Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

Article

Quality assessment of Crowdsourced Data (CSD) using semantics and Geographical Information Retrieval (GIR) techniques

Saman Koswatte [1,4], Kevin McDougall [2]* and Xiaoye Liu [3]

[1] University of Southern Queensland, Australia; Saman.Koswatte@usq.edu.au
[2] University of Southern Queensland, Australia; Kevin.McDougall@usq.edu.au
[3] University of Southern Queensland, Australia; Xiaoye.Liu@usq.edu.au
[4] Sabaragamuwa University of Sri Lanka, Sri Lanka; sam@geo.sab.ac.lk
* Correspondence: Kevin.McDougall@usq.edu.au; Tel. : +617-4631-2545

Abstract:

Crowdsourced Data (CSD) generated by citizens is becoming more popular as its potential utilisation in many applications is increasing due to its currency and availability. However, the quality of CSD, including its relevance, is often questioned as the data is not generated by professionals nor follows standard data collection procedures. The quality of CSD can be assessed according to a range of attributes including its relevance. Information relevance has been explored through using in Geographic Information Retrieval (GIR) techniques to identify relevant information. This research tested a relevance assessment approach for CSD by adapting relevance assessment techniques available in the GIR domain. The thematic and geographic relevance were assessed using the Term Frequency-Inverse Document Frequency (TF-IDF), Vector Space Model (VSM) and Natural Language Processing (NLP) techniques. The thematic and geographic specificities of the queries were calculated as 0.44 and 0.67 respectively, which indicates the queries used were more geographically specific than thematically specific. The Spearman's rho value of 0.62 indicated that the final ranked relevance lists showed reasonable agreement with a manually classified list and confirmed the potential of the approach for CSD relevance assessment for other possible crowdsourced data analysis.

Keywords: Crowdsourced Data, Relevance, Semantics, Geographic Information Retrieval, Natural Language Processing

1. Introduction

The traditional methods of geographic information production have continued to change as new software tools and methods emerge as a result of the technological, infrastructure, communication and Information Technology (IT) developments of the modern world. Geographic information collected and voluntarily produced by untrained citizens using modern information and communication tools is often termed as Volunteered Geographic Information (VGI) [1]. Some crowdsourced data (CSD) can be considered as a subset of VGI when the user location is considered, however, CSD often has limited location information compared to VGI [2]. This form of new data has gained increased attention due its potential utilisation in many applications. The information currency and availability of CSD is high however, its quality including its reliability (credibility) and usability (relevance) are still unclear.

The quality of geospatial data has long been considered in the field of geospatial information management where assessment parameters and techniques are often defined. However, CSD does not follow standard data collection procedures nor is the data generated by skilled geospatial professionals. Therefore, CSD often does not have a clear data structure or metadata and therefore

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

the application of traditional spatial data quality assessment parameters and techniques may be problematic. Researchers are therefore testing new parameters and methods for CSD quality assessment and have proposed credibility and relevance as possible quality indicators. In our recent work [3], we have presented a CSD credibility assessment approach using a spam email detection technique. Choosing the most relevant geospatial information is important if high quality outcomes are expected in geospatial data dependant applications, as not all CSD may be related to the task at hand. Data that is not relevant or has a low relevance is of limited use for applications such as emergency management. In large datasets, data that is of low relevance may exists and therefore, relevance analysis of CSD is important prior to utilising this data in applications such that require trustworthy data.

Geographic relevance is applied in many of today's human information inquiry activities e.g. in search engines. Geographic relevance can be defined as 'a relation between a geographic information need and the spatio-temporal expression of the geographic information objects needed to satisfy it' [4]. The fields of Information Retrieval (IR) and modern web based Geographic Information Systems (GIS) have now matured to provide professional outputs for their own information requests. These developments suggest that the combined use of GIS and IR systems to handle the requests on geo-textual information are now more effective [5].

This paper discusses the use of a GIR technique used in the IT domain to analyse the relevance of CSD. The paper is structured as follows: Section two discusses the background of CSD relevance and its analysis. Section three describes the methods used in the study. Section four details the results of the study and section five discusses their implications. Finally, section six provides some concluding remarks and some future suggestions for research.

2. Data Relevance in respect to Crowdsourced Data and Use of Geographic Information Retrieval for Crowdsourced Data Relevance Analysis

Relevance is naturally cognitive and 'the greater the cognitive effects the greater the relevance and the smaller the processing efforts to derive these effects, the greater the relevance' [6]. It is highly dependent on the end user's requirements regardless of being a product or information. The context of relevance has long been studied in diverse fields including philosophy, communication, logic, psychology, artificial intelligence, natural language processing, documentation, information science and information retrieval [7]. Saracevic [7] identified five types of relevance namely: (1) topical or cognitive relevance, (2) algorithmic relevance, (3) pertinence or intellectual relevance, (4) situational relevance and (5) motivational or affective relevance. This research proposes to focus on the situational relevance which can be defined as 'usefulness of the viewed and assessed information' towards the task at hand and information needs of the user [8] and is more appropriate to assessing the CSD relevance in a post-disaster management context.

Geographic information retrieval seeks to retrieve geographically relevant documents [8-15] or identify unambiguous geographic associations [16] based on the user's requirements. Simple word or toponym matching is not adequate for geographic information retrieval purposes [15]. Therefore, toponym matching based on semantic similarity measures may be the most appropriate approach.

2.1. Adapting Geogprahic Information Retrieval Process for Crowdsourced Data Relevance Analysis

The key objective of Geographic Information Retrieval (GIR) is to identify the place names or toponyms within a corpus (a large structured set of text, e.g. web sites, documents or social media posts) and their corresponding geographic location[8]. On the one hand, it is a process that manages the imprecision and ambiguity as geographic names are often ambiguous [17]. On the other hand, it is a process of ranking the relevance in two dimensions namely thematic and geographic [8] with the assumption that they are independent from each other [12].

Although, the GIR field is relatively new [12], numerous mechanisms have been proposed such as weighted geo-textual similarity measures [8], extended vector space model [5], probabilistic models [10], dynamic assessment of the specificity of the users' search context [18], and semantic and ontology based models [19] to identify relevant geographic information. Similarly, techniques can be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

applied along with natural language processing techniques to detect relevance of data with very low signal-to-noise ratios [20] and even in a near-real-time context [21]. De Sabbata and Reichenbacher [10] suggested that GIR concepts can be utilised to estimate the relevance of geographic objects based on user context by converting geographic distances into similarity scores.

Monterio et al. [21] highlighted four techniques associated with the various stages of GIR based search engine pipelines, namely, (1) geographic indexing, (2) query expansion, (3) recognition and use of place names and (4) geographic ranking. A number of key challenges lie in the area of analysing and processing sets of documents and queries, textual-geographical indexing and ranking the documents using the relevance criteria [12].

2.1.1. Managing Thematic Relevance

The presence of relevant terms in a document provides an indication of the relevance of the document for a selected task. From an information analysis perspective, the terms can be weighted based on the importance of the task at hand. A commonly used weighting method is the Term Frequency-Inverse Document Frequency (TF-IDF) model. In this model, higher weights are assigned for specific terms appearing more frequently in a document. This is based on the premise that the more frequently a given term appears, the more likely that document is relevant to the search. Conversely, a low weight will be assigned to more commonly available terms in the whole document set.

2.1.2. Managing the Geographic Relevance

Managing the geographic relevance or discovering and disambiguating toponyms that exist in the text document has been identified as the process of Geographic Scope Resolution (GSR) [21,22]. Generally, GSR consists of three tasks namely (1) geo-parsing (identifying toponyms), (2) reference resolution (toponym resolution) and (3) ground referencing (mapping toponyms to a footprint) [21]. Common geo-parsing methods include gazetteer lookup based (searching and tested the location terms against a Gazetteer), rule based (identifying location terms based on pre-defined rules) and machine learning based methods (trained to detect location terms based on correlation measures with reference data i.e. training corpus) [23]. The reference resolution process which maps the relevant toponyms is mandatory when ambiguities occur [21].

This research suggests that the Natural Language Processing (NLP) based gazetteer lookup approaches are viable for semantically extracting location information from CSD. The geographic information retrieval can be performed by NLP software such as GATE[1]. GATE is a robust and scalable open-source java based tool developed by the University of Sheffield, United Kingdom for text processing including semantic processing. This type of work may be supported by an ontological gazetteer for both toponym identification and ambiguity resolution. The ground referencing (geo-tagging) can also be assisted by the ontological gazetteer.

Usually after the GSR process, there is a need to calculate the geographic focus of a message. Different approaches are available for geographic focus detection such as measuring the geographic similarity and relevance ranking. The geographic similarity measures can be calculated based on region overlaps [24] or calculating a non-linear normalised distance between the scopes of the document and the query [8,17,25]. Andrade and Silva [8] explored a model which combined the ontological geographic relevance calculations whilst Zaila and Montesi [17] proposed a model based on topological relations, metric proximity calculations and ontological geographic similarity calculations.

[1] https://gate.ac.uk

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

2.2. Relevance Ranking and Merging the Thematic and Geographic Relevance

In order to prepare a final relevance ranked list of messages it is important to consider both the calculated thematic and geographic relevance lists. A combined relevance ranked list would also allow the faster retrieval of the geographic information identified. In GIR research the weighted sum method for relevance fusion is commonly utilised [8,17-19]. It is often advantageous to consider the specificity of the query scope in assessing the CSD thematic relevance as suggested by Yu and Cai [18]. They also reported that Dempster-Shafer's method of evidence combination shows superior results in their experimental study which was also very close to human judgments in many cases.

2.3. Quality Assessment of the Crowdsourced Data Relevance Analysis

Quality assessment is essential to confirm the validity of the approach utilised. There are various quality metrics to test the performance and quality of the results from these types of analyses. Measures such as recall and precision are popular measures in these classification systems. However, precision is often regarded as a more important measure than recall in rank based IR systems if the user does not intended to retrieve all of the relevant records [26]. In relation to the information retrieval, precision refers to the fraction of correctly identified documents that are relevant to the query in relation to all retrieved documents (https://wikipedia.org). The precision can be calculated by:

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (1)$$

Other measures including Average Precision (AP), Mean Average Precision (MAP) and Precision at K are the measures often used in modern web based information retrieval systems as other indicators may not represent meaningful measures where thousands of relevant documents are present. Average Precision refers to the precision averaged across all values of recall between 0 and 1. CSD differs from general spatial data and analysing its relevance remains challenging. The details of the methods applied to test the relevance of a selected CSD dataset is explained in the next section.

3. Materials and Methods

Previous research showed that CSD relevance analysis has been investigated using a variety of methods. However, the suitability of each approach depends on the data and the application. This research selected Geographic Information Retrieval techniques to assess the CSD relevance for post-flood emergency management through thematic and geographic relevance analysis. The geographic information retrieval processes were implemented using a Java framework, Lucene IR software and the GATE natural language processing software. The Ushahidi Crowdmap dataset of 2011 Australian floods was used as the testing dataset. From the Crowdmap reports, 200 random messages were selected for this analysis for faster data manipulation and to better understand the system's behaviour. After the pre-possessing of the initial data set, 182 reports remained for the thematic and geographic relevance analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

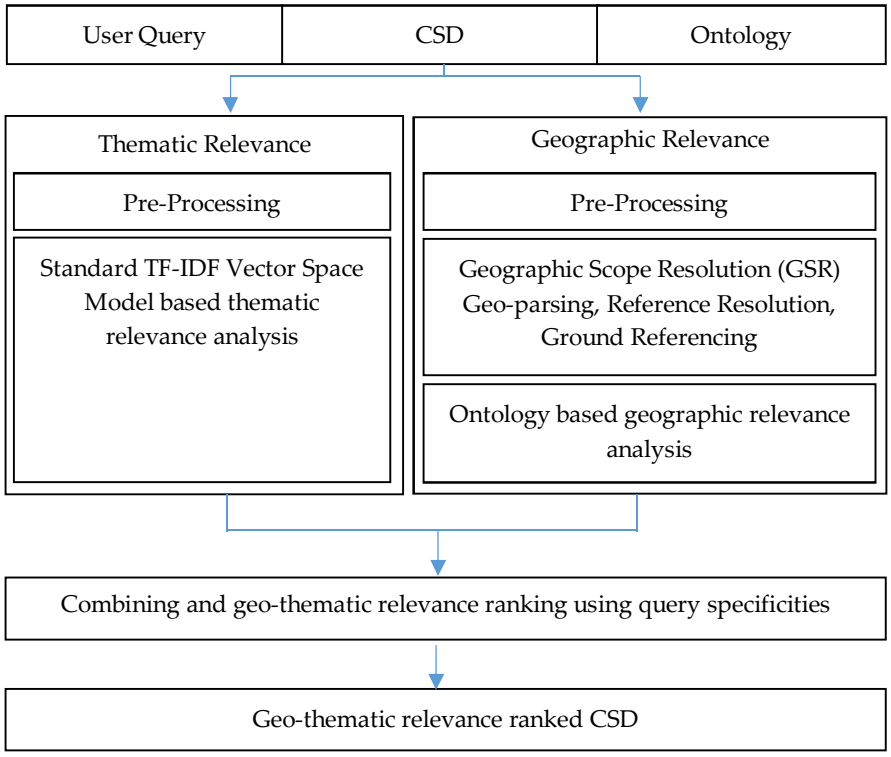

Figure 1. CSD relevance detection approach adapted from Zaila and Montesi's [17] GIR architecture

Figure 1 depicts the overall CSD relevance analysis approach adopted in this research. Five queries (Table 1) were set to extract flood related information within the Toowoomba (a city in Queensland affected by the 2011 Australian floods). The selected CSD dataset was analysed based on two key relevance dimensions i.e. the thematic relevance and the geographic relevance, utilising the user queries and ontology developed in our previous work [27].

Table 1. User queries

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

3.1. Thematic Relevance Analysis

3.1.1. Pre-processing

Initially, the pre-processing of CSD was carried out to prepare the unstructured raw dataset for further processing. This included actions such as duplicate removal, tokenizing, stop-word removal (i.e. removing common terms similar to prepositions etc.), stemming and lemmatization (i.e. bringing the word to its base form such as changing 'flooding' to 'flood') and removing non-words such as numbers, white spaces etc.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

3.1.2. Term Frequency-Inverse Document Frequency Vector Space based Model for Thematic Relevance Analysis

The Term Frequency-Inverse Document Frequency Vector Space Model (TF-IDF VSM) was utilised to analyse textual data (i.e. a document or query) to test the relevance to a particular task. This model is used in many of the information retrieval applications including GIR and this study utilised the Lucene[2] open-source keyword matching information retrieval system which is based on the standard TF-IDF VSM model. Lucene is a high-performance, fully featured text search engine library written entirely in Java.

The CSD thematic relevance analysis was conducted using two Java programs which were constructed based on Lucene 6.0 API and its standard analyser. The first Java program was developed for indexing the dataset and the second program was used to perform the searching using the TF-IDF VSM model.

The TF-IDF model utilised a weighting function where the importance of terms or words in a document is statistically estimated using the following process.

Firstly, the Term Frequency (TF) of term t was calculated by:

$$TF(t) = \frac{\text{Number of times the term } t \text{ occurs in a message}}{\text{Total number of terms in the message}} \quad (2)$$

Next, the Inverse Document Frequency (IDF) of the term t was determined by:

$$IDF(t) = \log_e \left[\frac{\text{Total number of messages}}{\text{Total number of messages where the term } t \text{ exists}}\right] \quad (3)$$

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

$$(TF - IDF)_{t,m} = TF_{t,m} * IDF_{t,m} \quad (4)$$

Finally, the thematic similarity score $Sim_T(q,m)$, which represents the similarity between the message m for the term t and the query q, was calculated using:

$$Sim_{T_{(q,m)}} = \sum_{t \epsilon q} (TF - IDF)_{t,m} \quad (5)$$

After the TF-IDF values of the message terms were calculated, the message was represented in a Vector Space Model (VSM) which is an algebraic model for representing text documents. In this process, each document is represented by vectors of identifiers i.e. index terms weighted based on their importance using a model such as the TF-IDF model. The axes of the vector space are denoted by the terms of the message.

3.2. Geographic Relevance Analysis

The next stage was the geographic relevance analysis which included the Geographic Scope Resolution (GSR) process (i.e. geo-parsing, reference resolution and ground referencing) and was performed using a natural language processing based gazetteer lookup approach. These tasks were carried out using the GATE software.

3.2.1. Pre-processing

The selected sample of the CSD dataset had to first undergo pre-processing to filter inappropriate content such as duplicates. However, tokenizing, stemming and lemmatizing pre-processing tasks which were used in the thematic relevance analysis were not performed during the

[2] http://lucene.apache.org

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

pre-processing of geographic relevance analysis as they were undertaken within the GATE software through a morphological analysis.

3.2.2. Geographic Scope Resolution (GSR)

During the GSR process, the geo-parsing was undertaken to identify and tag toponyms. These toponyms were then utilised for the geographic reference resolution to identify the best (i.e. most appropriate) toponym for the CSD report. The geographic reference resolution is more challenging when ambiguities such as geo-geo or geo-non-geo occur (that is when different locations share the same place name and some locations have non geographic terms such as people's names). Mostly, these situations consist of relationship terms such as 'near', 'between', 'crossing' and 'south of' etc. and contain contextually important information that can be resolved using context based semantic processing. The queries were split into triples to form <what, relation, where> relations by concatenating the individual tokens. In both of the above tasks, the possible toponyms in the message content were identified by searching the semantic Queensland local gazetteer (QLDGazOnto) reference list developed in our previous work [3].

The next step of the GSR process was the ground referencing. Initial ground referencing was performed using a Java program based on the Google geo-coding API. The Java Annotation Pattern Engine (JAPE) transducers were used to process the geographic references to identify appropriate locations according to the relationships. JAPE is a component of GATE which is useful for pattern-matching, semantic extraction, and many other operations in text processing. There were a number of issues identified during the processing including missing locations and ambiguities of the generated locations. Several JAPE rules were developed (see Figure 2) to resolve the ambiguities and to tag the messages and then to allocate location coordinates with the help of QLDGazOnto ontological gazetteer.

```
Phase: OntoMatching // phase name
Input: Lookup
Options: control = applet // control type
Rule: GeoTag      // rule name
({Lookup.class == Place}c
) //search for place names in the semantic gazetteer
:place-->
:place.Mention = {class = :place.Lookup.class, inst = :place.Lookup.inst}
//match and tag with toponym
```

Figure 2. Example of JAPE rule used for semantic tagging

3.2.3. Ontology Based Geographic Relevance Analysis

After completing the GSR process, the next task was to calculate the geographic similarity measures. The geographic similarity measures between the messages and queries were used to determine the relatedness of the CSD messages for the selected task at the identified location. The geographic similarities were calculated using equation (6) below by considering the geographic scope of the query and the geographic scope of each CSD report using the QLDGazOnto ontology information.

The similarity $Sim_G(q,m)$ between the geographic scope of the query (S_q) and geographic scope of the message (S_m) based on the ontology information was calculated using:

$$Sim_G(q,m)(S_q,S_m) = K \times \{Insd(S_q,S_m) + Proxm(S_q,S_m)\} + (1-K) \times Sib(S_q,S_m) \quad (6)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

The value for the variable K in the equation was identified as 0.8 after manual testing and normally lies between 0 and 1.

In the above equation, the component 'inside (*Insd*)' computed the weight if the scope of the message (S_m) was inside the scope of the query (S_q) based on the number of decedents in the ontology as:

$$Insd(S_q, S_m) = \frac{NumberOfDecendents(S_m)+1}{NumberOfDecendents(S_q)+1} \text{ and 0 otherwise} \quad (7)$$

If the scopes spatially overlap, then equation (6) returns values between 0 and 1. It is at a maximum when both scopes are equal and a minimum when the message scope has no decendents. $NumberOfDecendents(S_m) + 1$ returns the number of scopes spatially inside S_m plus the scope itself which can be derived from the ontology.

The component 'proximity (*Proxm*)' was assessed based on the inverse distance, where the distance was normalized by the diagonal of the Minimum Bounding Rectangle (MBR) of the query scope as:

$$Proxm(S_q, S_m) = \frac{1}{1+\frac{Dist(S_q,S_m)}{Diagonal(S_q)}} \quad (8)$$

Finally, the component 'siblings (*Sib*)' were tested to check whether the scope of the message (S_m) and scope of the query (S_q) were siblings by:

$$Sib(S_q, S_m) = 1 \text{ if } S_m \text{ and } S_q \text{ are siblings in the ontology, 0 otherwise} \quad (9)$$

For example, if the message scope (S_m) and query scope (S_q) are polygons representing 'Brisbane North' and 'Toowoomba' (Figure 3) respectively;

1. The function 'inside (*Insd*)' returns no value as the scopes do not spatially overlap.
2. The function 'proximity (*Proxm*)' returns a value based on the Distances 'D' and 'd' as indicated in the Figure 3.
3. The function 'siblings (*Sib*)' returns the value 1 as the two scopes are both siblings of the larger region in the ontology.

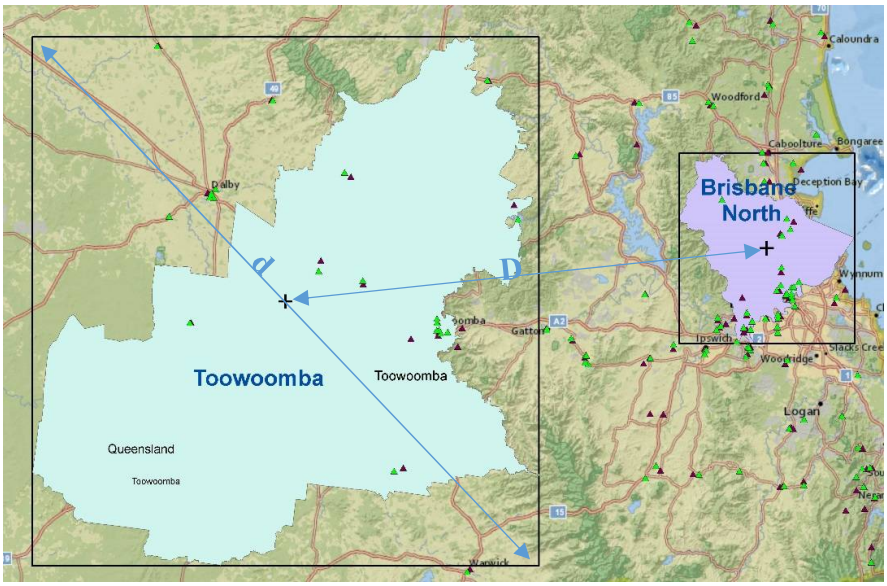

Figure 3. Proximity calculations in geographic relevance analysis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

3.3. Combining the Geographic and Thematic Relevance Rankings

Finally, the geographic and thematic relevance lists were merged to create the final geo-thematic relevance ranked list using the equations below which considered the thematic and geographic specificities of the queries. The specificity provided an indication of the quality of thematic and geographic relatedness of the queries considered.

The thematic specificity Spc_T of query $q = \{t_1, t_2, \dots, t_n\}$ was calculated by:

$$Spc_T = -\sum_{t \in q} \omega_t * CTM(t) \log\left(\frac{N_t+1}{N}\right) \quad (10)$$

Where: t_k be the k^{th} term of the query q,
ω_t is the weight for each term,
$CTM(t)$ is the Conceptual Term Matrix of term t from the WordNet[3] ontology,
N_t is the number of messages containing term t and N is the total number of messages in the dataset.

The conceptual term matrix $CTM(t)$ was calculated by firstly extracting conceptual information representatives of the term t (i.e. number of senses, number of synonyms, level number and number of sibblings) from the WordNet ontology in the form of integer values in CTM. Next, the weighting was performed to transform the values into weights based on the importance of the different information types and then the combined weighted values were calculated to give the final single score in $CTM(t)$.

The geographic specificity $Spc_G(q, m)$ of geo-referenced query q was calculated by:

$$Spc_G = -\log\left(\frac{Area(G_q)}{Area(G_M)}\right) \quad (11)$$

Where: G_q is the geometry representative of the associated geographic scope of query q,
$Area(G_q)$ is the area of the geographic scope of q,
$Area(G_M)$ is the area of the coverage of all messages in the dataset.

The final rank as a weighted sum of individual scores was calculated by:

$$Rel(q, m) = \omega_T * Sim_T(q, m) + \omega_G * Sim_G(q, m) \quad (12)$$

Where ω_T and ω_G are normalized weights of the two relevance scores and calculated by:

$$\omega_T = \frac{1}{\ln(e+Spc_T)} \quad (13)$$

$$\omega_G = \frac{1}{\ln(e+Spc_G)} \quad (14)$$

In addition to the thematic and geographic relevance analysis, a reference data set was constructed manually to classify messages from the total message data set that were considered to be relevant or not relevant to the disaster being investigated. This data set was utilised to test the accuracy of the classification processes.

The results of the process are discussed in the next section.

[3] https://wordnet.princeton.edu/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

4. Results

In the CSD relevance analysis, 182 Ushahidi Crowdmap messages were selected for the geo-thematic relevance analysis after the initial pre-processing.

4.1. Results of the Thematic Relevance Analysis

The quality of thematic relevance analysis used the Lucene benchmark quality assessment package. In this analysis two configuration files were constructed, one containing the queries and the other containing a manually classified test reference collection. The test reference collection consisted of relevant and non-relevant sets of messages for each query. These configuration files were used for the quality analysis along with the indexed file of CSD messages.

Table 2. Quality assessment results of thematic relevance analysis

No.	Query	# hits	Average Precision	P@5	P@10
1	Road closed flood Toowoomba	120	0.655	0.600	0.300
2	Highway closed	69	0.897	0.800	0.600
3	Evacuation centre open	21	0.595	0.400	0.300
4	Heavy rainfall Toowoomba	45	0.911	0.800	0.600
5	Flash flooding Toowoomba	55	0.903	0.800	0.500

Table 2 shows the performance test results of the thematic relevance analysis using the Lucene software. This research selected the Average Precision (AP), Mean Average Precision (MAP) and Precision at a Kcertain level K (P@K) metrics to analyse the quality of the CSD relevance assessment. The AP for a query q refers to the average precision for each relevant message retrieved and the MAP is the mean average precision of all Q queries.

The Average Precision (AP) and Mean Average Precision (MAP) were calculated by:

$$AP = \frac{\sum_{k=1}^{n}(P(K)\times(Rel(K))}{number\ of\ relevant\ messages} \quad (15)$$

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \quad (16)$$

Where K is the rank in the retrieved message list and $P(K)$ is the precision at cut off K in the list and $Rel(K)$ is an indicator function which provides 1 if the message at position K is relevant and 0 otherwise.

The measure Precision at K (P@K) reports the fraction of messages ranked in the top K results marked as relevant. Generally, in ranked lists in information retrieval systems such as web searches, there will be thousands of possible records and the user may not be interested in seeing all the records. Therefore, considering the top most records (i.e. 20 records at the top of the list) is more appropriate in ranked lists such as in a geo-thematic relevance ranked lists produced in this system.

The table 2 also shows the number of hits (i.e. the number of messages identified relevant to the each query) along with the average precision, precision at level 5 (P@5) and precision at level 10 (P@10) of the analysis. According to the Lucene benchmark quality results, the average precision of the relevance of the message retrieval to the queries was generally above or close to 0.6 which indicates the system performed well. The P@5 was generally above 0.4 and the minimum value was 0.3 which meant the system was better at identifying relevant documents at the top levels. The MAP of the quality assessment was calculated as 0.792 which is a good indication of systems performance for relevance assessment as the value 1 indicates the best performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

4.2. Results of the Geographic Relevance Analysis

The ground referencing of the geographic relevance assessment was performed using a Java program based on Google geo-coding API. The location availability of CSD messages were close to 90% (i.e. 163 out of 182 messages) after the GSR process. The geographic similarities were calculated with the value of K set to 0.8. The geographic scope of the queries was selected as Toowoomba local government area which was a polygon feature. In the case of a polygon feature, it can use the minimum bounding rectangle (MBR) or convex hull as the feature representing the geographic scope. Both the MBR and convex hull options (Figure 4) were tested in calculating locations inside and within proximity. The results identified that if the MBR was used instead of the convex hull, there was a 46% increase in area and therefore the selection of 12 additional points which did not belong to the Toowoomba local government area.

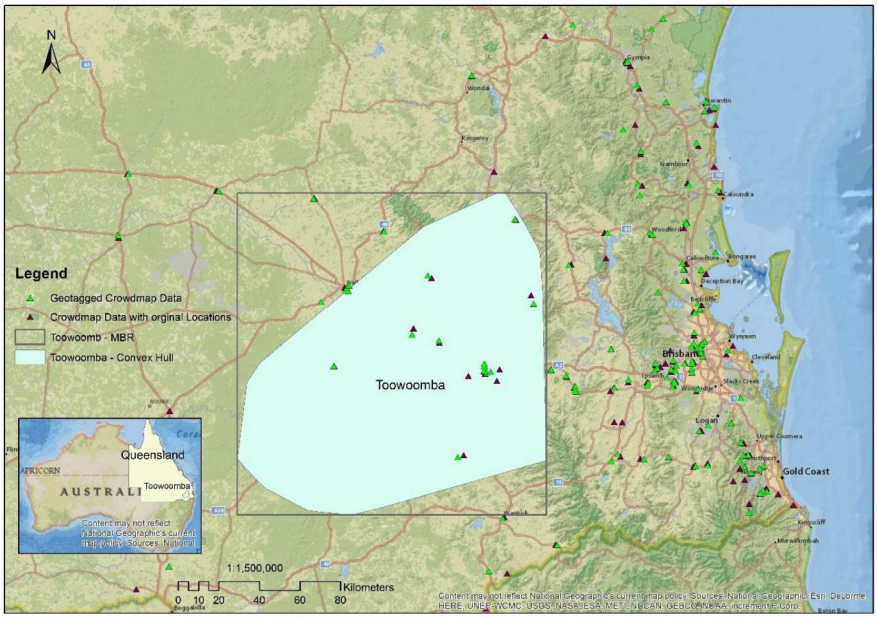

Figure 4. Crowdmap data and Toowoomba local government area using MBR and convex-hull

4.3. Results of the Final Geo-Thematic Relevance Ranking

The thematic specificity and geographic specificity of the analysis were calculated as 0.44 and 0.67 respectively. The values indicate that the queries used were less thematically specific and more geographically specific with the value 1 indicating the highest specificity. Table 3 shows the part of the CSD reports of the final geo-thematic relevance ranked list.

Table 3. Part of the final geo-thematic relevance ranked list

No.	CSD Report
1	Flash flooding has caused a shopping centre in Toowoomba to be closed.
2	Flash flooding caused landslide at Toowoomba range.
3	Flash flooding in Toowoomba region experiencing roadways cut off in town. Recent heavy falls within the last hour have managed to cut off some minor and major roads in Toowoomba CBD and surrounding suburbs.
4	The Warrego Highway at the Toowoomba Range is closed in both directions. Motorists are advised to seek an alternative route.
5	The Warrego Highway is presently closed at Jondaryan following heavy rain in the area.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

6	Toowoomba Regional Council crews and SES personnel are assessing road damage after today's severe flash flooding in Toowoomba. The main areas impacted were in the vicinity of East and West creeks which run through the centre of the city.
7	Flash flooding has caused a library to be evacuated.
8	The Clifton-Leyburn Road is OPEN WITH CAUTION from Clifton to Condamine River to all vehicles. There is no access to the Toowoomba-Karara Road and Ryeford-Pratten Road due to flood waters and pavement damage. Drivers are urged not to enter floodwaters.
9	Water bird habitat damaged-fences down at Toowoomba water bird habitat.
10	Road closed on Griffiths Street East of Mort Street.

The final rankings results were then compared with a human ranked list which was a manually re-ranked list of the final ranked list without considering the auto generated rank to see the system's ability to analyse the relevance compared to a human. Spearman's rho, which is a commonly used statistical test to compare agreement between two rankings where the value 1 indicates perfect match and -1 indicates complete inverse ranking [18], was calculated. The Spearman's rho was 0.62 which indicates the two lists agree with each other and confirms the validity of the approach for CSD relevance assessments.

5. Discussion

Understanding spatial data quality is essential for establishing confidence in the quality of the outputs of any spatial data dependent project. This research tested an information relevance assessment methodology for Crowdsourced data for the purpose of post-disaster management. In disasters such as floods, speedy identification of relevant and credible spatial information is important and is required to support victims and save lives.

This research analysed the CSD relevance based on two dimensions of the data relevance, namely the thematic relevance and geographic relevance. The thematic relevance assessment tested the degree to which the CSD was thematically relevant to the user queries. The results of the thematic analysis showed that the classification system performed well in analyzing CSD thematic relevance in respect to the defined user queries. However, the results would be different if the user queries were changed or used a different set of terms which may or may not be considered relevant. Therefore, it is suggested that future research into the sensitivity of the user queries be considered in order to normalize this impact. A possible solution may be to introduce a learning mechanism for the system that may consider the different query terms and the results of relevant thematic assessment.

The research used the Natural Language Processing (NLP) based gazetteer lookup for Geographic Scope Resolution (GSR) in the thematic relevance assessment. It applied stop-word and common-word filters to minimise the effect of frequently occurring terms. However, it identified drawbacks in the application of these filters. For example, the removal of terms such as 'Can' in toponyms such as 'Tin Can Bay' can render the geographic term unusable. Therefore, it is important to further understand similar effects and to identify precautions to prevent the removal of important terms. For geo-tagging purposes, the research used the Google geo-coding service with the support of the local semantic gazetteer (QLDGazOnto) for ambiguity resolution. This proved very useful in resolving geo-geo and geo non-geo ambiguities (e.g. Killarney in Ireland and Killarney in Australia, John Krebs is a personal name and there is a bridge called John Krebs Bridge in Murgon, Queensland, Australia).

The geographic relevance analysis assessed how geographically relevant the CSD was to the user queries. During this process, it was important to generate locations of the CSD using the Geographic Scope Resolution (GSR) process as the CSD locations were often missing. The ground referencing process of the GSR process utilised the Google geo-coding tool to assign locations for the identified locations. However, this tool was not capable of fully determining the local toponyms. This issue was rectified by using a local gazetteer and a JAPE rule based approach. In future, a local geo-coder should be utilised and would most likely improve the geo-coding of the results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

After the thematic and geographic relevancies were determined, the results were merged to calculate the combined geo-thematic relevance based on the thematic and geographic specificities of the queries utilised. During this process, it was identified that the queries were more geographically specific than thematically specific. This resulted in a higher weighting of the final ranked list towards the geographically relevant results. Although this approach is understood and was considered appropriate, this may not be the case in all analyses and it may be important to balance the thematic and geographic specificity in particular situations. In some cases, it may be appropriate to determine any bias towards the thematic or geographic relevance in the initial stages of the processing. This may assist in alerting the user in regard to balancing of the thematic and geographic specificity of the query terms. However, it will be important to consider whether it is a good approach to control the freedom of the user in setting their own queries.

6. Conclusions

CSD in general is curated by different people with different experiences and different knowledge levels using heterogeneous devices. In the Crowdmap content, people communicated similar incidents in various ways i.e. the intended meaning of road closures reported such as 'road closed, no go zone, water over the road, road under water, road flooded, road impassable, highway cut, water across road, etc.'. Identifying similar meanings using a keyword based search is challenging. This research tested the use of a semantic based thematic relevance assessment for highly unstructured and heterogeneous data such as CSD. It is recommended that further research should be directed towards understanding of the initial queries and their structure to improve the outcomes of the relevance analysis.

The research also considered query specificity for the final geo-thematic relevance joining and ranking. This was useful in identifying contextually more relevant CSD messages for flood disaster managers and other stakeholders. It is suggested that further work be completed to test and compare the performance and usefulness of other available geo-thematic relevance combination approaches.

Finally, it is noted that the GIR field is a fast-growing research area and new techniques are emerging regularly. This research suggests the need to test innovative and more stable approaches used in GIR to validate the applicability of similar approaches for CSD relevance studies.

Acknowledgments: Authors wishes to acknowledge the Australian Government for providing support for the research work through the Research Training Program (RTP) and Monique Potts, ABC – Australia for providing the 2011 Australian Flood's Ushahidi Crowdmap data.

Author Contributions: Saman Koswatte undertook the review of literature, contributed to the design of the methodology, completed the analysis of results and drafted sections of the paper. Kevin McDougall provided inputs to the methodology and analysis of the results including overall supervision and drafted part of the paper. Xiaoye Liu provided inputs to the methodology and analysis of results and guidance.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Goodchild, M.F. Citizens as sensors: the world of volunteered geography. *GeoJournal* **2007**, *69*, 211-221.
2. Koswatte, S.; McDougall, K.; Liu, X. Ontology driven VGI filtering to empower next generation SDIs for disaster management. In *Research at Locate 14*, Winter, S.; Rizos, C., Eds. Canberra, Australia, 2014.
3. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing* **2008**, *7*, 12-18.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

4. Raper, J. Geographic relevance. *Journal of Documentation* **2007**, *63*, 836-852.
5. Cai, G. In *GeoVSM: An integrated retrieval model for geographic information*, Proceedings of International Conference on Geographic Information Science (GIScience 2002), Boulder, CO, USA, 2002; Egenhofer, M.J.; Mark, D.M., Eds. Springer: Boulder, CO, USA, pp 65-79.
6. White, H.D. Relevance theory and citations. *Journal of Pragmatics* **2011**, *43*, 3345-3361.
7. Saracevic, T. In *Relevance reconsidered*, Proceedings of the Second conference on conceptions of library and information science (CoLIS 2), Copenhagen, 1996; Copenhagen, pp 201-218.
8. Andrade, L.; Silva, M.J. In *Relevance Ranking for Geographic IR*, Proceedings of Workshop on Geographic Information Retrieval - SIGIR '06, Seattle, USA, 2006; Seattle, USA.
9. Larson, R.R. In *Geographic information retrieval and spatial browsing*, Proceedings of Clinic on Library Applications of Data Processing - 1995, Illinois, USA, 1996; Smith, L.C.; Gluck, M., Eds. Illinois, USA.
10. De Sabbata, S.; Reichenbacher, T. In *A probabilistic model of geographic relevance*, Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR-10), Zurich, Switzerland, 2010; ACM: Zurich, Switzerland, p 23.
11. Janowicz, K.; Raubal, M.; Kuhn, W. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science* **2011**, *2011*, 29-57.
12. Kumar, C. In *Relevance and ranking in geographic information retrieval*, Proceedings of the Fourth BCS-IRSG conference on Future Directions in Information Access, 2011; British Computer Society: pp 2-7.
13. Wang, C.; Xie, X.; Wang, L.; Lu, Y.; Ma, W.Y. In *Detecting geographic locations from web resources*, Proceedings of Workshop on Geographic information retrieval at conference on Information and Knowledge Management (CIKM '05), Bremen, Germany, 2005; ACM: Bremen, Germany, pp 17-24.
14. Jones, C.B.; Purves, R.S. Geographical information retrieval. *International Journal of Geographical Information Science* **2008**, *22*, 219-228.
15. Jones, C.B.; Alani, H.; Tudhope, D. Geographical information retrieval with ontologies of place. In *Spatial information theory*, Springer: 2001; pp 322-335.
16. Amitay, E.; Har'El, N.; Sivan, R.; Soffer, A. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM: Sheffield, United Kingdom, 2004; pp 273-280.
17. Zaila, Y.L.; Montesi, D. Geographic information extraction, disambiguation and ranking techniques. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, ACM: Paris, France, 2015; pp 1-7.
18. Yu, B.; Cai, G. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, ACM: Lisbon, Portugal, 2007; pp 49-54.
19. Martins, B.; Silva, M.J.; Andrade, L. Indexing and ranking in Geo-IR systems. In *Proceedings of Workshop on Geographic information retrieval*, ACM: Bremen, Germany, 2005; pp 31-34.
20. Stowe, K.; Paul, M.; Palmer, M.; Palen, L.; Anderson, K. In *Identifying and Categorizing Disaster-Related Tweets*, Proceedings of conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 2016; Austin, Texas, USA, pp 1-6.
21. Monteiro, B.R.; Davis, C.A.; Fonseca, F. A survey on the geographic scope of textual documents. *Computers & Geosciences* **2016**, *96*, 23-34.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 April 2018 doi:10.20944/preprints201804.0338.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 256; doi:10.3390/ijgi7070256

626 22. Alexopoulos, P.; Ruiz, C.; Villazon-terrazas, B. KLocator: An Ontology-Based
627 Framework for Scenario-Driven Geographical Scope Resolution. *International*
628 *Journal on Advances in Intelligent Systems* **2013**, *6*, 177-187.
629 23. Leidner, J.L.; Lieberman, M.D. Detecting geographical references in the form of
630 place names and associated spatial natural language. *SIGSPATIAL Special* **2011**, *3*,
631 5-11.
632 24. Frontiera, P.; Larson, R.; Radke, J. A comparison of geometric approaches to
633 assessing spatial similarity for GIR. *International Journal of Geographical*
634 *Information Science* **2008**, *22*, 337-360.
635 25. Lieberman, M.D.; Samet, H.; Sankaranarayanan, J.; Sperling, J. In *STEWARD:*
636 *architecture of a spatio-textual search engine*, Proceedings of the 15th annual ACM
637 international symposium on Advances in geographic information systems, 2007;
638 ACM: p 25.
639 26. Inkpen, D. Information retrieval on the internet. 2007.
640 27. Koswatte, S.; McDougall, K.; Liu, X. VGI and crowdsourced data credibility analysis
641 using spam email detection techniques. *International Journal of Digital Earth* **2017**,
642 1-13.
643