Creating corroborated crisis reports from social media data through formal concept analysis

ANDREWS, Simon <http://orcid.org/0000-0003-2094-7456>, GIBSON, Helen <http://orcid.org/0000-0002-5242-0950>, DOMDOUZIS, Konstantinos <http://orcid.org/0000-0003-3679-3527> and AKHGAR, Babak <http://orcid.org/0000-0003-3684-6481>

Published version

ANDREWS, Simon, GIBSON, Helen, DOMDOUZIS, Konstantinos and AKHGAR, Babak (2016). Creating corroborated crisis reports from social media data through formal concept analysis. Journal of Intelligent Information Systems, 47, 287-312.

J Intell Inf Syst
DOI 10.1007/s10844-016-0404-9

Creating corroborated crisis reports from social media data through formal concept analysis

Simon Andrews[1,2] · **Helen Gibson**[1] · **Konstantinos Domdouzis**[1] · **Babak Akhgar**[1]

Received: 25 September 2015 / Revised: 20 March 2016 / Accepted: 28 March 2016

Abstract During a crisis citizens reach for their smart phones to report, comment and explore information surrounding the crisis. These actions often involve social media and this data forms a large repository of real-time, crisis related information. Law enforcement agencies and other first responders see this information as having untapped potential. That is, it has the capacity extend their situational awareness beyond the scope of a usual command and control centre. Despite this potential, the sheer volume, the speed at which it arrives, and unstructured nature of social media means that making sense of this data is not a trivial task and one that is not yet satisfactorily solved; both in crisis management and beyond. Therefore we propose a multi-stage process to extract meaning from this data that will provide relevant and near real-time information to command and control to assist in decision support. This process begins with the capture of real-time social media data, the development of specific LEA and crisis focused taxonomies for categorisation and entity extraction, the application of formal concept analysis for aggregation and corroboration and the presentation of this data via map-based and other visualisations. We demonstrate that this novel use of formal concept analysis in combination with context-based entity extraction has the potential to inform law enforcement and/or humanitarian responders about on-going crisis events using social media data in the context of the 2015 Nepal earthquake.

Keywords Formal concept analysis · Crisis management · Disaster response · Visualisation · Entity extraction

✉ Helen Gibson
h.gibson@shu.ac.uk

[1] CENTRIC, Sheffield Hallam University, 153 Arundel Street, Cantor Building, Sheffield, S1 2NU, UK

[2] The Conceptual Structures Research Group, Department of Computing, Sheffield Hallam University, Sheffield, UK

Published online: 06 May 2016

1 Introduction

The use of social media is now ubiquitous and, while the services and platforms used may vary, their overarching goal is the same: to “allow the creation and exchange of User Generated Content” (Kaplan and Haenlein 2010). As this ubiquity has increased, the potential for social media posts to be re-purposed for a number of other applications has increased. This includes sentiment analysis of major events, e.g., elections (Tumasjan et al. 2010), use in open-source intelligence (Omand et al. 2012), to track disease outbreaks (Brownstein et al. 2009), predict box-office hits (Asur and Huberman 2010) and for situational awareness during crises (Vieweg et al. 2010). Here we focus on its potential applications to crisis situations and, in particular, the complete process from data acquistion and information extraction to visualisation and analysis.

Because of this ubiquity, during a crisis people naturally reach for their smartphones to report, comment and explore information surrounding the crisis creating a large volume of social media data. Those interested in this data extends beyond the public and journalists to law enforcement agencies (LEAs), first and humanitarian responders. These services see untapped potential within this data, especially within command and control, and exploiting it has the capacity to increase the effectiveness of the crisis response. However, the sheer volume of social media output during a crisis situation—certain crisis situations have generated millions of tweets (Zubiaga and Ji 2014)—makes it impossible for decision makers to utilise it without help. Therefore we must use the computational power we possess to make the analysts’ jobs easier, to present them with aggregated data that draws out the key features and relationships in the data but leaves them with the final decision.

Achieving this does not just merely require collecting social media data or simply disseminating information on social media. In this paper we propose a workflow for extracting social media data and from it producing and visualising aggregated reports about specific events during a crisis. Bearing this in mind, we explain and develop a multi-stage process, set against the backdrop of real crisis tweets from the Nepal earthquake of 2015, which demonstrates how we can go from raw social media data to aggregated and specific crisis concepts explorable through a dashboard interface.

A key element of this process is the production of crisis concepts through formal concept analysis (FCA) (Ganter and Wille 1999). FCA uses a binary object-attribute matrix that denotes whether a particular attribute is present or absent for a particular object. This matrix is then used to group objects which share common attributes to produce a classification hierarchy. Thus we can use FCA to identify tweets which share similar or identical information.

The overall process uniquely combines the generation of attributes through the means of categorisation and context-based entity extraction before utilising them in a newly updated version of In-Close (Andrews 2011) developed specifically for the task of mining crisis concepts efficiently, to product formal (crisis) concepts (natural groupings of attributes) that are then mapped and visualised. Furthermore, little previous work has considered the effect of temporality in the construction and analysis of formal concepts when using real-time data, and within our analysis and presentation of the crisis concepts we explore some methods of introducing this temporality and avenues for future research.

The process proceeds as follows: firstly, relevant data from social media must be identified and captured in real-time. This data must then be classified and categorised to bring out keywords, themes and relationships corresponding to actions, events, locations, and time.

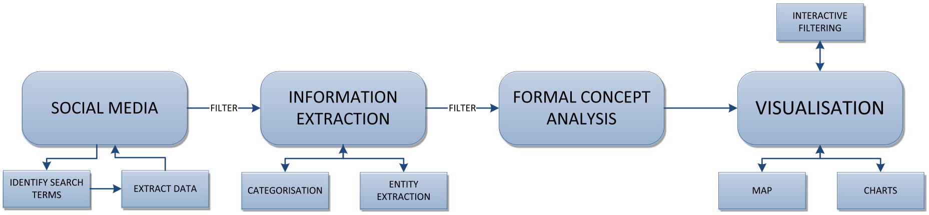


Fig. 1 The information processing pipeline taking raw social media data to visualised crisis-concepts

We then propose the application of a Formal Concept Analysis (FCA)-based solution. By reformatting the data so that each action, event, relationship, location and time period is converted into Boolean attributes, FCA can be applied to this data to create 'crisis concepts': aggregated information sources where each source contains the same crisis information. Consequently, the volume of information presented to the analyst is significantly reduced, and the information that is presented is information made more credible by the fact that it is corroborated by several sources. The final step is visualising the crisis concepts. In a crisis, probably the most crucial piece of information concerning an event is its location. By ensuring that locations are extracted during the information gathering process, crisis concepts can be presented as a map-based visualisation. Typical search and filter features can then easily be provided for the analyst to focus on particular types of information and thus further aid in the decision making process. This process is demonstrated in Fig. 1.

Section 2 reviews related work on crisis categorisation, entity extraction, and formal concept analysis. Section 3 outlines the process for creating crisis concepts while Section 4 demonstrates this in the context of the 2015 Nepal earthquake. Section 5 concludes the paper with a discussion on future directions.

2 Background and related work

2.1 Project Athena

Athena[1] is an EU FP7 funded project aiming to bring citizens, first responders and LEAs together to tackle crisis situations using social media and custom mobile applications. A key outcome will be a suite of prototype software tools that will support citizens, first responders and LEAs in achieving these goals. These components comprise a data processing centre, a mobile application and a command and control intelligence dashboard. An initial consideration of how FCA may be applied within the Athena project was given in (Andrews et al. 2013) and here we explore that potential with a larger and concrete dataset.

The envisaged Athena system has the following workflow. The crisis information processing centre scans the social media landscape for relevant posts about the ongoing crisis, these reports are augmented by citizen and first responder reports made via the mobile appli-

[1] http://www.projectathena.eu/

cation. These posts and reports are then filtered, analysed and aggregated and passed to the dashboard which is housed the strategic command and control centre during a major crisis event. The dashboard will present the information extracted from social media through a real-time crisis map whilst other parts of the interface will allow users to query and visualise the data to obtain better situational awareness of the current crisis. Completing the cycle between LEAs, first responders and back to citizens the dashboard will facilitate the updating of the mobile crisis map making it accessible to citizens and first responders who can then also track the crisis in real time.

In this paper, we restrict ourselves to presenting the workflow associated with FCA. Namely, the extraction of social media posts, categorisation and entity extraction, FCA and, finally, how this can be visualised on a dashboard. By combining these components together we demonstrate a novel mechanism to aggregate and display crisis related information from social media.

2.2 Crises and social media

The potential in social media data for enhancing the understanding and management of a crisis situation has been explored extensively. Due to the volumes of data produced we are concerned with efficient methods for data gathering, processing, filtering and analysis. Imran et al. (2015) have already conducted a recent comprehensive review of processing crisis-related social media messages that we will not replicate here. Instead we focus on the two key elements for our system: the taxonomies used to categorise and extract entities that form the basis of our own context-based entity extraction and the mapping and visualisation functionalities that will take the FCA output and present it via a dashboard interface.

2.2.1 Crisis taxonomies, ontologies and entity extraction

The nature of social media means that its content is often colloquial and unstructured. Thus, it is necessary to tease out information through a range of natural language processing and semantic techniques. A key part of this is the construction of taxonomies and ontologies defining the information to be extracted. However, despite its centrality to data extraction there is currently no standardised method of categorising crisis information (Imran et al. 2015); although, classification of social media posts in a crisis is usually a two-stage process: initially the relevance of the post is determined (Vieweg 2012; Imran et al. 2013; Olteanu et al. 2015) followed by the classification of these posts into more specific categories.

Within the crisis management literature there is a wide variety of suggestions for how social media content can be categorised. Imran et al. (2015) suggest six possible approaches for classifying information: by facts versus opinion, type of information, information source, credibility, time or location. Our categorisation taxonomy is aligned with the “type of information” style classifications which also means we concentrate on information that is factual. Most existing taxonomies focus on natural disasters and the humanitarian aspects of a crisis. Both the categorisation and concept extraction taxonomies developed as part of this research build on existing taxonomies but are constructed for the purpose of providing relevant information for LEAs. Thus they have a strong bias towards identifying particular incidents that may occur over the duration of a crisis event. Furthermore, as Athena is not aimed at one specific crisis but aims to cover multiple crises, the taxonomies also reflect this broadness and therefore cover naturally occurring events (such as natural disasters) and man-made events such as terrorism.

Vieweg (2012) defined three high level categories: social environment, built environment and physical environment divided into 32 sub-categories (such as caution and advice, causalities and damage, donations, people missing, found or seen, and information source). Okolloh (2009) also provide a set of 10 categories these, along with Vieweg's, provide the basis for many of the other taxonomical classifications. Olteanu et al. (2015) developed their categories along three high-level crisis dimensions—hazard, temporal development and geographic spread—which are sub-divided into more specific categories. For example, hazard is sub-divided into natural (disaster) and human-induced categories covering both intentional (e.g., terrorism, shootings) and accidental (e.g., building collapse). These classifications are based on the International Disaster Database[2]—which has two high level disaster categories (natural and technological) and those from ready.gov.[3] The category structure for Athena is heavily influenced by these Olteanu et al.'s taxonomy and that of the International Disaster Database.

Categorisation can be carried out by using rule-based and machine learning techniques. Machine learning methods include the use of the Latent Dirichlet Algorithm (Caragea et al. 2011), a Naive Bayes classifier (Imran et al. 2013), logistic regression, random forests and support vector machines (Cobo et al. 2015) as well as those employed by AIDR (Artifical Intelligence for Disaster Response) (Imran et al. 2014). D-Sieve (Roy Chowdhury et al. 2015) applies a post-processing solution in order to produce stable named entities and hashtags which improves classification accuracy while Pohl et al. (2012) carry out sub-event detection to find crisis event hot-spots associated with Flickr and YouTube videos through TF-IDF and clustering by self-organising maps.

There is also the Management of a Crisis (MOAC) ontology[4] (Limbu et al. 2012; Ortmann et al. 2011), the integrated data for events analysis (IDEA)[5] coding for political, economic and social events for GEDLT[6] and the humanitarian data exchange (Keßler and Hendrix 2015) which provides a way of marking up simple spreadsheet data with specific tags to indicate language, places, people, crisis events, dates.

Another method is to have specific vocabulary that can be used to assign crisis data to particular crisis categories. By using VerbNet (Kipper et al. 2008), Vieweg linked certain verbs and sentence structure with categories while Olteanu et al. (2014) defined a crisis lexicon. Temnikova et al. (2015) have gone further by identifying 7000 terms that are routinely used on Twitter in relation to 23 different categories.

Beyond categorisation is information extraction. Imran et al. (2013) call these 'information nuggets' and include locations, times, objects and numbers. Nguyen et al. (2015) are able to summarise tweets containing actionable information that would answer the questions 'what?', 'where?' and 'how many?'. We perform similar information extraction but using custom built entity extraction rules that rely not only on the presence of an attribute but take into account the context of the entities being extracted and only extract entities when certain pre-conditions are met.

[2] http://www.emdat.be/classification

[3] http://www.ready.gov/be-informed

[4] http://www.observedchange.com/moac/ns/

[5] http://vranet.com/IDEA.aspx

[6] http://gdeltproject.org/

Our taxonomies have used these categories as inspiration but also needed to include incidents that are born of man-made scenarios such as terrorist attacks. We take a lexicographical approach to extracting entities and matching categories rather than a machine learning based one and these taxonomies are elaborated on further in Section 3.1.

2.2.2 Crisis maps and systems

A common method for dealing with crisis data extracted from social media is to place the data on to a map. Meier (2012) describes having a map as being "almost as good as having your own helicopter" as it provides "...immediate situational awareness, a third dimension, and additional perspective, a bird's eye view of events unfolding in time and space". However, Shanley et al. (2013) add a note of caution (and their concerns apply more widely to the reuse of social media data) by saying that future adoption of crisis mapping technologies need to overcome issues such as trust and credibility, privacy and security concerns, intellectual property rights, data access and distribution policies as well as liability and ethics issues. These are all issues that we are trying to tackle as part of the wider goals of the Athena project.

Goolsby (2010) suggests that different maps, or at least map layers should be available for different responders, for example citizens may require different information to first responders or LEAs. Birregah et al. (2012) take this further by defining a process that combines text mining with semantic approaches to generate a multi-layer crisis map where different layers are made available to different organisations. Currently most maps developed only have a single public facing layer used to assist in humanitarian relief efforts.

Ushahidi[7] (Okolloh 2009) was one of the first platforms deployed for crisis mapping. Ushahidi began by allowing people to send SMS messages to a specific Ushahidi number or enter reports via the website so that those messages could be verified by Ushahidi volunteers and plotted on the crisis map. Ushahidi was implemented during violence post the Kenyan elections of 2008 (Meier 2012) and in the aftermath of the 2010 Haiti earthquake (Heinzelman and Waters 2010). Ushahidi is still in development today and is provided as a standalone platform (as was done for the Nepal earthquake) or as a paid version. Crisis mapping often takes advantage of crowdsourcing; the Micromappers platform (MacKenzie 2013; Meier et al. 2014) asks volunteers to tag tweet text, images and videos for relevance and appropriate categorisation. Once tagged these are then placed on crisis maps which is exposed to citizens and humanitarian organisations. EmergencyAUS[8] maps incidents around Australia using official public information including emergency incidents, warnings, perimeters, earthquakes, weather warnings and traffic incidents. Users can make reports as well as verifying others' observations and marking them as helpful/unhelpful.

Tracking the progression of a crisis on social media over time is provided by CrisisTracker (Rogstadius et al. 2013) which provides a dashboard map and filtering using tags, named entities, and time as well as providing tweet statistics. TweetTracker (Kumar et al. 2011) provides similar functionality with filters, rankings, map-based visualisations, and a keyword trending engine while TwitInfo (Marcus et al. 2011) is real-time tweet exploration system that provides visualisation and summarisation of tweet data including maps, sub-event detection by monitoring tweet bursts, tracking of sentiment and presentation of

[7] http://www.ushadidi.com

[8] http://www.emergencyaus.info/

relevant tweets in a list style. Twitcident (Abel et al. 2012) is more LEA focused and monitors tweets on given topics and then visualisations them as lists, on maps, and charts of the number of mentions of entities. These can also all be viewed over time to monitor how the crisis is progressing.

Most of these crisis systems and applications only map the data they receive directly and do not include a flow of sophisticated pre-processing and analysis before mapping these posts. A key and novel feature of this implementation and the Athena project as a whole is the incorporation of social media data, crisis taxonomies, entity extraction and the application of FCA before the mapping and other visualisation takes place whereas many systems stop after the entity extraction phase or they miss it out completely and simply map the raw social media information. The inclusion of FCA is crucial to the data aggregation process and is introduced in the next section.

2.3 Formal concept analysis

Formal Concept Analysis (FCA) was proposed by Wille and Gartner in the 1990s (Ganter and Wille 1999; Wille 2005). FCA is a method for deriving a hierarchical classification of objects based on a set of binary attributes. The hierarchy begins with a concept containing the set all objects and no attributes and filters down to reach a final concept which contains the set of all attributes and no objects (unless there is an object that contains all attributes). In between tthere are a number of hierarchically structured groups of objects and attributes each one known as a formal concept.

To compute a concept hierarchy, one must first have a data set consisting of objects and their attributes. Each of these attributes are described in a binary format, that is, either an object has the attribute or it does not. This means that each object corresponds to a row of a binary object-attribute matrix where each column represents a single attribute.

This process can be described formally as having a set of objects G with a set of attributes M such that the binary relation $I \subseteq G \times M$. This means that for a specific object $g \in G$ with the attribute $m \in M$ the relationship gIm implies that object g has attribute m. If we then take a subset of objects $A \subseteq G$ then by using the derivation operator $'$ to define the set of attributes that are shared by all objects in A such that,

$$A' = \{j \in M \| \forall i \in A : iIj\}. \tag{1}$$

This means that given the set of object A, the set of attributes A' are all the attributes that are common to the attributes in A. Similarly, for a set of attributes $B \subseteq M$ the derivation operator is used to define the set of objects with the attributes in B as follows,

$$B' = \{i \in G | \forall j \in B : iIj\}. \tag{2}$$

That is given the set of attributes B the set of objects B' is the set of all objects possessing all the attributes in B. The set $\{A, B\}$ is then considered a formal concept if $A = B'$ and $B = A'$. In this case we would call A the extent and B the intent.

The computation of these concepts induces a hierarchy such that the further down the hierarchy one travels the more specialised each formal concept becomes due to the addition of further attributes to the concept. In this case the number of objects belonging to each concept will also decrease. The resulting list of concepts can be used to understand how attributes are grouped in the data and how common the appearance of that combination of attributes is.

In the context of our crisis application, G would represent the set of all tweets and M would represent the set of all attributes (as will be defined by our taxonomy). These attributes include the set of all categories as well as the entities we will extract. For example, A could be the attributes earthquake and avalanche and B' would be all of the tweets that match that contain those attributes.

FCA has been applied to a number of different applications areas. One of its original applications was to software maintenance and identifying classes in object oriented software (Tilley et al. 2005) but it has also been used to identify software faults (Cellier et al. 2008) and to identify which developer should be assigned to fix a software bug (Wermelinger et al. 2009).

Text processing is another typical application of FCA including analysis of text corpora in tourism, finance and real estate adverts (Cimiano et al. 2005; Cole and Eklund 2001) as well as knowledge discovery from databases, document and email repositories (Poelmans et al. 2010; Stumme et al. 1998; Cole et al. 2003), content based retrieval (Jay et al. 2008), exploration of social and web communities (Rome and Haralick 2005; Jay et al. 2008), uses in the semantic web domain (d'Aquin and Motta 2011; Beydoun 2009) and as part of a collaborative recommendation system (du Boucher-Ryan and Bridge 2006).

FCA has also been used to explore biological data. These applications include the analysis of gene expression data (Kaytoue et al. 2009, 2011), the identification of biomarkers for breast cancer (Motameny et al. 2008), enzyme classification (Coste et al. 2014) and the tracking of ecological traits within particular species (Bertaux et al. 2009). FCA has also been used in the detection of organised crime threats (Brewster et al. 2014), incidents of domestic violence (Poelmans et al. 2009) and terrorist threats (Elzinga et al. 2010).

Here we will use FCA to try to classify and interpret crisis related data. The next section will look at the construction of two crisis taxonomies which will provide the basis of the attributes to be used in our FCA application.

3 Creation of corroborated crisis concepts

3.1 Crisis taxonomies

Given the unstructured nature of most social media posts sophisticated natural language processing techniques are required to in order to extract meaning and assist in the sense-making process.

Since no two crisis are the same and that social media data used from one crisis for machine learning based activities do not necessarily apply effectively to another crisis (Imran et al. 2014) our taxonomies do not depend on past crisis data. Rather they are formed from a complex set of boolean rules that first, if possible, put each post into one more categories and secondly, extract a number entities based on their context, i.e., for the most part we do not only extract named entities but we extract these entities only when they appear in a particular context; also known as 'atomic fact extraction' (Chakraborty et al. 2014) or template-based information extraction. Chambers and Jurafsky (2011), and Lee and Bui (2000) explored some disaster related templates as part of their work but we believe that the rules we have constructed cover a wider range of crisis events and are the first to be combined with FCA to produce crisis reports.

Both the categorisation and entity extraction processes are supported by SAS Content Categorisation Studio (SAS 2015). This software provides a framework for developing rule based models for categorisation and entity extraction which integrates directly with SAS's Information Retrieval Studio interface that combines crawling of data from social media with instant categorisation and entity extraction. In this application the data generated by

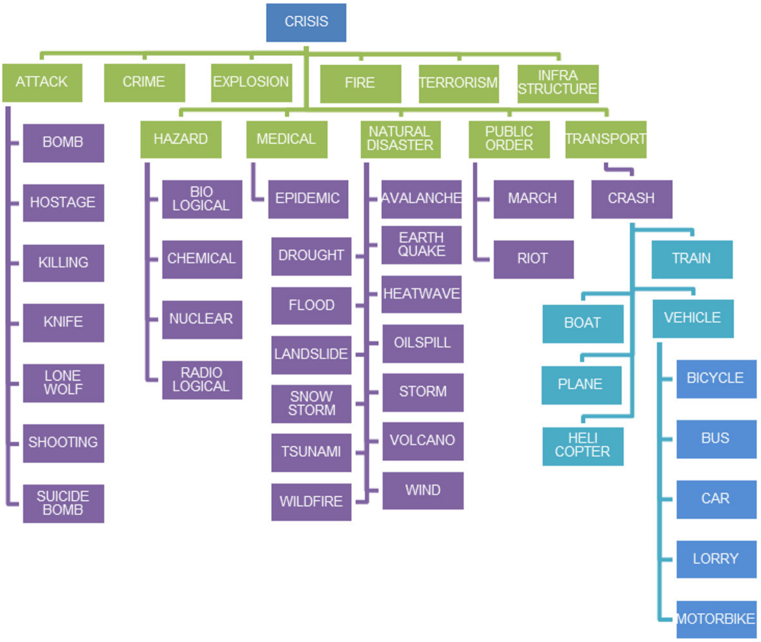


Fig. 2 The Athena crisis categorisation taxonomy

passing the social media data through the crisis taxonomies will form the attributes that are needed for our formal concept analysis.

3.1.1 Categorisation taxonomy and rule creation

Categorisation is the process of applying one or more labels to a document based on its content. Building on the categorisation taxonomies, as discussed in Section 2.2.1 we have developed a multi-level hierarchical taxonomy that aims to categorise individual tweets into particular crisis categories. As can be seen in Fig. 2, the taxonomy has a number of higher level categories which are then split up into a number of sub-level categories. The categories were developed based on those from Olteanu et al. (2015) and those from the International Disaster Database[9] customised to agree with the ambitions of Athena. The top-level of the taxonomy is composed of 11 categories: attack, crime, explosion, fire, terrorism, infrastructure, hazard, medical, natural disaster, public order and transport. Some of these categories are then further sub-divided into more specific sub-categories.

To be compatible with future versions of the system the top level of the categorisation hierarchy corresponds directly with the categories that the user is able select from when they submit a report via the Athena mobile app.

Each category is described by a series of pre-defined boolean rules, one of which a post must match in order to be given that particular category label. These generation of these rules consisted of the identification of textual clues which would indicate that the particular piece of text, in this case a tweet, is about that particular category. These rules can be a

[9] http://www.emdat.be/classification

simple one word match—such as any post containing the word earthquake is automatically added to the earthquake category as well as being more complex, for example matching two or more of the terms 'shaking', tremor', 'epicenter', 'aftershock' or other terms that are usually associated with an earthquake. Each post is not limited to matching only one category; however, due to the brevity of most social media postings usually only one or two categories are matched. Each category that appears in the document set may then be considered an attribute in the formal context.

3.1.2 Context-based entity extraction

Although categorisation is able to give an overview, categories are not able to identify specific information given in the tweet. In order to achieve this we implement a specific form of entity extraction where each entity is only extracted should it appear in a specific context, such as how it appears in a sentence structure. Therefore, we define a second entity extraction taxonomy designed to extract entities relating to the crisis from current tweets. We aim to extract five specific types of information: locations (i.e., where), people (divided into actions against people, and actions carried out by people) (i.e, who and by-who), actions relating to objects and actions themselves (i.e., what). Each of these entities is accompanied by an event. For example, the event 'avalanche' could be accompanied by a location. A shooting event may be accompanied by a location, a person entity identifying who has carried out the shooting and a second person object identifying who has been shot. The events and the associated entities that can be extracted alongside them are shown in Fig. 3. We have chosen these specific events as they match many of the categories we have chosen but enable us to extract some additional information alongside them where available. Furthermore, as the results of this process would be used in command and control centre

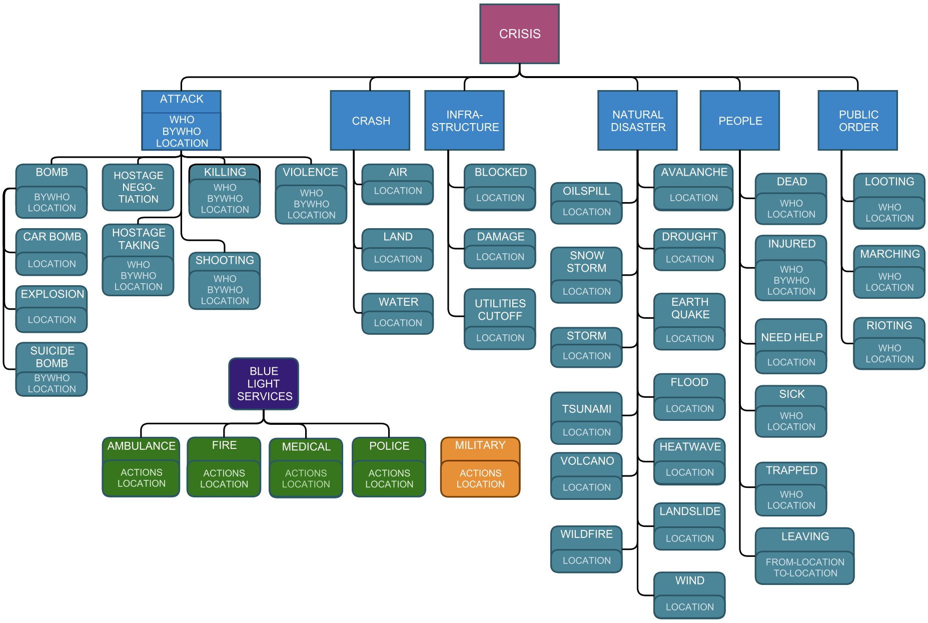


Fig. 3 The Athena context-based entity extraction taxonomy. Each *box* defines an event and, if defined, the types of entities that can be extracted with that event

for LEAs we try to extract information that would also be especially relevant to them such as the actions of people. We note that there are currently some inconsistencies between the categorisation and concept extraction taxonomies and we acknowledge that further work is required to harmonise these taxonomies.

The rules for the context-based entity extraction are are constructed similarly to those defined for categorisation except that this time we list a set of terms, often verbs, that define an action or an event as well as a list of nouns that define people, locations, and objects or we define the rule such that we search only for a noun in a specific context. For example, if we have a preposition such as ‘in’, ‘at’, ‘by’ or ‘near to’ we know that each of these words may be followed by a noun that defines a location and so we simply extract that noun. For example, when detecting the location of an earthquake we define a number of terms associated with an earthquake including earthquake, seismic activity, tremor, aftershock, epicenter and Richter scale. One of these terms must then appear within the same sentence as a location and that location must be preceded by one of the location related prepositions as listed above. Thus the rule is written as below. These rules are written for each of the entities that we want to extract.

```
CONCEPT_RULE : (
  SENT , EARTHQUAKE–TERMS ,
    (ORDDIST_4 , (OR, by , at , in , near , next to ,
    beside , inside , outside , close to , on ,
    under , opposite , in front of ) ,
      _c {LOCATION } ) )
```

In writing the rules this way we ensure that entities are only extracted when they appear in a specific context, i.e., in conjunction with a very specific event—in the example above that is an earthquake. There are few applications that use this type of entity extraction, with most relying on machine learning approaches; however, we feel that this approach gives us greater control over the entities extracted and is advantageous for FCA.

3.2 FCA and crisis concepts

In this article we use an FCA solution to analyse and extract important information gathered from tweets around a crisis. The background to formal concept analysis (FCA) was described in Section 2.3. Here we propose to use FCA by defining each post made to social media, in this case a single tweet, as an object and using the output from the categorisation and contextual extraction processes as the attributes. Once extracted, this data can then be coerced into a boolean object-attribute format that can be easily converted into a CXT file on which the FCA algorithm can run and the concepts can be exported. This section introduces the software In-Close—a fast formal concept miner and an adaptation to it for the benefit of this project in order to enable it to output formal ‘crisis concepts’.

3.2.1 In-Close

In-Close[10] is a fast formal concept miner (Andrews 2011) that is able to quickly calculate the formal concepts from large context files with hundreds of thousands of objects and thousands of attributes. In other words, capable of handling the volume of data that might be expected from crawling social media. In-Close is currently a command-line tool that takes

[10] http://sourceforge.net/projects/inclose/

a CXT file as input along with the ability to define the minimum intent and the minimum extent. In-Close uses a partial-closure canonicity test to avoid computation of repeated concepts (Andrews 2014), thus making it faster than other approaches in mining clusters from data (Andrews 2015).

In-Close had below-par performance when presented with multiple objects with similar attribtues. Part of the innovation of the Athena project was to eradicate this from In-Close and establish it as a fast-formal concept miner for all forms of data. This was achieved by introducing a new row sorting algorithm which takes advantage of Bentley and McIlroy's (1993) three way partitioning rather than the standard quicksort. This significantly improved In-Close's sort time for data with repeated attributes and is one of the innovations utilised in this paper.

Furthermore, In-Close did not originally out put the formal concepts it produced in a clear machine readable format. A key addition and contribution to In-Close for this work has been the ability to output crisis concepts using the structure of entities output from SAS and to do it in a JSON format. Moreover, the only options for controlling the concepts out put by In-Close prior to this work was the restrictions on the minimum extent (number of objects) or the minimum intent (number of attributes) that a particular formal concept must have. With the inclusion of the crisis concepts as described in the next section it is also now possible to restrict the concepts produced to having to have specific attributes as members. For example, the requirement that each concept should contain a location.

3.2.2 Crisis concepts

When the posts from social media are crawled they are run through SAS Information Retrieval Studio which is able to use the categorisation and context-based entity taxonomies as defined in the section above to mark-up the content with the matches made to the entries in these taxonomies. In converting output from SAS into the CXT format that can be used by In-Close the attributes either have the format: 'type_value' or 'event_type_value'. The first attribute format is used for defining categories, i.e., 'category_attack' whereas the second type of attributes are used to denote the named entities extracted and their context, for example, earthquake_loc_nepal has earthquake as the type of event, the type of named entity as location and that location as Nepal. For each type of event there may be a number of entity types and each entity type may have one or more values each defined as a separate attribute in the CXT file. Therefore as we construct the crisis concepts we need to take into account this structure.

The crisis concepts are output in JSON format and are constructed by examining each of the concepts generated through In-Close. For each of these concepts its constituent attributes are then examined. The attributes are then structured so that each event is grouped together with its associated types (location, who, bywho, action, what). The next section takes a concrete example of the Nepal earthquake beginning from the raw social media data through categorisation and entity extraction to the computation of the formal concepts and finally their visualisation on a crisis dashboard.

4 Case study: Nepal earthquake

At 11:56 NST (06:11 UTC) on 25th April 2015 a 7.8 magnitude earthquake hit Nepal. Its epicentre was said to be in small village Barpak in the Gorkha district. The earthquake caused devastation in areas such as the Kathmandu Valley. The severity of the earthquake meant that it triggered avalanches on Mount Everest and in the Langtang Valley. The earthquake also caused a further risk of landslides and resulted in a number of deaths, people being trapped under rubble and hundreds of thousands of people being displaced. A major aftershock also hit the region approximately 1 h later at 12:54 NST (7:09 UTC).

4.1 Background

As part of the live crisis response, a number of social media based and digital relief efforts were put into place. Facebook activated their safety check feature where people, detected as being in Nepal, could indicate that they were safe (Zuckerberg 2015). Google also activated their person finder functionality to allow individuals to report that they are safe or to search for the status of a person (Google 2015).

A number of crisis maps were produced in the initial response the earthquake these included QuakeMap,[11] the activation of the Micromappers platform to collect tweets[12] and images[13] while other maps showed images for a road survey,[14] geo-coded tweets with overlaid satellite imagery,[15] damage assessments,[16,17] route mapping indicating areas to avoid[18] and a series of maps showing numbers of injuries, deaths and buildings destroyed.[19]

4.2 Initial data processing

Between the 24 April and 27 April 2015 984,643 tweets were collected using the the Twitter Search API[20] via SAS Information Retrieval Studio with the search terms ‘nepalearthquake’, ‘katmandu’, ‘everest’ and ‘nepalquake’, Following a process where retweets were removed alongside tweets from before the earthquake occurred this resulted in a final set of 165,715 tweets. Each of these tweets were subjected to our categorisation and entity extraction schemes. In total 33 categories were detected. Table 1 shows the categories associated with 10 or more tweets. Unsurprisingly, earthquake and avalanche are the top categories.

This table also demonstrates one of the fallacies of natural language processing: the issue of word sense disambiguation. We see the category ‘shooting’ appearing quite commonly but on further examination of the source tweets it is clear that the words shot/shoots/shooting actual relate to the capture of photographs or video footage rather than the use of a gun. Strategies for attempting to tackle these mismatches is something we hope to rectify in future versions of our categorisation rules. Nevertheless, most categories are directly related to the type of information we would wish to collect—such as reports of infrastructure and medical issues. We also note that in our taxonomy the category ‘killing’ was intended to

[11] http://quakemap.org/

[12] http://maps.micromappers.org/2015/nepal/tweets/

[13] http://maps.micromappers.org/2015/nepal/images/

[14] http://www.nepalquake.org/status-map/

[15] http://vision.sdsu.edu/ec2/geoviewer/nepal-kathmandu#

[16] http://www.arcgis.com/home/webmap/viewer.html?webmap=9900baca7ab341a48834b81116a7f7ec&extent=85.148,27.7096,85.1901,27.7287

[17] https://geosprocket.cartodb.com/viz/bb5ed630-ee1e-11e4-8dca-0e018d66dc29/embed_map

[18] http://www.openrouteservice.org/disaster-nepal/

[19] http://www.codefornepal.org/2015/04/map-of-reported-nepalearthquake-deaths/

[20] https://dev.twitter.com/rest/public/search

Table 1 Number of tweets relating to each category

Crisis event	Count	Crisis event	Count
Earthquake	1484	Flood	19
Avalanche	1107	Crisis	18
Killing	1059	Explosion	18
Medical	371	Natural disaster	14
Infrastructure	177	Hostage	14
Landslide	65	Bomb	10
Shooting	32	Fire	10
Volcano	23		

represent some kind of attack but in fact here it is the earthquake or the avalanche which has done the killing.

Each tweet was also processed using our context-based entity extraction rules. We concentrated on extracting specific events and the entities associated with them. Those events that were detected more than 50 times are shown in Table 2.

Corresponding with these events are a number of entities. At this point some semi-automatic normalisation of entities took place using OpenRefine[21] standardising the data set so that places such as 'katmandu' and 'kathmandu' which have alternative spellings were consolidated into one entity. Entities that did not make sense, for example, a location erroneously appearing under the 'Who' type were also removed. This left us with a total of 339 locations, 50 people or organisations under the 'ByWho' entity, 210 people or organisations in the 'Who' entity, 41 entities for 'What' and eight in the 'Action' entity. In addition to these categories and entities we also extracted the time each tweet was made in UTC time and put these into bins of hour length. Thus in total there are 65 different time bins running from 06:00 on the 25th April to 22:00 on the 27th April.

Using these attributes we then created two CXT files: one which include time as an attribute and one which did not. We did this for two reasons: (1) we wanted to create a 'master' set of concepts that appear during the crisis that we could use to plot on the map but (2) we are also aware that in a live scenario the concepts produced may change over time as the crisis progresses and so we were interested in how this was represented in our dataset. Therefore our first CXT file without time included had 1489 attributes and 165,715 objects each of which represented an individual tweet. Our second dataset then had the additional 65 time attributes and, given that each object represents a specific tweet, only one of these time attributes is represented per object.

4.3 Nepal crisis concepts

As described in Section 3.2.2 the formal concept miner In-Close has been adapted to export the formal concepts created in a JSON format so that they can be used easily by web-based applications. In this case 4189 concepts are created from the file without time attribute while 14993 concepts are generated from the file with the time attributes. Analysing this data we find that 838 of the concepts that appear in the no time attribute dataset do not appear in

[21] http://openrefine.org/

Table 2 Number of times each context event is detected

Context event	Count	Context event	Count
Earthquake	1389	People trapped	227
People dead	1263	Medical	140
Avalanche	778	People sick	109
Damage	771	Blocked	104
Killing	558	Ambulance	92
Violence	444	People leaving from	80
People injured	278	Volcano	54
People need help	269	Hostage take	52

the time attribute dataset and, similarly, 177 'concepts'[22] that appear in the time attribute dataset do not appear in the dataset without the time attribute.

An example of a formal concept in this case could contain the **event people injured**, with one entity type **location**, which contains a single entity **Kathmandu**. There is then a further attribute containing the matched **categories** and and two further fields denoting the **number of sources** belonging to the concept and an array containing the identifiers of those sources. If this was a concept containing a time attribute then a further attribute would be included that would denote during which **hour** this concept appeared.

The next sections discuss how these concepts can be displayed in a dashboard interface that can be explored and used to gain situational information about the crisis.

4.4 Nepal FCA dashboard

Using the crisis concepts files generated our aim was then to produce a crisis dashboard interface that can be used explore the data and be used as a prototype interface for understanding how Athena might incorporate formal concepts as part of the final Athena dashboard

Within our dashboard there are four main components:

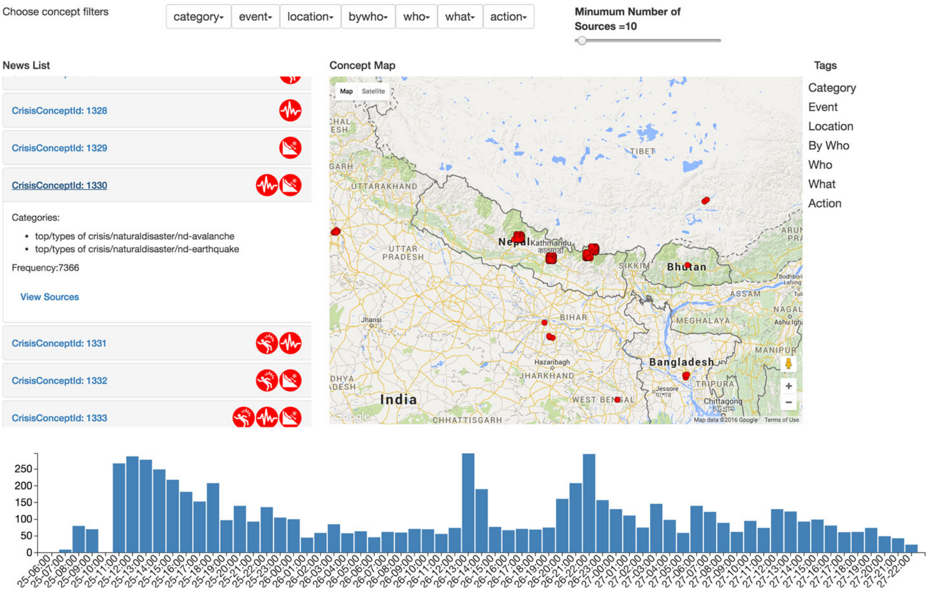
1. Crisis concept data filtering menu
2. Crisis concept map
3. Crisis concept list
4. Crisis concept timeline visualisation

The next sections explain the use and features of each of these components. Figure 4 shows an overview of the whole dashboard interface.

4.4.1 Filtering

FCA has enabled us to aggregate the data from individual tweets into formal concepts; however, even exploring these concepts solely via the map would be a cumbersome and time-consuming task. Therefore we provide some filtering options. The first allows the user

[22] In this case the actual number of concepts is greater than 177 as we consider two concepts as identical if all their attributes are the same except for the time attribute.



Choose concept filters

category▾ event▾ location▾ bywho▾ who▾ what▾ action▾

Minumum Number of Sources =10

News List

CrisisConceptId: 1328

CrisisConceptId: 1329

CrisisConceptId: 1330

Categories:

- top/types of crisis/naturaldisaster/nd-avalanche
- top/types of crisis/naturaldisaster/nd-earthquake

Frequency:7366

View Sources

CrisisConceptId: 1331

CrisisConceptId: 1332

CrisisConceptId: 1333

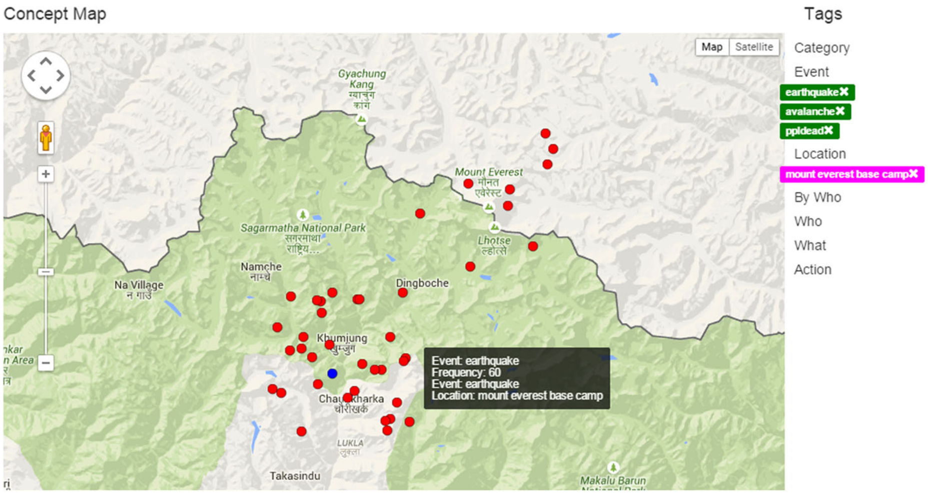
Concept Map

Tags

Category
Event
Location
By Who
Who
What
Action

Fig. 4 Overview of the Nepal concept explorer interface

to select one or more entities from the categories, event, locations, who, bywho and what filters. The attributes chosen are then shown in the tags column on the right-hand side of the map. Figure 5 shows the filtered map where three events have been chosen: **earthquake, avalanche and people-dead** alongside a single location attribute: **mount everest base camp**. The pins that match these concepts are shown on the map while all other pints are hidden. Currently, this filtering occurs on an 'AND' based operation.



Concept Map

Tags

Category
Event
earthquake✖
avalanche✖
ppldead✖
Location
mount everest base camp✖
By Who
Who
What
Action

Event: earthquake
Frequency: 60
Event: earthquake
Location: mount everest base camp

Fig. 5 Overview of the Nepal Concept Explorer Map where the pins are filtered based on the tags shown on the *right hand side*. More details about the content of the pin is shown in the tooltip

The second filtering option filters on the number of sources (i.e., tweets) that contribute to each formal concept with a default of ten sources. This option allows users to view only concepts that are more highly corroborated which may be beneficial in a crisis situation where a controller needs to view information that they think are most likely to be true so they can act quickly and accurately. These filtering options are provided in order to reduce the cognitive load on the analyst and enable them focus on specific pieces of information that may be relevant to them at that point.

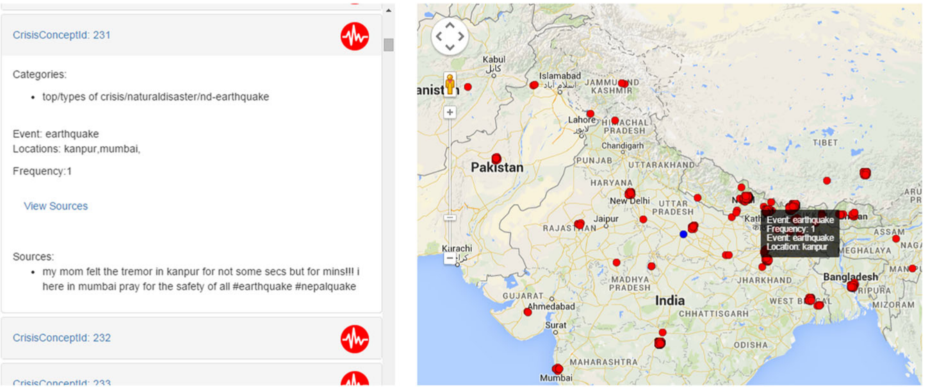
Future filtering options could include attribute filtering using an 'OR' operation to enable a user to see concepts across a number of attributes. Alternatively, attribute drop-down menus are currently statically built, if we filter by an 'AND' operation it may be useful then only show attributes where one or more crisis concepts would be returned by selecting that additional attribute. We are also be interested in exploring when formal concepts appear over time; therefore, the ability to filter on time ranges would also be useful. However, as discussed in the upcoming Section 4.5 there are some limitations associated with the time aspect of this FCA analysis.

4.4.2 Crisis map

For any situational awareness or decision support platform, including Athena, a crisis map is essential. A map facilitates an understanding of how the crisis is unfolding geographically. This is imperative for determining where and when resources are deployed and the available routes for getting there. Every tweet can include a geo-location; however, the amount of tweets with geo-locations attached are usually less than 1 % (Middleton et al. 2014) consequently we cannot rely on geolocation information for the mapping of the crisis concepts (although in future version we would like to explore how these can be incorporated into the concepts (see Gibson et al. 2014 for some initial thoughts). In this case we use pre-compute geo-coordinates based on the location attributes extracted during the contextual extraction phase.

Even with pre-computed geo-coordinates, plotting the crisis concepts required overcoming a number of issues. Firstly, each crisis concept may contain multiple locations meaning that there is not a one-to-one mapping between crisis concepts and pins. We chose to represent each crisis concept with multiple locations as multiple pins. In future versions we would like to develop a method for identifying when a crisis concept is composed of multiple pins. A second issue with not using geo-locations is that text-based locations usually refer to locations that cover a wide area, for example, Nepal, Kathmandu or Mount Everest . Therefore, mapping every pin to the Mount Everest location results in a significant number of pins being placed in exactly the same position. For this application we introduced a small amount of randomness to the placing of the pins. That is each pin appears within a radius of $\pm 0.1^\circ$ of its original geo-coordinate. This is purely for demonstration purposes and in the future we will be looking to map locations more accurately using geo-location information from reports. A final limitation is that not all concepts have any location data attached and thus cannot be mapped at all or that the location information that is extracted is too general, airports or train stations. These concepts currently only appear in the news list without having a corresponding pin on the map.

Figure 6 shows the map and news interface where no attributes have been added to the filter and the minimum number of sources is set to one. In this zoomed-out view we see a number of clusters of pins across the Indian Subcontinent each representing a concept and a location. These pins are not just all commenting on the earthquake but also people mentioning that they can feel the tremors of the earthquake further afield. Hovering over



CrisisConceptId: 231

Categories:

- top/types of crisis/naturaldisaster/nd-earthquake

Event: earthquake
Locations: kanpur,mumbai,

Frequency:1

View Sources

Sources:

- my mom felt the tremor in kanpur for not some secs but for mins!!! i here in mumbai pray for the safety of all #earthquake #nepalquake

CrisisConceptId: 232

CrisisConceptId: 233

Event: earthquake
Frequency: 1
Event: earthquake
Location: kanpur

Fig. 6 View of the news list after a pin has been selected and view sources has been expanded

a single pin gives brief information about the contents of the concept that it represents by identifying the event, the current location and the number of sources that pin has.

4.4.3 *Crisis concept new list*

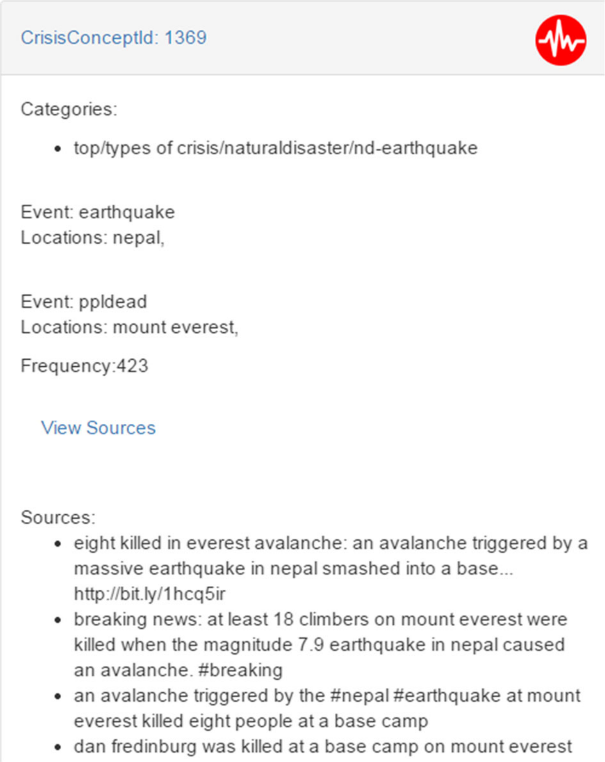
The crisis news list contains the text-based representation of each of the concepts. Each crisis concept is identified by its ID number in the heading alongside a visual icon for each category in the concept. The symbology is based on that provided by UN OCHA (OCHA 2012) with some additional symbols.

Clicking on the heading of a concept expands the concept so that its attributes are presented to the user. For each concept the categories are listed followed by each of the events included in the concept and their related entities. The number of sources are listed and the user has the option to expand the 'View Sources' panel so that they can view the full-text of the tweet which makes up each source. Figure 7 shows an example of a single news list item which has the earthquake symbol in its panel header, is involved in two events: earthquake and people dead that are listed in two different locations. The view sources panel is expanded to reveal the text of the first few tweets that are the objects for this particular concept.

Clicking on the header also has one additional effect: it updates the crisis timeline as described in the next section. An additional feature for future versions would be the ability to pan and zoom to the pin or pins on the map that this crisis concept represents.

4.5 Crisis concept timeline

The crisis concept timeline makes use of the second crisis concept file where each concept contains an additional time attribute. This means that where a crisis concept appears at multiple time periods we can track how the number of sources that make up these concepts can change over time. This enables us to track the emergence, the peak and the disappearance of certain events. Figure 8 demonstrates how a single concept can change over time. The height of each of the bars represents the number of sources that the concept has during that particular time period. Here we see that for the concept **people need help** and the location **nepal** that immediately after the earthquake struck appeals for help were limited but grew gradually over the hours and days after the earthquake struck.



CrisisConceptId: 1369

Categories:

- top/types of crisis/naturaldisaster/nd-earthquake

Event: earthquake
Locations: nepal,

Event: ppldead
Locations: mount everest,

Frequency:423

View Sources

Sources:

- eight killed in everest avalanche: an avalanche triggered by a massive earthquake in nepal smashed into a base... http://bit.ly/1hcq5ir
- breaking news: at least 18 climbers on mount everest were killed when the magnitude 7.9 earthquake in nepal caused an avalanche. #breaking
- an avalanche triggered by the #nepal #earthquake at mount everest killed eight people at a base camp
- dan fredinburg was killed at a base camp on mount everest

Fig. 7 Example of an expanded news list item

One of the issues with the crisis timeline is that using a separate file to generate a set of concepts with time attributes does not guarantee that the same set of concepts will be generated as were found from the main file. For example, if there are two concepts in the set of main concepts with attributes a, b and a, b, c respectively then in the list of concepts with time attributes there may be certain time points, say time t_3 where only the concept with attributes a, b, c appears. Therefore when we create the timeline graph for the concept with attributes a, b a gap appears at time t_3 where it appears that there are no sources with these attributes. However, it is just that all sources that appear at this time point that have the attributes a, b also have the attribute c and thus the concept with a, b as attributes is never generated. And, thus while this is technically correct, from an FCA standpoint, it gives a misleading view about the number of objects that have this set of attributes.

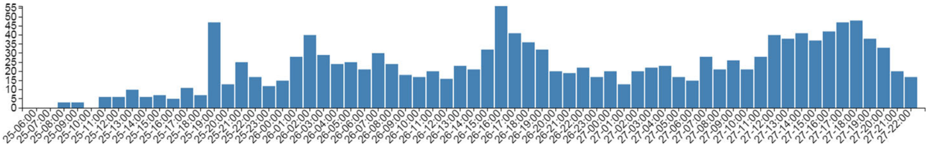


Fig. 8 Example of the crisis concept timeline for the concept with event—people need help and the corresponding location—Nepal

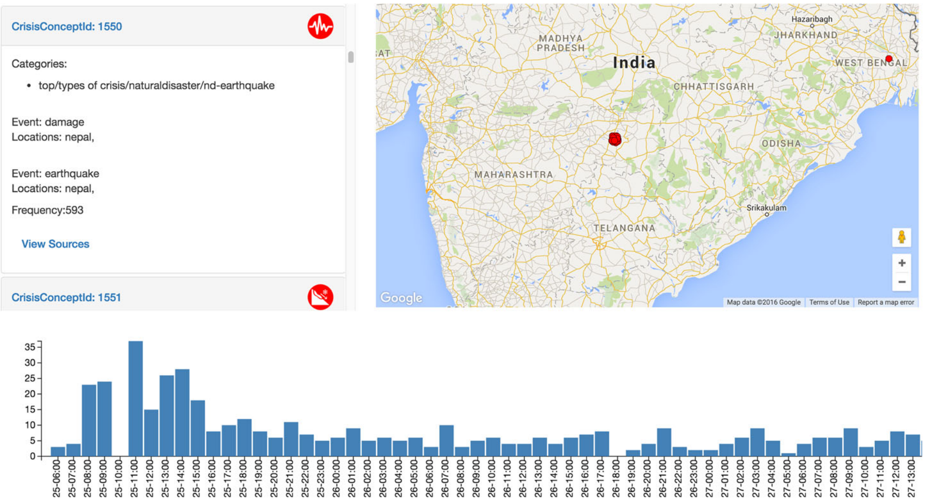
There are two potential strategies for overcoming this issue. The first would be to not create a second set of concepts using the time attributes but to use the sources assigned to each concept in the main concept file and match them to the time that they appear and perform the equivalent of a database ‘group by’ operation to determine how many of these appear in each time period and use that data to construct the graph. The second method would be a more FCA based solution where the concept lattice of the main concept file could be traversed so that if a concept was found not to have an entry for that specific time period the sub-concept(s) of that particular concept could be identified from the main set of concepts and then if the sub-concept appeared for the time-attribute concepts the the number of sources for that concept could be used instead.

4.6 Crisis concept exploration

The purpose of creating the categorisation and entity extraction rules, applying the rules to our social media data, extracting crisis concepts and visualising them is to improve the situational awareness and decision support capability of LEAs during a crisis situation. The previous sections have demonstrated how each of the components we have proposed work together as well as some of their limitations. We cannot examine the whole crisis within the confines of this paper; however, some examples of the type of information an analyst would have access to are described below.

Figure 6 demonstrates the extent of the impact of the earthquake and from that we can see individuals who have posts from small towns that are not at the centre of attention in the response but may still be in need of help. This includes reports about someone being stuck in the Langtang National Park and reports of a landslide in Sikkim.

We introduced the crisis concept timeline to illustrate how considering temporality as part of FCA enables us to understand how crisis events emerge and disappear. We can see from Fig. 9 that the mention of the damage caused by the earthquake in Nepal is very prevalent at the beginning of the crisis and then it tails off after the first few hours whereas Fig. 10 shows us that mentions of death are not high during the first few hours as reaches its peak about 15 h after the initial earthquake struck. A useful feature here would be an



CrisisConceptId: 1550

Categories:

- top/types of crisis/naturaldisaster/nd-earthquake

Event: damage
Locations: nepal,

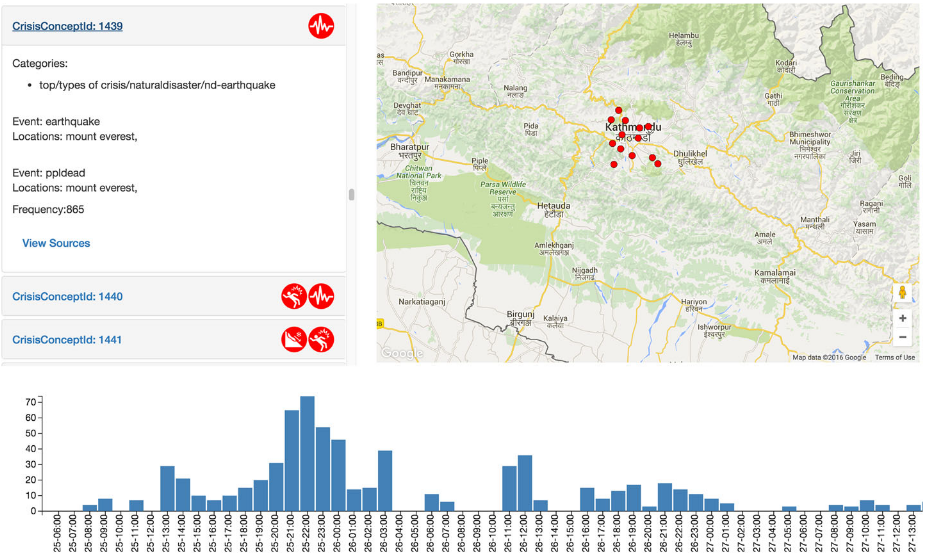
Event: earthquake
Locations: nepal,

Frequency:593

View Sources

CrisisConceptId: 1551

Fig. 9 Occurrence of crisis concepts mentioning damage, earthquake and Nepal over time



CrisisConceptId: 1439

Categories:

- top/types of crisis/naturaldisaster/nd-earthquake

Event: earthquake
Locations: mount everest,

Event: ppldead
Locations: mount everest,

Frequency:865

View Sources

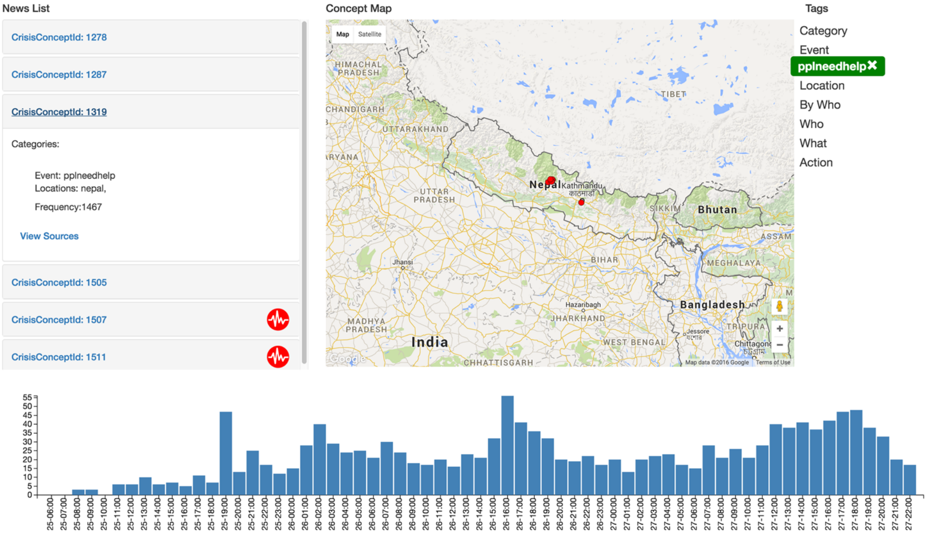
CrisisConceptId: 1440

CrisisConceptId: 1441

Fig. 10 Occurrence of crisis concepts mentioning deaths in Nepal due to the earthquake

alert that would highlight to the analyst when a particular crisis concept is suddenly much more highly corroborated than in the previous time period or that a new crisis concept has emerged.

A key element of the dashboard is enabling the analyst to drill down into specific information through the source and the attribute filters. Figure 11 shows an example of when we are looking for instances of where people are requesting help or it is being reported that they need help. We can see on the map there are a number of requests only mentioning Nepal



News List

CrisisConceptId: 1278

CrisisConceptId: 1287

CrisisConceptId: 1319

Categories:

Event: pplneedhelp
Locations: nepal,

Frequency:1467

View Sources

CrisisConceptId: 1505

CrisisConceptId: 1507

CrisisConceptId: 1511

Concept Map

Tags

Category
Event
pplneedhelp ✖
Location
By Who
Who
What
Action

Fig. 11 Highly corroborated crisis concepts containing the attribute ‘people need help’

as well as there being a specific request for help from the Kathmandu area. These requests must have a minimum of 50 sources in this case showing that they are highly corroborated. This filtering enables analysts to quickly identify where specific incidents are occuring and reduce the information overload that exists when viewing all concepts.

5 Discussion

In this article we have explored how crisis concepts can be generated using data from Twitter about a crisis event and passing them through our categorisation and context-based entity extraction taxonomies in order to create attributes for FCA. We have then shown how In-Close was optimised and adapted to out put formal crisis concepts which can then be displayed on a dashboard and explored through a series of filtering, map-based and visualisation operations. This is the first time that template-based entity extraction and formal concept analysis have been applied together to understand social media data within the crisis management domain. Using the lessons learned from this prototype we are now able to form the basis of a version of the Athena dashboard where those in command and control can monitor new concepts being created in a timely fashion and use these concepts to enhance their situational awareness knowing that the information extracted will be relevant for that particular crisis and that the number of sources cited for each concept gives a degree of corroboration and validity to the concept.

Nevertheless, before these concepts can be fully integrated into the Athena dashboard a number of issues still have to be resolved. Firstly, in this case we only used locations extracted from the text as our locations for the pins. This resulted in many pins being placed in the same position which was partially resolved by introducing a small amount of randomness into the final positioning; however, for concepts to be used in a fully live dashboard the locations need to be accurate as dashboard operators will be deploying first responders and resources to these areas and thus the location needs to be as precise as possible. Secondly, the pins used to represent the concepts are also an issue. Here we used only a simple pin type where all pins were the same size no matter which concepts they represented. Future improvements could include the ability to map pin size to the number of sources that it represents or to include this number as part of the pin's iconography, to include a category symbol as part of the pin head although in this case we would need to resolve which symbol would be used when a pin is assigned multiple categories. Another issue with the pins in this interface is the many to one relationship with a concept. Either a suitable method needs to be found to generate a simple one-to-one mapping (such as using the most specific location information we have available) or by identifying which pins form part of the same concept and to link them. Another option may be to extract the individual geo-locations from the raw tweet data and use that to plot the pins onto the map. The third main issue is how to construct the time based concepts to account for when a concept does not appear in a particular time bin despite there being objects with those attributes. Two potential solutions were discussed based on either calculating the number of objects for each concept in each time period on the fly using the master concept list or by utilising the lattice hierarchy to determine which objects should belong to concepts missing from certain time periods.

The next steps will be to look at how these results can be incorporated into the main Athena system and to work with analysts to understand whether the information provided is giving them the situational awareness that they require.

Acknowledgments This work is part of the Athena project, co-funded by the European Union Seventh Framework Programme SEC Call 1 - FP7-SEC-2012.6.1-30.

References

Abel, F., Hauff, C., Houben, G.J., Stronkman, R., & Tao, K. (2012). Twitcident: fighting fire with information from socialweb streams. In *Proceedings of the 21st international conference companion on World Wide Web—WWW '12 Companion* (pp. 305–308).

Andrews, S. (2011). In-close2, a high performance formal concept miner. In S. Andrews, S. Polovina, R. Hill, & B. Akhgar (Eds.) *Conceptual structures for discovering knowledge. Lecture notes in computer science* (Vol. 6828, pp. 50–62). Berlin: Springer.

Andrews, S. (2014). A partial-closure canonicity test to increase the efficiency of cbo-type algorithms. In N. Hernandez, R. Jäschke, & M. Croitoru (Eds.) *Graph-based representation and reasoning. Lecture notes in computer science* (Vol. 8577, pp. 37–50).

Andrews, S. (2015). A 'Best-of-Breed' approach for designing a fast algorithm for computing fixpoints of Galois Connections. *Information Sciences*, *295*, 633–649.

Andrews, S., Yates, S., Akhgar, B., & Fortune, D. (2013). The athena project: using formal concept analysis to facilitate the actions of responders in a crisis situation. In *Strategic intelligence management: national security imperatives and information and communication technologies* (pp. 167–180). Elsevier Amsterdam.

Asur, S., & Huberman, B. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT)* (Vol. 1, pp. 492–499).

Bentley, J.L., & McIlroy, M.D. (1993). Engineering a sort function. *Software: Practice and Experience*, *23*(11), 1249–1265.

Bertaux, A., Le Ber, F., Braud, A., & Trémolières, M. (2009). Identifying ecological traits: a concrete fca-based approach. In S. Ferré, & S. Rudolph (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 5548, pp. 224–236). Berlin: Springer.

Beydoun, G. (2009). Formal concept analysis for an e-learning semantic web. *Expert Systems with Applications*, *36*(8), 10952–10961.

Birregah, B., Top, T., Perez, C., Châtelet, E., Matta, N., Lemercier, M., & Snoussi, H. (2012). Multi-layer crisis mapping: a social media-based approach. In *Proceedings of the workshop on enabling technologies: infrastructure for collaborative enterprises, WETICE* (pp. 379–384).

Brewster, B., Andrews, S., Polovina, S., Hirsch, L., & Akhgar, B. (2014). Environmental scanning and knowledge representation for the detection of organised crime threats. In N. Hernandez, R. Jäschke, & M. Croitoru (Eds.) *Graph-based representation and reasoning. Lecture notes in computer science* (Vol. 8577, pp. 275–280). Berlin: Springer International Publishing.

Brownstein, J.S., Freifeld, C.C., & Madoff, L.C. (2009). Digital disease detection—harnessing the web for public health surveillance. *New England Journal of Medicine*, *360*(21), 2153–2157.

Caragea, C., Mcneese, N., Jaiswal, A., Traylor, G., Kim, H.W., Mitra, P., Wu, D., Tapia, A.H., Giles, L., Jansen, B.J., & Yen, J. (2011). Classifying text messages for the Haiti earthquake. In *Proceedings of the 8th international ISCRAM conference (May)*.

Cellier, P., Ducassé, M., Ferré, S., & Ridoux, O. (2008). Formal concept analysis enhances fault localization in software. In R. Medina, & S. Obiedkov (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 4933, pp. 273–288). Berlin: Springer.

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text mining and analysis: practical methods, examples, and case studies using SAS*. SAS Institute.

Chambers, N., & Jurafsky, D. (2011). Template-based information extraction without the templates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies—volume 1, association for computational linguistics, Stroudsburg, PA, USA, HLT '11* (pp 976–986). http://dl.acm.org/citation.cfm?id=2002472.2002595.

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, *24*(1), 305–339.

Cobo, A., Parra, D., & Navón, J. (2015). Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In *Proceedings of the 24th international conference on World Wide Web Companion, WWW '15 companion, international world wide web conferences steering committee* (pp 1189–1194).

Cole, R., & Eklund, P. (2001). Browsing semi-structured web texts using formal concept analysis. In H. Delugach, & G. Stumme (Eds.) *Conceptual structures: broadening the base. Lecture notes in computer science* (Vol. 2120, pp. 319–332). Berlin: Springer.

Cole, R., Eklund, P., & Stumme, G. (2003). Document retrieval for e-mail search and discovery using formal concept analysis. *Applied Artificial Intelligence*, *17*(3), 257–280.

Coste, F., Garet, G., Groisillier, A., Nicolas, J., & Tonon, T. (2014). Automated enzyme classification by formal concept analysis. In C. Glodeanu, M. Kaytoue, & C. Sacarea (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 8478, pp. 235–250). Berlin: Springer International Publishing.

d'Aquin, M., & Motta, E. (2011). Extracting relevant questions to an rdf dataset using formal concept analysis. In *Proceedings of the sixth international conference on knowledge capture, K-CAP '11* (pp. 121–128). New York: ACM.

du Boucher-Ryan, P., & Bridge, D. (2006). Collaborative recommending using formal concept analysis. *Knowledge-Based Systems*, *19*(5), 309–315.

Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., & Morsing, S. (2010). Terrorist threat assessment with formal concept analysis. In *2010 IEEE international conference on intelligence and security informatics (ISI)* (pp. 77–82).

Ganter, B., & Wille, R. (1999). *Formal concept analysis: mathematical foundations*. Berlin: Springer.

Gibson, H., Andrews, S., Domdouzis, K., Hirsch, L., & Akhgar, B. (2014). Combining big social media data and fca for crisis response. In *2014 IEEE/ACM 7th international conference on utility and cloud computing (UCC)* (pp. 690–695).

Google (2015). Google person finder. https://google.org/personfinder/2015-nepal-earthquake/.

Goolsby, R. (2010). Social media as crisis platform. *ACM Transactions on Intelligent Systems and Technology*, *1*(1), 1–11.

Heinzelman, J., & Waters, C. (2010). Crowdsourcing crisis information in disaster-affected Haiti. Tech. rep., United States Institute of Peace.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster- Related messages in social media. In *ISCRAM 2013 Conference Proceedings - 10th international conference on information systems for crisis response and management* (pp. 791–801). Germany: Baden-Baden.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: Artificial intelligence for disaster response. In *Proceedings of the companion publication of the 23rd international conference on World Wide Web companion, international World Wide Web conferences steering committee, republic and Canton of Geneva, Switzerland, WWW Companion '14* (pp 159–162).

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), 671–6738. doi:10.1145/2771588.

Jay, N., Kohler, F., & Napoli, A. (2008). Analysis of social communities with iceberg and stability-based concept lattices. In R. Medina, & S. Obiedkov (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 4933, pp. 258–272). Berlin: Springer.

Kaplan, A.M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68.

Kaytoue, M., Duplessis, S., Kuznetsov, S., & Napoli, A. (2009). Two fca-based methods for mining gene expression data. In S. Ferré, & S. Rudolph (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 5548, pp. 251–266). Berlin: Springer.

Kaytoue, M., Kuznetsov, S.O., Napoli, A., & Duplessis, S. (2011). Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences*, *181*(10), 1989–2001.

Keßler, C., & Hendrix, C. (2015), The humanitarian exchange language: coordinating disaster response with semantic web technologies. *Semantic Web Journal*, accepted.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation*, *42*(1), 21–40.

Kumar, S., Barbier, G., Abbasi, M.A., & Liu, H. (2011). Tweettracker: an analysis tool for humanitarian and disaster relief. In *Fifth international AAAI conference on weblogs and social media. ICWSM '11.*

Lee, J., & Bui, T. (2000). A template-based methodology for disaster management information systems. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*, 2000. doi:10.1109/HICSS.2000.926635.

Limbu, M., Wang, D., Kauppinen, T., & Ortmann, J. (2012), Management of a Crisis (MOAC) vocabulary specification. http://www.observedchange.com/moac/ns/.

MacKenzie, D. (2013). Aftermath of a typhoon. *New Scientist*, *220*(2943), 6–7.

Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., & Miller, R.C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on human factors in computing systems—CHI '11* (pp. 227–236).

Meier, P. (2012). Crisis mapping in action: how open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map & Geography Libraries*, *8*(February 2015), 89–100.

Meier, P., Lucas, J., Mackinnon, J., & Gonzalez, D.L. (2014). MicroMappers. http://micromappers.com/.

Middleton, S.E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17.

Motameny, S., Versmold, B., & Schmutzler, R. (2008). Formal concept analysis for the identification of combinatorial biomarkers in breast cancer. In R. Medina, & S. Obiedkov (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 4933, pp. 229–240). Berlin: Springer.

Nguyen, M.T., Kitamoto, A., & Nguyen, T.T. (2015). Tsum4act: a framework for retrieving and summarizing actionable tweets during a disaster for reaction. In T. Cao, E. P Lim, Z.H. Zhou, T. B. Ho, D. Cheung, & H. Motoda (Eds.) *Advances in knowledge discovery and data mining. Lecture notes in computer science* (Vol. 9078, pp. 64–75). Berlin: Springer International Publishing.

OCHA (2012). OCHA launches 500 free humanitarian symbols. http://www.unocha.org/top-stories/all-stories/ocha-launches-500-free-humanitarian-symbols.

Okolloh, O. (2009). Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action (January)*, 65–70.

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). CrisisLex: a lexicon for collecting and filtering microblogged communications in crises. In *Proceedings of the 8th international conference on weblogs and social media* (p. 376). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8091/8138.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: social media communications across crises. In *Proceedings of the ACM 2015 conference on computer supported cooperative work and social computing (CSCW '15)*. Vancouver: ACM.

Omand, D., Bartlett, J., & Miller, C. (2012). Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, *27*(6), 801–823.

Ortmann, J., Limbu, M., Wang, D., & Kauppinen, T. (2011). Crowdsourcing Linked Open Data for Disaster Management. In *Proceedings of the 10th international semantic web conference (January)*, 2010.

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2009). In P. Perner (Ed.) *Advances in data mining. Applications and theoretical aspects. Lecture notes in computer science* (Vol. 5633, pp. 247–260). Berlin: Springer.

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2010). Formal concept analysis in knowledge discovery: a survey. In M. Croitoru, S. Ferré, & D. Lukose (Eds.) *Conceptual structures: from information to intelligence. Lecture notes in computer science* (Vol. 6208, pp. 139–153). Berlin: Springer.

Pohl, D., Bouchachia, A., & Hellwagner, H. (2012). Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference companion on World Wide Web—WWW '12 Companion* (p. 683). doi:10.1145/2187980.2188180.

Rogstadius, J., Vukovic, M., Teixeira, C.A., Kostakos, V., Karapanos, E., & Laredo, J.A. (2013). CrisisTracker: crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, *57*(5), 41–413.

Rome, J., & Haralick, R. (2005). Towards a formal concept analysis approach to exploring communities on the world wide web. In B. Ganter, & R. Godin (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 3403, pp. 33–48). Berlin: Springer.

Roy Chowdhury, S., Purohit, H., & Imran, M. (2015). D-sieve: a novel data processing engine for efficient handling of crises-related social messages. In *Proceedings of the 24th international conference on World Wide Web, WWW '15 companion* (pp. 1227–1232). New York: ACM. doi:10.1145/2740908.2741731.

SAS (2015). SAS Content Categorization Studio. http://support.sas.com/software/products/ccs/index.html.

Shanley, L.A., Burns, R., Bastian, Z., & Robson, E.S. (2013). Tweeting up a storm the promise and perils of crisis mapping. *Photogrammetric Engineering & Remote Sensing*, 865–879.

Stumme, G., Wille, R., & Wille, U. (1998). Conceptual knowledge discovery in databases using formal concept analysis methods. In J. Żytkow, & M. Quafafou (Eds.) *Principles of data mining and knowledge discovery. Lecture notes in computer science* (Vol. 1510, pp. 450–458). Berlin: Springer.

Temnikova, I., Castillo, C., & Vieweg, S. (2015). EMTerms 1. 0: a terminological resource for crisis tweets. In *ISCRAM 2015 proceedings of the 12th international conference on information systems for crisis response and management*.

Tilley, T., Cole, R., Becker, P., & Eklund, P. (2005). A survey of formal concept analysis support for software engineering activities. In B. Ganter, G. Stumme, & R. Wille (Eds.) *Formal concept analysis. Lecture notes in computer science* (Vol. 3626, pp. 250–271). Berlin: Springer.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: what 140 characters reveal about political sentiment. In *ICWSM* (pp. 178–185).

Vieweg, S.E. (2012). Situational awareness in mass emergency: a behavioral and linguistic analysis of microblogged communications.

Vieweg, S., Hughes, A.L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems, CHI '10* (pp. 1079–1088). New York: ACM.

Wermelinger, M., Yu, Y., & Strohmaier, M. (2009). Using formal concept analysis to construct and visualise hierarchies of socio-technical relations. In *31st international conference on software engineering—companion volume, 2009. ICSE-Companion 2009* (pp. 327–330).

Wille, R. (2005). Formal concept analysis as mathematical theory of concepts and concept hierarchies. In *Formal concept analysis* (pp. 1–33). Berlin: Springer.

Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, *4*(1), 163. doi:10.1007/s13278-014-0163-y.

Zuckerberg, M. (2015). Nepal earthquake safety check activated, https://www.facebook.com/zuck/posts/10102050030813611?pnref=story.