity or bias, bad structure, and un-professionalism. As a result of this, disaster responders are reluctant to use such data for their critical decision making actions. Using Natural Language Processing (NLP) and Geographic Information Retrieval (GIR) techniques, this study evaluated the quality of CSD, focusing on the thematic relevance. The study examined a proof of concept on relevance assessment based on an improved set of user queries utilizing crowdsourced messages from the 2011 Australian floods (Ushahidi Crowd-map). The findings show that the approach was effective in generating a thematically rated list of CSD messages for post-flood disaster managers to confidently take actions. The study's future work will consider thematic and geographic specificities and semantic context of the modified queries. Moreover, it is expected to test the approach with similar geospatial crowd-map data, and finally to check the possibility of integrating the derived information with authoritative datasets such as Spatial Data Infrastructures (SDIs).

## 1 Introduction

Crowdsourcing is the act of obtaining inputs from a large number of people and has become one of the latest spatial data collection methods. It has been in existence well before the digital age and has now become a novel data collection method in the modern world. A key objective of crowdsourcing is to source personal properties and services, including ideas from a large group of citizens, relatively open and faster means, mainly through the internet. Generally, in crowdsourcing, a particular work is distributed among the participants to obtain a cumulative result (Goodchild, 2007).

Crowdsourcing is relatively an easier and inexpensive means to collect data through the support of the community and more interestingly, it is a real-time data collecting method. Crowdsourced Data (CSD) contains more updated information and leads to faster solutions than traditional methods. However, anybody can intentionally provide inaccurate data and hence not all CSD are legitimate. The key challenges of CSD lie with its quality, including credibility and relevance (Foody et al., 2015).

CSD generally consists of data sets that were generated with the aid of a large group of people. Social media is a popular platform for collaborative mapping and crowdsourcing. Crowdsourced Spatial Data (Spatial CSD) or Volunteered Geographic Information (VGI) are special form of crowd generated content through the web based tools (Goodchild, 2007) that includes a kind of geographical information. Therefore, it may be possible to analyze Spatial CSD using spatial analysis tools. Although CSD contains location information, they cannot always be considered as complete spatial data due to their incomplete structures and missing metadata (O'Donovan et al., 2012).

Effective disaster management needs to have real-time, reliable, and high quality spatial data. During disaster management, available data sources such as authoritative data may not optimally be configured to achieve the objectives of the disaster management in full due to completeness, access, and availability issues. The other

* Corresponding author: Department, Sri Lanka.
*E-mail address:* naosanda@stdgeo.sab.ac.lk (N.A.V.O. Kaushalye).

option is to use other forms of data, such as CSD, in tandem with government maintained authoritative data. The use of data provided through CSD creates an opportunity to have higher levels of currency or further depths. Although CSD are freely available, this data may pose many problems due to lack of accuracy, reliability and structure related issues which need to be addressed prior to use in critical applications (Ciepłuch et al., 2010).

According to previous studies, multi-criteria ratings, scoring, and validation based on spatial-temporal clustering, hybrid computational and hybrid and manual processes, opinion mining and sentiment analyzing, and rule-based reasoning approaches were all used to analyze the CSD significance. Each approach's suitability is determined by the data used and applications in hand.

The rest of this paper is organized as follows. Section 2 describes the related work followed by the methodology of Crowdsourced data relevance analysis. A detailed discussion of the results obtained is provided in the section 4. Finally, the study is concluded in the section 5 with information about future directions.

## 2 Related Work

Natural disasters are becoming more common and more intense around the world (Crooks and Wise, 2013; Ogie et al., 2019). A recent report published by the International Federation of Red Cross and Red Crescent Societies (IFRC) (2020) states that climate and weather related natural disasters are increasing and noticed almost 35% rise since 1960. It also states that climate and weather-related disasters have killed more than 410,000 people during the last decade. Floods are also a kind of natural phenomena, and both the number of incidents and the number of people impacted have risen dramatically in recent years (Hirata et al., 2018). With the recent developments of information technology and other infrastructures, a shift can be observed in the firsthand reporting of natural disasters from official and standard media to the crowd who are often at the disaster scene. Similarly, relief organizations have changed their strategies towards crowd opinions when collecting situational information (Sakurai and Murayama, 2019). This kind of information is now widely available as Crowdsourced Data (CSD), published by the crowd through social networks and related services.

Social media and other Information Technology (IT) based tools such as Social Networking Services (SNS) (Facebook, MySpace), microblogs (Twitter), video and pictures (e.g. YouTube, Flicker), and other collaborations (e.g. Wikipedia) (Kavota et al., 2020) played a significant role in recent disasters by enabling citizens to seek for support, sharing information on support and resources, alerting fellow citizens about threats, breakdowns, road closures and alternative routes and many more. Examples are Ushahidi based Crisis Map of the Czech republic (Pánek et al., 2017), Australian Broadcasting Corporation (ABC)'s Ushahidi based flood crisis map (Crowd-map) (Potts et al., 2011), and Sahana information sharing system which is originally developed by a Sri Lankan group during the 2004 Indian Ocean Tsunami and later successfully utilized during the 2010 Earthquake in Haiti (Sakurai and Murayama, 2019).

CSD are easily available and usually comes in the form of big data. However, researchers are continuously questioning about the reliability and relevance (Basiri et al., 2019; Senaratne et al., 2016), and the heterogeneity or the bias (Fonte et al., 2017) of CSD. This prevents CSD from being used as a direct source of information for disaster management purposes due to the reason that using this information in its raw form may lead to inaccurate decision making by the relevant officials. However, researchers have identified the possibility of improving the quality and reliability of CSD using diverse approaches. Koswatte et al. (2018) analyzed the relevance of a flood related crowdsourced dataset using Geographic Information Retrieval (GIR) and Natural Language Processing (NLP) techniques, with the help of specially designed set of queries. They have noted that their queries were more geographically specific than thematically specific. Additionally, they have highlighted the importance of modifying the user queries for better relevance ranking results which has been addressed in this research.

The disaster management activities, including post-flood disaster management, depend highly on location information. To utilize the CSD for post-flood disaster management purposes, it is required to retrieve only the relevant and reliable information from these massive datasets. The Information Retrieval (IR) concepts discussed in the Information Technology (IT) domain can easily be adapted with certain modification to match with CSD retrieval. GIR is a type of information retrieval that employs geographic indexing and retrieval, and it is a method that handles the geographic imprecision and uncertainty (Zaila, 2015). GIR's main goal is to find place names or toponyms within a corpus (i.e., a broad organized collection of text, such as web pages, documents, or social media posts) and their corresponding geographic position, which is referred to as 'concept@location.' (Andrade and Silva, 2006). Several mechanisms have been proposed for GIR purposes, including weighted geo-textual similarity measures, extended vector space model, probabilistic models, dynamic assessment of the specificity of the users' search context, and semantic and ontology-based models (Andrade and Silva, 2006; Cai, 2002; De Sabbata and Reichenbacher, 2010; Kumar, 2011; Yu and Cai, 2007). These techniques can be used in conjunction with NLP techniques to identify relevance of data for particular purpose (such as post-flood disaster management) even with data that have very low signal-to-noise ratio, such as social media data (Stowe et al., 2016). This research utilized the Term Frequency-Inverse Document Frequency (TF-IDF) numerical statistical model available in GIR domain, along with NLP techniques to deal with relevance ranking of CSD.

## 3 Methodology

The crowdsourced information can easily be gathered from the public by using diverse social networking and related tools through the internet. In this study, the quality

assessment of CSD is done by the relevance ranking method. Natural Language Processing techniques have been used in this study with the python libraries, where the dataset was selected related to a flood that occurred in Queensland, Australia, in 2011.

### 3.1 Study Area and Ushahidi Crowd-map

The state of Queensland, Australia, was struck by a series of floods in 2011. Thousands of people were forced to flee their homes as a result of this flooding, while over 200,000 people were affected in at least 90 cities. The cost of the damage was originally projected to be around A$1 billion, but it was later increased to A$2.38 billion. The Australian Broadcasting Corporation (ABC) (www.abc.net.au, accessed 20 May 2019) created a customized version of the Ushahidi Crowd-map (Fig. 1) to report and map disaster communications during the 2011 Australian floods. The Ushahidi ("testimony" in Swahili) software is a crisis mapping tool created by African citizen journalists to report on election-related violence in Kenya during the 2008 election defeat. It allowed people to report crisis information via the Internet or mobile platforms using SMS. During the 2011 Australian floods, people used this Crowd-map to exchange flood information.

The quality of CSD for post-flood emergency management was evaluated using thematic relevance analysis techniques in this study. The testing dataset was the Ushahidi Crowd-map dataset of the 2011 Australian floods.

For this study, 10,000 random messages were selected from the Crowd-map reports for the relevance assessment to be used for the post-flood disaster management purposes.

Fig. 2 below shows the overall CSD thematic relevance analysis procedure used in this research. Initially, it was required to define a set of user queries to process the CSD, and different user queries were established by carefully modifying Koswatte et al.'s (2018) five user queries (Table 1), as per their recommendation, to extract information related to floods within Toowoomba (which is an Australian town in Queensland that was flooded in 2011). Such queries were selected to obtain information for people interested in specific information about flooding within the study area.

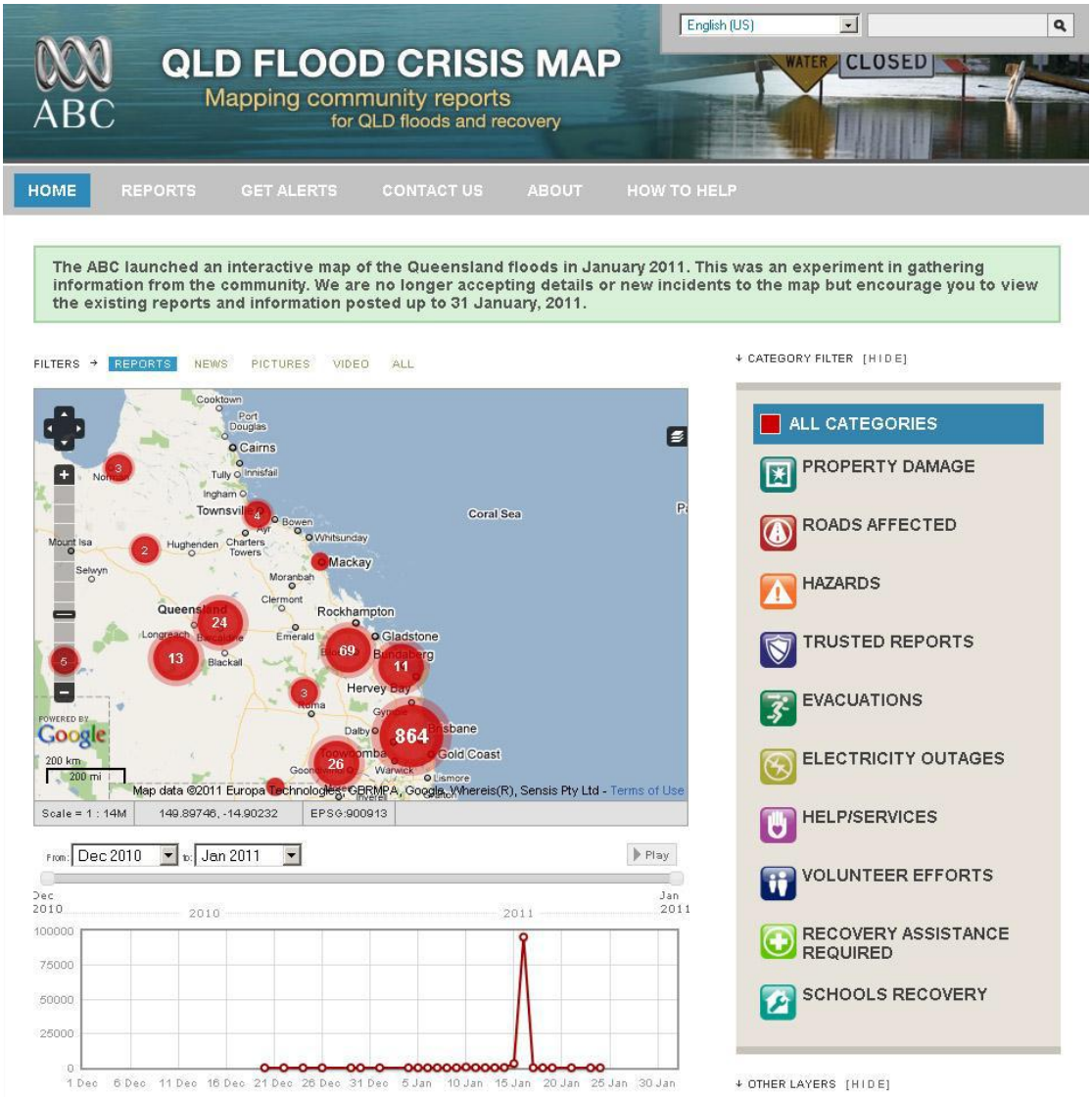


**Fig. 1:** Ushahidi Crowd-map Australian Floods 2011 (Potts et al., 2011).

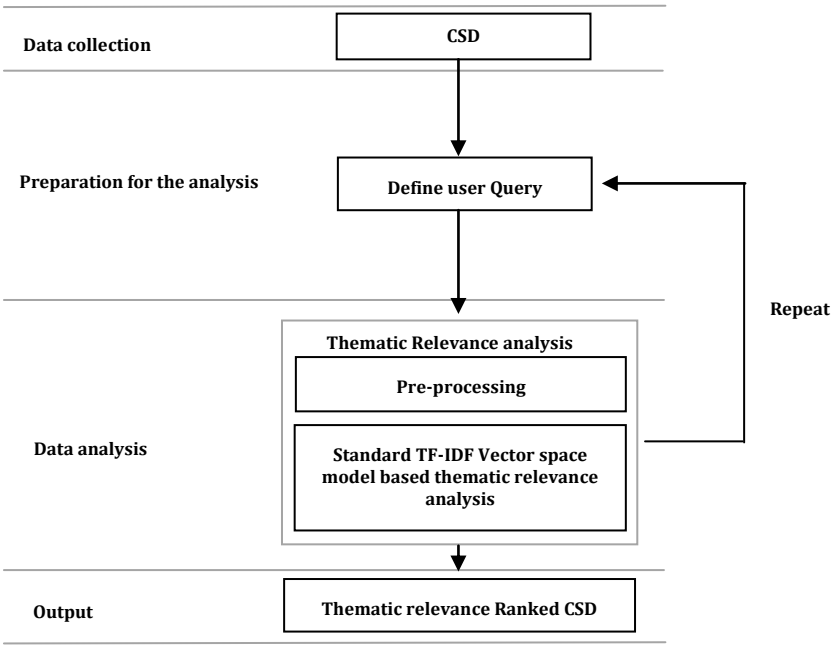


**Fig. 2:** Thematic relevance ranking procedure.

**Table 1:** User Queries adapted from (Koswatte et al., 2018).

| No. | Query |
|---|---|
| 1 | Road closed flood Toowoomba |
| 2 | Highway closed |
| 3 | Evacuation center open |
| 4 | Heavy rainfall Toowoomba |
| 5 | Flash flooding Toowoomba |

CSD were then pre-processed to prepare the unstructured raw dataset for subsequent processing. Duplicate elimination, tokenizing, stop-word removal (i.e., eliminating common terms similar to prepositions, and so on), stemming and lemmatization (i.e., changing "drinking" to "drink"), and removing non-words such as numbers, white spaces, and so on, were all part of the process.

To check the information relevance to a particular task, the Term Frequency – Inverse Document Frequency Vector Space Model (TF-IDF VSM) was often used to evaluate textual data (i.e., a document or query). This research used the Natural Language Toolkit (NLTK), Pandas library in the Jupyter Notebook (Anaconda) open-source keyword matching information retrieval system based on this TF-IDF VSM. Term Frequency (TF) is a metric that measures how often a given term appears in each text in the corpus. It is the ratio of the number of times a word appears in a document to the total number of words in the document. This increases as the number of times the word appears in the document increases. Each and every document has its own TF and the TF-IDF model employed a weighting method in which the significance of phrases, or words, in a text was measured statistically, using the procedure outlined below.

$$\mathrm{TF}(t) = \frac{\text{Number of times the term } t \text{ occurs in a message}}{\text{Total number of terms in the message}} \quad (1)$$

Next, the inverse document frequency (IDF) of the term t was determined by,

$$\mathrm{IDF}(t) = \log_e \left[\frac{\text{Total number of messages}}{\text{Total number of messages where the term exist}}\right] \quad (2)$$

Then the $(\mathrm{TF\text{-}IDF})_{t,m}$ weight for term t in message m was calculated using the following equation.

$$(TF - IDF)_{t,m} = TF_{t,m} * IDF_{t,m} \quad (3)$$

Finally, the thematic similarity score $\mathrm{Sim}_{T(q,\ m)}$ was determined using the similarity between the message m for the term t, and the query q was calculated by the following equation.

$$Sim_{T(q,m)} = \sum_{t \in q}(TF - IDF)_{t,m} \quad (4)$$

The procedure was repeated with modifying queries which were defined by the researchers Koswatte et al. (2018) until the best values were obtained for the number of hits (i.e., the number of CSD messages retrieved by each query) and thematic similarity score.

## 4 Results and Discussion

During the CSD thematic analysis, 10,000 Ushahidi Crowdmap messages were selected after the pre-processing of over 49,000 messages. The pre-processing consisted of actions such as duplicate removal, stemming and

lemmatizing, which generate the root form of the inflected words by removing words less than 3 characters, and removing the white space which is the section of a document that is unused or space around an object.

After the pre-preprocessing, the Term Frequency, and Inverse Document Frequency (TF-IDF) values were calculated using "TfidfVectorizer" in the python toolkit. The TF-IDF is a weighting factor that is often used in information retrieval. This weight is a statistical measure used to determine the importance of a word in a text. Simply put, the proportion to the number of times a word appears in the text, the value is compensated for by the frequency of the word within the whole corpus. Search engines often use the TF-IDF weighting as a key method in scoring and rating the importance of a document within the context of user queries. This study used it for calculating the relevance of words in the CSD messages for flood disaster management purpose.

A matrix was formed, which included TF-IDF values related to each word of a selected CSD message. For the relevance calculations, this study used different sets of the modified versions of the criteria (Table 1, user queries) developed by Koswatte et al. (2018) for assessing the relevance of Queensland, Australia, floods in 2011. The above procedure was repeated while carefully changing the queries, whilst keeping the balance between thematic and geographic scopes, and checking resultant number of hits and the thematic similarity scores. The most appropriate set of queries were selected based on these values. Table 2 shows the final set of queries determined by the study suitable for CSD relevance ranking.

**Table 2:** Final successful user queries after the experiments.

| No. | Query |
|---|---|
| 1 | Road closed flood Toowoomba |
| 2 | Highway closed |
| 3 | Evacuation center open |
| 4 | Heavy rainfall |
| 5 | Flash flooding Toowoomba |

The sum of TF-IDF value for each CSD message was then calculated. The final ranked message list was derived using the thematic relevance ranking method explained in the methodology section using the final user queries stated in the Table 2. The first 25 ranked messages of the list are shown in Table 3, with their sum of TF-IDF values.

**Table 3:** Ranked Message list.

| Rank | Message | SUM of TF-IDF |
|---|---|---|
| **1** | Roads in Toowoomba are not open. The road from Toowoomba to Brisbane is closed and Flagstone | 1.132254 |
| **2** | Dakabin Toowoomba Gympie shelters are closed due to flooding | 0.756556 |
| **3** | Secondary evacuation Center operational Centre Brisbane | 0.616403 |
| **4** | Breakfast creek road flooded and closed bnefloods | 0.593037 |
| **5** | Have just seen firsthand the aftermath at Toowoomba. Hard to believe the sheer force of flash flood | 0.582444 |
| **6** | Soaked Victoria warned of flash flooding | 0.529276 |
| **7** | Amazing footage of Toowoomba flood Help QLDERS with Flood Relief Appeal | 0.526146 |
| **8** | Welsby st New Farm #road flooded and closed | 0.504992 |
| **9** | Cafe open or closed let us know we can help put the word out here | 0.504744 |
| **10** | Flash flooding alert for Melbourne yes Melbourne | 0.476357 |

The quality is a key aspect when using CSD for critical applications such as in disaster management because the authoritativeness is always questionable in CSD. This study developed a technique to generate a thematically ranked message list for post-flood disaster management by ranking the thematic relevance of the CSD. During natural disasters such as floods, it is critical to support victims to save lives and quickly identify an appropriate resource which all are highly dependent on accurate and relevant spatial information.

## 5 Conclusions

Different people with different backgrounds and expertise levels typically design CSD using heterogeneous tools. The reporting of traffic delays during disasters, for instance, is recorded in different ways, such as road closed-down, no-go zones, road floods, roads underwater, road flooding, highway shutdown, water over the roads, and so on. The

Crowd-map material includes similar incidents in numerous forms. It is very challenging to distinguish similar meanings only through keyword search. This work tested the thematic significance of a highly unstructured and heterogeneous set of CSD using Natural Language Processing tools, and the Geographic Information Retrieval technique, along with a set of improved user queries.

Collecting spatial data through Crowdsourcing is very useful for updating spatial datasets and supporting authoritative data collections, ultimately to support decision making in disaster management. This research conducted a CSD relevance ranking analysis through an automated method of thematic relevance ranking. This will support in identifying CSD communications that were more relevant in the sense of post-flood disaster management. Previous research has identified that user queries are playing a very significant role in the process of CSD relevance assessment. Therefore, this research considered it seriously and put an effort to test the outcomes of using different sets of user queries. Finally, it identified a better set of user queries that can support the relevance analysis of CSD for the post-flood disaster management.

Future directions of this research plan to test the impact of these modifications over the thematic and geographic specificity of the queries, and semantically improve the context. Implementing a machine learning algorithm that takes into account the various terms of each of the queries, as well as the effects of the related thematic and geographic assessment, may be a viable solution. Through machine learning and semantic implementations with further automation, the study will seek less ambiguous, fast processing and improved results. There are other similar geospatial crowdsourcing platforms used in crisis situations such as MicroMappers, Digital Humanitarian Network and Google's Crisis Response program. This research is planned to be extended by analyzing the CSD contents collected through other crowd-maps and similar platforms using the approach developed in this research. Further, it is planned to check the possibility of integrating the improved CSD with authoritative data such as Spatial Data Infrastructures (SDIs).

## Acknowledgement

The authors wish to acknowledge Monique Potts, Australian Broadcasting Corporation (ABC) – Australia, for providing the Ushahidi Crowd-map data of 2011 Queensland Floods, Australia.

## Author Contributions

Conceptualization, methodology, analysis, writing original draft preparation, N.A.V.O.K., and writing, review and editing, S.K. All authors have read and agreed to the published version of the manuscript.

## Conflict of Interest

The authors declare no conflict of interest.

## References

Andrade, L., Silva, M. J., Relevance Ranking for Geographic IR. Proceedings of Workshop on Geographic Information Retrieval - SIGIR '06, Seattle, USA, 2006.

Basiri, A., Haklay, M., Foody, G., Mooney, P., Crowdsourced geospatial data quality: Challenges and future directions. Taylor & Francis, 2019.

Cai, G., GeoVSM: An integrated retrieval model for geographic information. In: M. J. Egenhofer, D. M. Mark, Eds.), Proceedings of International Conference on Geographic Information Science (GIScience 2002). Springer, Boulder, CO, USA, 2002, pp. 65-79.

Ciepłuch, B., Jacob, R., Mooney, P., Winstanley, A. C., Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010. University of Leicester, 2010, pp. 337.

Crooks, A. T., Wise, S., 2013. GIS and agent-based models for humanitarian assistance. Computers, Environment and Urban Systems. 41**,** 100-111.

De Sabbata, S., Reichenbacher, T., A probabilistic model of geographic relevance. Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR-10). ACM, Zurich, Switzerland, 2010, pp. 23.

Fonte, C. C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J. J., Bayas, J.-C. L., See, L., Vatseva, R., 2017. Assessing VGI data quality. Mapping and the citizen sensor. 137-163.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D., Comber, A., 2015. Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. The Cartographic Journal. 52**,** 336-344.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal. 69**,** 211-221.

Hirata, E., Giannotti, M., Larocca, A., Quintanilha, J., 2018. Flooding and inundation collaborative mapping–use of the Crowdmap/Ushahidi platform in the city of Sao Paulo, Brazil. Journal of Flood Risk Management. 11**,** S98-S109.

IFRC, World Disaster Report 2020. International Federation of Red Cross and Red Crescent Societies: Geneva, Switzerland, 2020, 2020.

Kavota, J. K., Kamdjoug, J. R. K., Wamba, S. F., 2020. Social media and disaster management: Case of the north and south Kivu regions in the Democratic Republic of the Congo. International Journal of Information Management. 52**,** 102068.

Koswatte, S., McDougall, K., Liu, X., 2018. Relevance assessment of crowdsourced data (CSD) using semantics and geographic information retrieval (GIR) techniques. ISPRS International Journal of Geo-Information. 7**,** 256.

Kumar, C., Relevance and ranking in geographic information retrieval. Proceedings of the Fourth BCS-IRSG conference on Future Directions in

Information Access. British Computer Society, 2011, pp. 2-7.

O'Donovan, J., Kang, B., Meyer, G., Höllerer, T., Adalii, S., Credibility in context: An analysis of feature distributions in twitter. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing. IEEE, 2012, pp. 293-301.

Ogie, R. I., Clarke, R. J., Forehead, H., Perez, P., 2019. Crowdsourced social media data for disaster management: Lessons from the PetaJakarta. org project. Computers, Environment and Urban Systems. 73**,** 108-117.

Pánek, J., Marek, L., Pászto, V., Valůch, J., 2017. The Crisis Map of the Czech Republic: the nationwide deployment of an Ushahidi application for disasters. Disasters. 41**,** 649-671.

Potts, M., Lo, P., McGuinness, R., Ushahidi Queensland Floods Trial Evaluation Paper: A collaboration between ABC Innovation and ABC Radio. ABC Australia, 2011.

Sakurai, M., Murayama, Y., 2019. Information technologies and disaster management–Benefits and issues. Progress in Disaster Science. 2**,** 100012.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., Haklay, M., 2016. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science. 1-29.

Stowe, K., Paul, M., Palmer, M., Palen, L., Anderson, K., Identifying and Categorizing Disaster-Related Tweets. Proceedings of conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 2016, pp. 1-6.

Yu, B., Cai, G., A query-aware document ranking method for geographic information retrieval. Proceedings of the 4th ACM workshop on Geographical information retrieval. ACM, Lisbon, Portugal, 2007, pp. 49-54.

Zaila, Y. L., Different tools for handling geographic information retrieval problems. Proceedings of the 6th Symposium on Future Directions in Information Access. British Computer Society, Thessaloniki, Greece, 2015, pp. 62-66.