International Journal of Geo-Information

MDPI

Article

Relevance Assessment of Crowdsourced Data (CSD) Using Semantics and Geographic Information Retrieval (GIR) Techniques

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

[1] School of Civil Engineering and Surveying, University of Southern Queensland, Darling Heights 4350, Australia; Saman.Koswatte@usq.edu.au or sam@geo.sab.ac.lk (S.K.); Xiaoye.Liu@usq.edu.au (X.L.)
[2] Department of RSGIS, Faculty of Geomatics, Sabaragamuwa University of Sri Lanka, P.O. Box 02, Belihuloya 70140, Sri Lanka
* Correspondence: Kevin.McDougall@usq.edu.au; Tel.: +617-4631-2545

Received: 26 April 2018; Accepted: 26 June 2018; Published: 29 June 2018

Abstract: Crowdsourced data (CSD) generated by citizens is becoming more popular as its potential utilization in many applications increases 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 characteristics including its relevance. In this paper, information relevance has been explored through using geographic information retrieval (GIR) techniques to identify the most highly relevant information from a set of crowdsourced data. This research tested a relevance assessment approach for CSD by adapting relevance assessment techniques available in the GIR domain. Thematic and geographic relevance were assessed by analyzing the frequency of selected terms which appeared in CSD reports using natural language processing techniques. The study analyzed crowdsourced reports from the 2011 Australian flood's Crowdmap to examine a proof of concept on relevance assessment using a subset of this dataset based on a defined set of queries. The results determined that the thematic and geographic specificities of the queries were 0.44 and 0.67, respectively, which indicated 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. In particular, this research has contributed to the field of CSD relevance assessment through an integrated thematic and geographic relevance ranking process by using a user-query specificity approach to improve the final ranking.

Keywords: crowdsourced data; relevance; semantics; geographic information retrieval; natural language processing

1. Introduction

The traditional methods of geographic information production have continued to evolve as new software tools and methods emerge as a result of technological, infrastructure, communication, and information technology developments. Crowdsourced data (CSD) is often used to describe the contributions of and comments by the crowd through social media and other specific platforms regarding a particular activity or event. 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 CSD can be considered as a subset of VGI when user location is considered, as CSD often has limited location information compared to VGI [2]. This form of new

data has gained increased attention due to its potential utilization in many applications such as in routing and navigation domains [3–5] and in disaster management [6–8]. The information currency and availability of CSD is high; however, its quality, including its reliability (credibility) and usability (relevance), is still unclear [9].

The quality of geospatial data has long been considered in the field of geospatial information management, where assessment parameters and techniques are often defined [10]. 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 so the application of traditional spatial data-quality assessment parameters and techniques may be problematic. Researchers are therefore exploring new parameters and methods for CSD quality assessment and have identified credibility and relevance as possible quality indicators [10–16]. Choosing the most relevant geospatial information is important if high-quality outcomes are expected in geospatial data dependent applications, as not all CSD may be related or relevant 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 exist and, therefore, relevance analysis of CSD is important prior to utilizing this data in applications that require relevant and 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" [17]. The fields of information retrieval 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 information retrieval systems to handle the requests on geo-textual information are now more effective [18] and include techniques such as rule-based spatial information retrieval techniques [19]. However, most current research approaches in the geographic information retrieval (GIR) domain adopt a simple geographic filter, which is often not sufficient where the CSD is highly variable and the expectation of the end user may be quite demanding and require multiple themes as part of a user request. It therefore requires further processing and the integration of multiple approaches to assess both the thematic and spatial relevance.

This paper discusses the use of a GIR technique used in the information technology domain to analyze the relevance of CSD. The study analyzed crowdsourced reports from the 2011 Australian flood's Crowdmap to examine a proof of concept on relevance assessment using a subset of this dataset based on a defined set of queries. The paper is structured as follows: Section 2 discusses the background of CSD relevance and its analysis. Section 3 describes the methods used in the study. Section 4 details the results of the study and discusses their implications. Finally, Section 5 provides some concluding remarks and some future suggestions for research.

2. Review of Current Literature

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" [20]. 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 [21]. Saracevic [21] 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. Another useful perspective of relevance is "situational awareness", which is "about the knowledge state of individuals tasked with monitoring and interpreting a situation and making decisions about how to act" [22], such as in the case of disaster responding. This research proposes to focus on situational relevance, which can be defined as the "usefulness of the viewed and assessed information" towards the task at hand and information

needs of the user [23], which is more appropriate to assessing the CSD relevance in a post-disaster management context.

Geographic information retrieval seeks to retrieve geographically relevant documents [24–31] or identify unambiguous geographic associations [32] based on the user's requirements. Simple word (term) or toponym matching is generally not adequate for geographic information retrieval purposes [31]. Therefore, toponym matching based on semantic similarity measures may often be the most appropriate approach.

2.1. Adapting Geographic 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., websites, documents, or social media posts) and their corresponding geographic location [25]. It is a process that manages imprecision and ambiguity, as geographic names are often ambiguous [33]. Often, it also includes a process of ranking the relevance in two dimensions, namely, thematic and geographic [25], with the assumption that they are independent of each other [28].

The researchers working in the field of GIR during the last decade or so "have developed more or less complete process chains" [34], such as weighted geo-textual similarity measures [25], extended vector space models [18], probabilistic models [26], dynamic assessment of the specificity of the users' search context [35], visually and computationally supported sense-making [36,37], and semantic and ontology-based models [38], to identify relevant geographic information. Similarly, techniques can be applied alongside natural language processing techniques to detect the relevance of data with very low signal-to-noise ratios [39] and even in a near-real-time context [40]. De Sabbata and Reichenbacher [26] suggested that GIR concepts can be utilized to estimate the relevance of geographic objects based on user context by converting geographic distances into similarity scores.

Monterio et al. [40] 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 analyzing and processing sets of documents and queries, textual-geographical indexing, and ranking the documents using the relevance criteria [28].

2.1.1. Managing the 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) [40,41]. Generally, GSR consists of three tasks, namely: (1) geo-parsing (identifying toponyms); (2) reference resolution (disambiguating toponyms); and (3) ground referencing (mapping toponyms to a footprint) [40]. Common geo-parsing methods include gazetteer lookup-based (searching and testing the location terms against a Gazetteer), rule-based (identifying location terms based on predefined rules), and machine-learning-based methods (trained to detect location terms based on correlation measures with reference data, i.e., training corpus) [42]. The reference resolution process which maps the relevant toponyms is mandatory when ambiguities occur [40].

This research suggests that the natural-language-processing-based gazetteer lookup approaches are viable for semantically extracting location information from CSD based on the experience of previous research [43]. Geographic information retrieval can be performed by natural language processing software such as GATE v8.1 (https://gate.ac.uk). GATE v8.1 is a robust and scalable open-source Java-based tool developed by the University of Sheffield, United Kingdom, for semantic text processing. This type of work may be supported by an ontological gazetteer for both toponym identification and ambiguity resolution.

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 [44] or calculating a nonlinear normalized distance between the scopes of the document and the query [25,33,45]. Andrade and Silva [25] explored a model which combined the ontological geographic relevance calculations, whilst Zaila and Montesi [33] proposed a model based on topological relations, metric proximity calculations, and ontological geographic similarity calculations.

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 utilized [25,33,35,38]. Often, there is a higher influence from the user queries over the geographic and thematic relevance scores calculated [25,35]. Therefore, it is important to consider the scope of the end user queries at the stage of combining the geographic and thematic relevance. Yu and Cai [35] suggested that it is often advantageous to consider the specificity of the query scope in assessing the CSD thematic relevance. They also reported that the Dempster–Shafer 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 utilized. 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 in these classification systems. However, precision is often regarded as a more important measure than recall in rank-based information retrieval systems if the user does not intend to retrieve all of the relevant records [46]. 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 [47]. 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. Average precision refers to the precision averaged across all values of recall between 0 and 1. CSD differs from general spatial data and analyzing its relevance remains challenging.

Senaratne, Mobasheri, Ali, Capineri, and Haklay [10] proposed a new category of VGI quality assessment to include "data mining approaches". They also state that the studies falling into this new category are largely independent of the geographic theories and that, currently, there has been limited research in the geographic domain of data mining. This paper addressed this research gap by extending the GIR approaches proposed Zaila and Montesi [33], Andrade and Silva [25], and Yu and Cai [35] for semantically assessing the thematic and spatial relevance of the CSD. This research analyzed how these available techniques could be integrated to assess CSD relevance and utilized a

user-query specificity approach to improve the final ranking. The details of the methods utilized to test the relevance of a selected CSD dataset are explained in the next section.

3. Materials and Methods

Previous research showed that CSD relevance analysis has been investigated using a variety of methods, including multicriteria rating [15,48], scoring and validation based on spatiotemporal clustering [11,49], opinion mining and sentiment analysis [50,51], hybrid computational and manual methods [52], and rule-based reasoning approaches [19]. 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 postflood emergency management through thematic and geographic relevance analysis. The geographic information retrieval processes were implemented using a Java framework, the Lucene v6.0 information retrieval software, and the GATE v8.1 natural language processing software. The Ushahidi Crowdmap dataset of 2011 Australian floods was used as the testing dataset. Ushahidi (meaning "testimony" in Swahili) software is a crisis-mapping platform which was originally developed by citizen journalists in Africa to report election-related violence in Kenya during the election fallouts in 2008. It enabled people to report crisis information through the internet or mobile platforms via SMS [53,54]. During the 2011 Australian floods, people utilized the Ushahidi-based Crowdmap developed by the Australian Broadcasting Corporation to share flood information [54]. From the Crowdmap reports, 200 random messages were selected for this analysis for faster data manipulation and to better understand the system's behavior. After the preprocessing of the initial dataset, 182 reports remained for the thematic and geographic relevance analysis.

Figure 1 depicts the overall CSD relevance analysis approach adopted in this research. Five queries (Table 1) were defined to extract flood-related information within Toowoomba (a city in Queensland affected by the 2011 Australian floods). These queries were selected with the aim of retrieving information which might be useful to people interested in relevant flood-related information within the geographic study area. The selected CSD dataset was analyzed based on two key relevance dimensions, i.e., the thematic relevance and the geographic relevance, utilizing the user queries and ontology developed in our previous work [16].

Table 1. User queries.

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

3.1. Thematic Relevance Analysis

3.1.1. Preprocessing

Initially, the preprocessing 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 (terms) to its base form, such as changing "flooding" to "flood") and removing nonwords such as numbers, white spaces, etc.

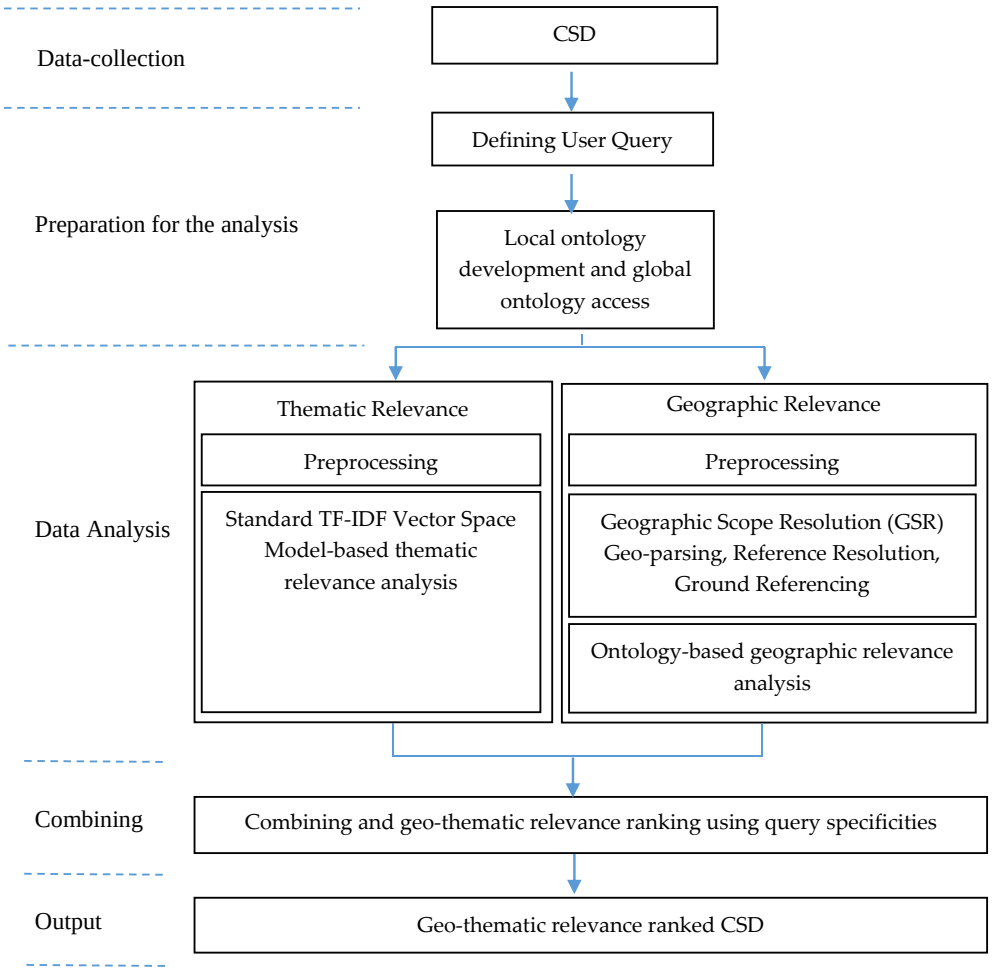

Figure 1. CSD relevance detection approach adapted from Zaila and Montesi's [33] GIR architecture.

3.1.2. Term Frequency Thematic Relevance Analysis

The Term Frequency–Inverse Document Frequency Vector Space Model (TF-IDF VSM) was utilized to analyze textual data (i.e., a document or query) to test the relevance to a particular task. This model is used in many information retrieval applications including GIR. This study utilized the Lucene v6.0 (http://lucene.apache.org) open-source keyword-matching information retrieval system, which is based on this term frequency model. Lucene v6.0 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 the Lucene v6.0 API and its standard analyzer. The key reasons for selecting this tool were its free and open-source nature and its support for term frequency calculations. The first Java program was used for indexing the dataset and the second program was used to perform the searching using the TF-IDF VSM model.

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

Firstly, the term frequency (–) 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 $(TF - IDF)_{t,m}$ 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 frequency 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, including 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 v8.1 software as it supported semantic processing and ontology development and editing.

3.2.1. Preprocessing

The selected sample of the CSD dataset had to first undergo preprocessing to filter inappropriate content such as duplicates. However, tokenizing, stemming, and lemmatizing preprocessing tasks, which were used in the thematic relevance analysis, were not performed during the preprocessing of geographic relevance analysis, as these tasks were undertaken within the GATE v8.1 software.

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 utilized 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 ambiguities occur (that is, when different locations share the same place name or where locations share the same name as a nongeographic term, such as a person's name). Mostly, these situations consist of spatial 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. This is to be incompatible with the semantic triples which were defined in the form of subject, predicate, and object. For example, the query "Road closed flood Toowoomba" can be put in a triple <Road, closed, Toowoomba> to form the <what, relation, where> relationship. In the GSR process, the possible toponyms in the message content were identified by searching the semantic Queensland local gazetteer (QLDGazOnto) reference list developed in our previous work [55].

The next step of the GSR process was the ground referencing or geo-coding. This was performed using the Java Annotation Pattern Engine (JAPE). JAPE is also 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 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})
//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 Equations (6)–(9) proposed by Andrade and Silva [25]:

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

The value for the variable K, which is used to maintain the similarity scores between 0 and 1, was set to 0.8 after manual testing.

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 descendants in the ontology as

$$Insd(S_q,S_m) = \frac{NumberOfDescendants(S_m)+1}{NumberOfDescendants(S_q)+1} \text{ IF scope } S_m \text{ is inside } S_q, \text{ and 0 otherwise} \quad (7)$$

If the scopes spatially overlap, then Equation (7) 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 descendants. $NumberOfDescendants(S_m)+1$ 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

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

The $Dist(S_q,S_m)$ is the distance between the scope of query and the scope of the message and is denoted by (D) in Figure 3. The diagonal (S_q) is the diagonal distance of MBR of the scope of the query and is denoted by (d) in Figure 3.

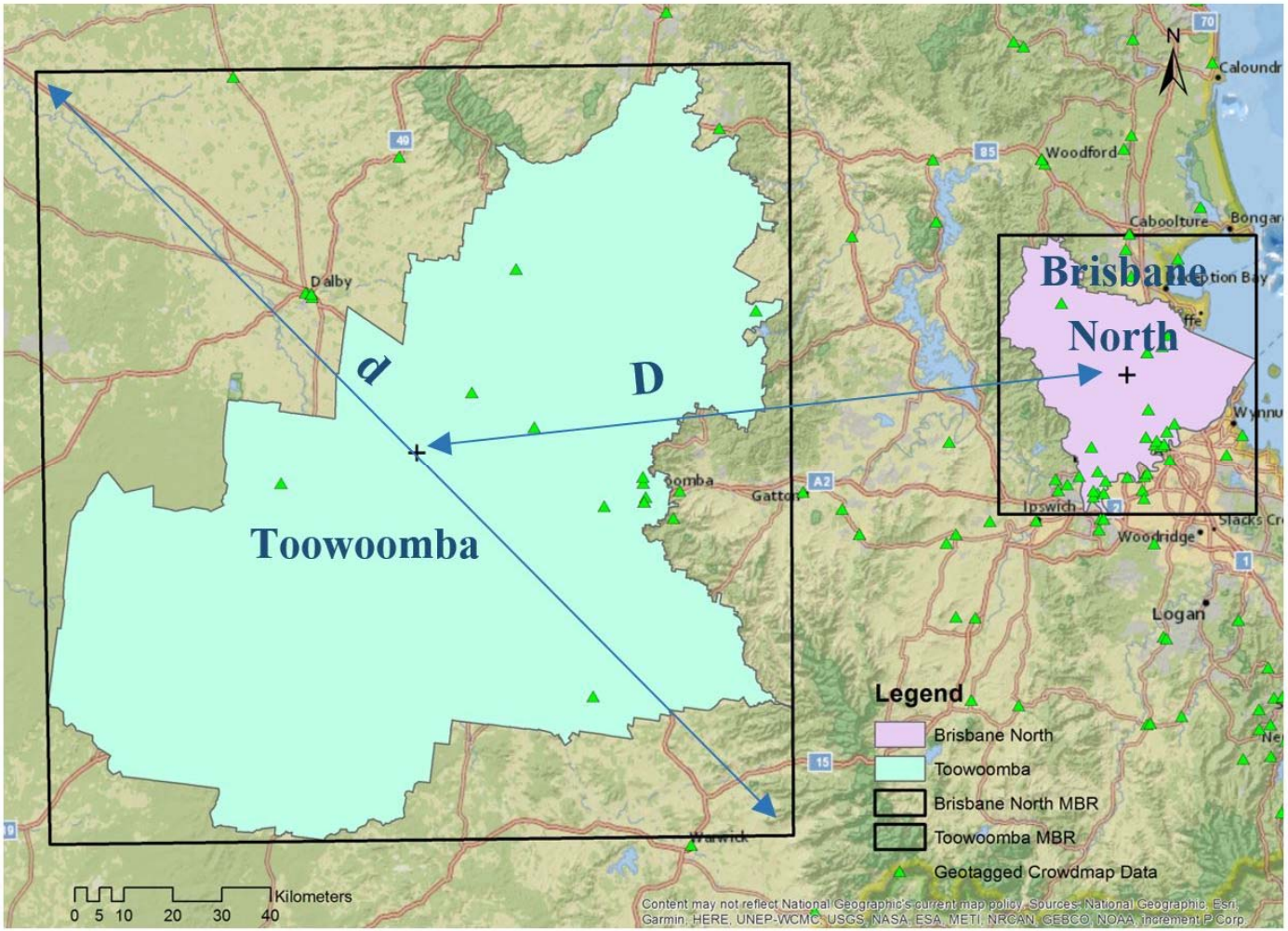

Figure 3. Proximity calculations in geographic relevance analysis.

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

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

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 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.

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 by considering the thematic and geographic specificities of the queries. The specificity provides an indication of the quality of the thematic and geographic relatedness of the queries considered. Yu and Cai [35] proposed Equations (10)–(14) to calculate the thematic and geographic specificities.

The thematic specificity Spc_T of query $q = \{t_1, t_2, \ldots, t_n\}$ 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 is the kth 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 (https://wordnet.princeton.edu) 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 siblings) from the WordNet ontology in the form of integer values. 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 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, and

$Area(G_M)$ 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 dataset was constructed manually to classify messages from the total message dataset that were considered to be relevant or not relevant to the disaster being investigated. This dataset was utilized to test the accuracy of the classification processes. The integration of the thematic and geographic ranking approaches provided an opportunity to improve the overall ranking of the relevance of the terms to the specific queries.

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

4. Results and Discussion

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

4.1. Results of the Thematic Relevance Analysis

The quality of thematic relevance analysis was reported from the statistics in the Lucene v6.0 software. In this analysis, two input 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 nonrelevant sets of messages for each query and provided a quality check on the relevance or otherwise of each message in relation to the query. These files were used for the quality analysis along with the indexed file of CSD messages.

Table 2 shows the performance test results of the thematic relevance analysis using the Lucene v6.0 software. This research selected the average precision (AP), mean average precision (MAP), and precision at a certain level K (P@K) metric to analyze 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.

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 center 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

The AP and MAP were calculated using the equations proposed by Liu [56] as

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

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

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 returns 1 if the message at position K is relevant and 0 otherwise.

The measure 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.

Table 2 also shows the number of hits (i.e., the number of messages identified relevant to 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 v6.0 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 highest performance.

4.2. Results of the Geographic Relevance Analysis

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 the Toowoomba local government area, which was a polygon feature. The process can use the polygon feature of the administrative local government area, a minimum bounding rectangle (MBR), or a simplified convex hull of the polygon as the feature representing the geographic scope. All these options were tested in calculating locations inside and within proximity (Figure 4). The convex hull (Figure 4) was deemed the appropriate option as it approximated the geographic extent of the local government area and also encapsulated points close to the edge of the administrative boundary. The MBR was not appropriate as it increased the selection area by 46% and the additional points were not relevant to the Toowoomba area. There was no detectable difference in processing time.

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 more geographically specific than thematically specific, with the value 1 indicating the highest specificity. In reality, these results will vary with the user queries utilized. Table 3 shows the part of the CSD report of the final geo-thematic relevance ranked list after running all the queries listed in Table 1.

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

Rank	CSD Report
1	Flash flooding has caused a shopping center 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.
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 center 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 ranking results were then compared with a manually ranked list to compare the system's ability to analyze the relevance of a query compared to a human. Spearman's rho, which is a commonly used statistical test to compare agreement between two ranked lists where the value 1 indicates a perfect match and -1 indicates a complete inverse ranking [35], was calculated. The Spearman's rho was 0.62, which indicates a good positive agreement between the two lists and supports the validity of the approach for the CSD relevance assessment.

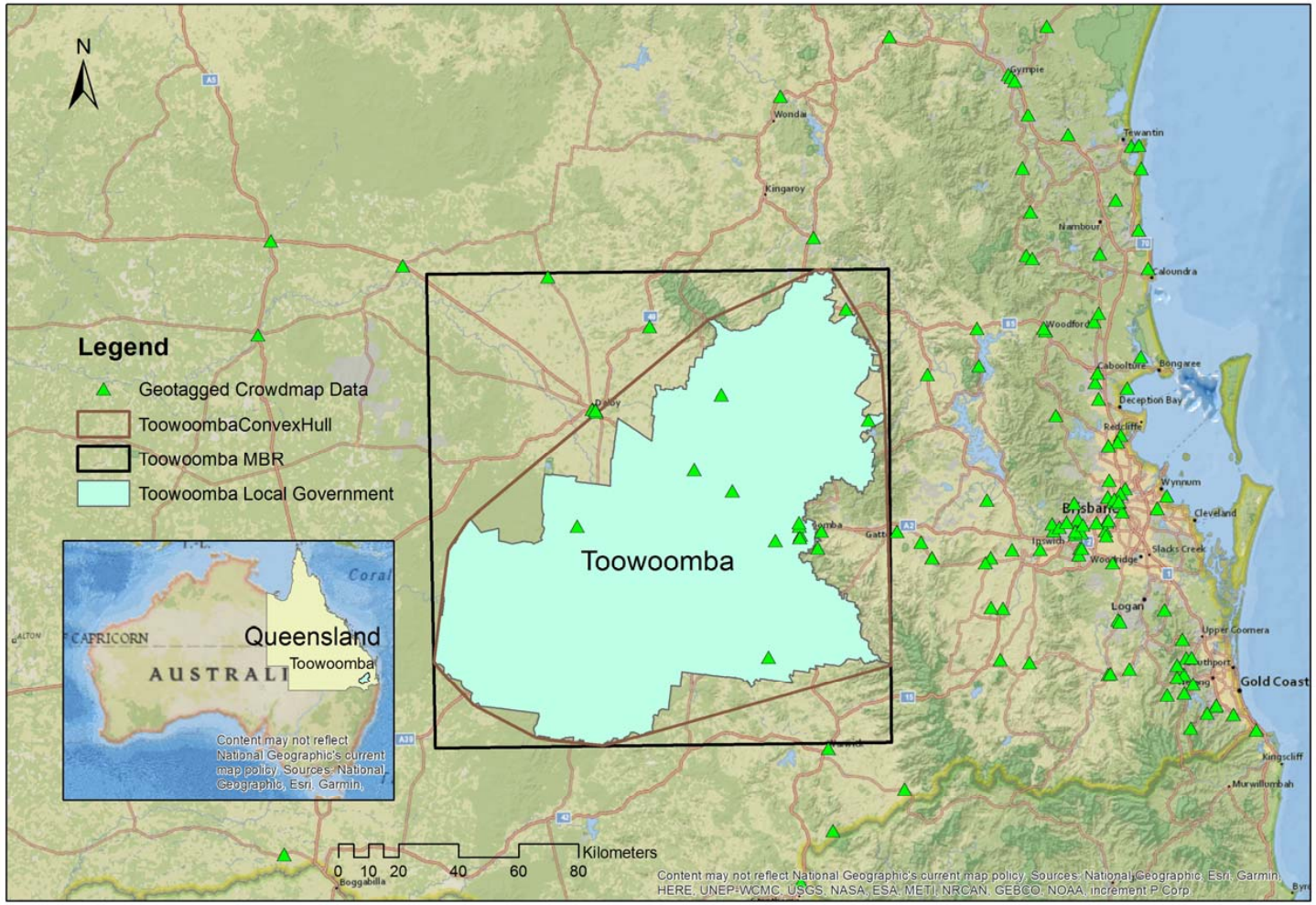

Figure 4. Crowdmap data and Toowoomba local government polygon, MBR, and simplified convex hull.

4.4. 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, timely identification of relevant and credible spatial information is important to support victims and save lives.

This research analyzed the CSD relevance based on two dimensions of 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-based gazetteer lookup for geographic scope resolution in the thematic relevance assessment. It applied stop-word and common-word filters to minimize the effect of frequently occurring terms. However, it identified limitations in the application of these filters. For example, the removal of terms such as "can" in toponyms such as "Tin Can Bay" can render a true 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 utilized 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 utilized and would most likely improve the geo-coding of the results.

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 utilized. 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 users in setting their own queries.

5. Conclusions

CSD in general is curated by different people with different experiences and different knowledge levels using heterogeneous devices. In the Crowdmap content, people communicate similar incidents in different ways, for example, the intended meaning of road closures can be reported in different ways, 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.

In particular, this research has contributed to the field of CSD relevance assessment through an integrated thematic and geographic relevance ranking process by using a user-query specificity approach to improve the final ranking. This was useful in identifying contextually more relevant CSD messages for the proposed application of post-flood disaster management. 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.

Author Contributions: S.K. undertook the review of literature, contributed to the design of the methodology, completed the analysis of results and drafted sections of the paper. K.M. provided inputs to the methodology and analysis of the results including overall supervision and drafted part of the paper. X.L. provided inputs to the methodology and analysis of results and guidance.

Funding: This research received no external funding.

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.

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. [CrossRef]
2. Koswatte, S.; McDougall, K.; Liu, X. Ontology driven VGI filtering to empower next generation SDIs for disaster management. In Proceedings of the Research @ Locate 2014, Canberra, Australia, 7–9 April 2014.
3. Keler, A.; Mazimpaka, J.D. Safety-aware routing for motorised tourists based on open data and VGI. *J. Location Based Serv.* **2016**, *10*, 64–77. [CrossRef]
4. Zipf, A.; Mobasheri, A.; Rousell, A.; Hahmann, S. Crowdsourcing for individual needs—The case of routing and navigation for mobility-impaired persons. In *European Handbook of Crowdsourced Geographic Information*; Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., Purves, R., Eds.; Ubiquity Press: London, UK, 2016; p. 325.
5. Prandi, F.; Soave, M.; Devigili, F.; De Amicis, R.; Astyakopoulos, A. Collaboratively Collected Geodata to Support Routing Service for Disabled People. In Proceedings of the 11th International Symposium on Location-Based Services, Vienna, Austria, 26–28 November 2014; pp. 67–79.
6. Haworth, B.; Bruce, E. A review of volunteered geographic information for disaster management. *Geogr. Compass* **2015**, *9*, 237–250. [CrossRef]
7. Horita, F.E.; de Albuquerque, J.P. An approach to support decision-making in disaster management based on volunteer geographic information (VGI) and spatial decision support systems (SDSS), In Proceedings of the 10th International ISCRAM Conference, Baden-Baden, Germany, 12–15 May 2013.
8. Granell, C.; Ostermann, F.O. Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Comput. Environ. Urban Syst.* **2016**, *59*, 231–243. [CrossRef]
9. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
10. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2016**, 1–29. [CrossRef]

11. Spinsanti, L.; Ostermann, F. Automated geographic context analysis for volunteered information. *Appl. Geogr.* **2013**, *43*, 36–44. [CrossRef]
12. O'Donovan, J.; Kang, B.; Meyer, G.; Hollerer, T.; Adalii, S. Credibility in context: An analysis of feature distributions in twitter. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; pp. 293–301.
13. Parker, C.J.; May, A.; Mitchell, V. Relevance of volunteered geographic information in a real world context, In Proceedings of the GISRUK 2011 Conference, Portsmouth, UK, 26–29 April 2011.
14. Flanagin, A.J.; Metzger, M.J. The credibility of volunteered geographic information. *GeoJournal* **2008**, *72*, 137–148. [CrossRef]
15. Cowan, T. A Framework for Investigating Volunteered Geographic Information Relevance in Planning. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2013.
16. Koswatte, S.; McDougall, K.; Liu, X. VGI and crowdsourced data credibility analysis using spam email detection techniques. *Int. J. Digit. Earth* **2017**, 1–13. [CrossRef]
17. Raper, J. Geographic relevance. *J. Doc.* **2007**, *63*, 836–852. [CrossRef]
18. Cai, G. GeoVSM: An integrated retrieval model for geographic information. In *International Conference on Geographic Information Science (GIScience 2002)*; Egenhofer, M.J., Mark, D.M., Eds.; Springer: Boulder, CO, USA, 2002; pp. 65–79.
19. Mobasheri, A. A rule-based spatial reasoning approach for OpenStreetMap data quality enrichment; case study of routing and navigation. *Sensors* **2017**, *17*, 2498. [CrossRef] [PubMed]
20. White, H.D. Relevance theory and citations. *J. Pragmat.* **2011**, *43*, 3345–3361.
21. Saracevic, T. Relevance reconsidered. In Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2), Copenhagen, Denmark, 13–16 October 1996; pp. 201–218.
22. MacEachren, A.M.; Jaiswal, A.; Robinson, A.C.; Pezanowski, S.; Savelyev, A.; Mitra, P.; Zhang, X.; Blanford, J. Senseplace2: Geotwitter analytics support for situational awareness. In Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), Providence, RI, USA, 23–28 October 2011; pp. 181–190.
23. Borlund, P. The concept of relevance in IR. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 913–925. [CrossRef]
24. Larson, R.R. Geographic information retrieval and spatial browsing. In Proceedings of the 1995 Clinic on Library Applications of Data Processing, Urbana, IL, USA, 10–12 April 1995; Smith, L.C., Gluck, M., Eds.; University of Illinois at Urbana–Champaign: Champaign, IL, USA, 1996.
25. Andrade, L.; Silva, M.J. Relevance Ranking for Geographic IR. In Proceedings of the Workshop on Geographic Information Retrieval, Seattle, WA, USA, 10–11 August 2006.
26. De Sabbata, S.; Reichenbacher, T. A probabilistic model of geographic relevance. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Zurich, Switzerland, 18–19 February 2010; p. 23.
27. Janowicz, K.; Raubal, M.; Kuhn, W. The semantics of similarity in geographic information retrieval. *J. Spat. Inf. Sci.* **2011**, *2011*, 29–57. [CrossRef]
28. Kumar, C. Relevance and ranking in geographic information retrieval. In Proceedings of the Fourth BCS-IRSG conference on Future Directions in Information Access, Koblenz, Germany, 31 August 2011; pp. 2–7.
29. Wang, C.; Xie, X.; Wang, L.; Lu, Y.; Ma, W.Y. Detecting geographic locations from web resources. In Proceedings of the Workshop on Geographic Information Retrieval, Bremen, Germany, 31 October–5 November 2005; pp. 17–24.
30. Jones, C.B.; Purves, R.S. Geographical information retrieval. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 219–228. [CrossRef]
31. Jones, C.B.; Alani, H.; Tudhope, D. Geographical information retrieval with ontologies of place. In *Spatial Information Theory*; Springer: London, UK, 2001; pp. 322–335.
32. 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, Sheffield, UK, 25–29 July 2004; pp. 273–280.

33. Zaila, Y.L.; Montesi, D. Geographic information extraction, disambiguation and ranking techniques. In Proceedings of the 9th Workshop on Geographic Information Retrieval, Paris, France, 26–27 November 2015; pp. 1–7.
34. Purves, R.S.; Clough, P.; Jones, C.B.; Hall, M.H.; Murdock, V. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Found. Trends Inf. Retr.* **2018**, *12*, 164–318. [CrossRef]
35. 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, Lisbon, Portugal, 6–10 November 2007; pp. 49–54.
36. Tomaszewski, B.; Blanford, J.; Ross, K.; Pezanowski, S.; MacEachren, A.M. Supporting geographically-aware web document foraging and sensemaking. *Comput. Environ. Urban Syst.* **2011**, *35*, 192–207. [CrossRef]
37. Tomaszewski, B.M.; MacEachren, A.M.; Pezanowski, S.; Liu, X.; Turton, I. Supporting humanitarian relief logistics operations through online geo-collaborative knowledge management. In Proceedings of the 2006 International Conference on Digital Government Research, San Diego, CA, USA, 21–24 May 2006; pp. 358–359.
38. Martins, B.; Silva, M.J.; Andrade, L. Indexing and ranking in Geo-IR systems. In Proceedings of the Workshop on Geographic Information Retrieval, Bremen, Germany, 4 November 2005; pp. 31–34.
39. Stowe, K.; Paul, M.; Palmer, M.; Palen, L.; Anderson, K. Identifying and Categorizing Disaster-Related Tweets. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 1–5 November 2016; pp. 1–6.
40. Monteiro, B.R.; Davis, C.A.; Fonseca, F. A survey on the geographic scope of textual documents. *Comput. Geosci.* **2016**, *96*, 23–34. [CrossRef]
41. Alexopoulos, P.; Ruiz, C.; Villazon-terrazas, B. KLocator: An Ontology-Based Framework for Scenario-Driven Geographical Scope Resolution. *Int. J. Adv. Intell. Syst.* **2013**, *6*, 177–187.
42. Leidner, J.L.; Lieberman, M.D. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Spec.* **2011**, *3*, 5–11. [CrossRef]
43. Koswatte, S.; McDougall, K.; Liu, X. Semantic Location Extraction from Crowdsourced Data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, 543–547. [CrossRef]
44. Frontiera, P.; Larson, R.; Radke, J. A comparison of geometric approaches to assessing spatial similarity for GIR. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 337–360. [CrossRef]
45. Lieberman, M.D.; Samet, H.; Sankaranarayanan, J.; Sperling, J. STEWARD: Architecture of a spatio-textual search engine. In Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, Seattle, WA, USA, 7–9 November 2007; p. 25.
46. Inkpen, D. Information Retrieval on the Internet. 2007. Available online: http://www.site.uottawa.ca/diana/csi4107/IR_draft.pdf (accessed on 05 December 2015).
47. Buckland, M.; Gey, F. The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* **1994**, *45*, 12. [CrossRef]
48. Criscuolo, L.; Carrara, P.; Bordogna, G.; Pepe, M.; Zucca, F.; Seppi, R.; Oggioni, A.; Rampini, A. Handling quality in crowdsourced geographic information. In *European Handbook of Crowdsourced Geographic Information*; Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., Purves, R., Eds.; Ubiquity Press: London, UK, 2016; pp. 57–74.
49. Spinsanti, L.; Ostermann, F. Validation and relevance assessment of volunteered geographic information in the case of forest fires. In Proceedings of the Validation of Geo-Information Products for Crisis Management Workshop (ValGeo 2010), Ispra, Italy, 11–13 October 2010.
50. Cambria, E.; Rajagopal, D.; Olsher, D.; Das, D. Big social data analysis. *Big Data Comput.* **2013**, *13*, 401–414.
51. Barbier, G.; Zafarani, R.; Gao, H.; Fung, G.; Liu, H. Maximizing benefits from crowdsourced data. *Comput. Math. Organ. Theory* **2012**, *18*, 257–279. [CrossRef]
52. Lewis, S.C.; Zamith, R.; Hermida, A. Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *J. Broadcast. Electron. Media* **2013**, *57*, 34–52. [CrossRef]
53. Okolloh, O. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Particip. Learn. Action* **2009**, *59*, 65–70.
54. Potts, M.; Lo, P.; McGuinness, R. *Ushahidi Queensland Floods Trial Evaluation Paper: A Collaboration between ABC Innovation and ABC Radio*; ABC Australia: Ultimo, Australia, 2011.

55. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
56. Liu, T.-Y. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.* **2009**, *3*, 225–331. [CrossRef]