Journal of Systems and Information Technology

A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations
Konstantinos Domdouzis, Babak Akhgar, Simon Andrews, Helen Gibson, Laurence Hirsch,

Article information:

Users who downloaded this article also downloaded:

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald
for Authors service information about how to choose which publication to write for and submission
guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company
manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as
well as providing an extensive range of online products and additional customer resources and
services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the
Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for
digital archive preservation.

*Related content and download information correct at time of download.

A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations

Konstantinos Domdouzis
Department of Computing, Sheffield Hallam University, Sheffield, UK

Babak Akhgar
Centre of Excellence in Terrorism, Resilience, Intelligence and Organized Crime Research (CENTRIC), Sheffield Hallam University, Sheffield, UK

Simon Andrews
Department of Computing, Sheffield Hallam University, Sheffield, UK

Helen Gibson
Centre of Excellence in Terrorism, Resilience, Intelligence and Organized Crime Research (CENTRIC), Sheffield Hallam University, Sheffield, UK, and

Laurence Hirsch
Department of Computing, Sheffield Hallam University, Sheffield, UK

Abstract

Purpose – A number of crisis situations, such as natural disasters, have affected the planet over the past decade. The outcomes of such disasters are catastrophic for the infrastructures of modern societies. Furthermore, after large disasters, societies come face-to-face with important issues, such as the loss of human lives, people who are missing and the increment of the criminality rate. In many occasions, they seem unprepared to face such issues. This paper aims to present an automated social media and crowdsourcing data mining system for the synchronization of the police and law enforcement agencies for the prevention of criminal activities during and post a large crisis situation.

Design/methodology/approach – The paper realized qualitative research in the form of a review of the literature. This review focuses on the necessity of using social media and crowdsourcing data mining techniques in combination with advanced Web technologies for the purpose of providing solutions to problems related to criminal activities caused during and after a crisis. The paper presents the ATHENA crisis management system, which uses a number of data mining techniques to collect and analyze crisis-related data from social media for the purpose of crime prevention.

Findings – Conclusions are drawn on the significance of social media and crowdsourcing data mining techniques for the resolution of problems related to large crisis situations with emphasis to the ATHENA system.

Originality/value – The paper shows how the integrated use of social media and data mining algorithms can contribute in the resolution of problems that are developed during and after a large crisis.

Keywords Analysis, Sentiment, Social media, Crisis, Crowdsourcing, ATHENA

Paper type Research paper

1. Introduction

Modern societies are characterized by increased crime rates, which are expressed in a multi-faceted manner. As societies develop, so does crime. The prevention of any type of crime is crucial for the maintenance of social stability and the further intellectual and financial growth of societies. To fight crime, the causes of criminal behavior must be identified and understood as well as the ways crime planning is realized. The advancement of information and communication technologies has enhanced criminal activities, but at the same time, it provided a powerful tool in the fight against crime. Specifically, the use of such technologies has allowed criminal activities to be realized in a quicker and more secretive way; however, they also enhanced the ways the law enforcement agencies (LEAs) monitor and prevent criminal activities. Each type of crime is characterized by different requirements and can be analyzed under the context of different circumstances (Kleemans *et al.*, 2012). Especially the issue of criminal activities during and post a crisis is very sensitive, as the police and LEAs have to face a number of consequences caused because of the crisis.

For example, after the Haiti earthquake, the crimes rates have increased. This is because of prisoner escapes during the earthquake and the presence of armed youth gangs that try to gain control of vulnerable areas. Displaced people who live in tents around Port-au-Prince are vulnerable to crime, while the danger of rapes for women is very high. Based on data provided by the Haitian National Police, 5,136 prisoners escaped, including 700 gang members (Berg, 2010).

Globalization and technological advances have led to the quick and complex evolution of crime. Cyber-crime is nowadays a well-known term and affects millions of business and individuals online. The total loss from cyber-crime in years 2000, 2001 and 2002 increased to $265 million, $378 million and $450 million, respectively, while the total loss from 1997 to 2002 was $2 billion. Examples of cyber-crimes are hacking of company databases and theft of financial, product or research and development data (Nykodym *et al.*, 2005). The terrible 2015 Nepal earthquake was exploited by internet fraudsters, who sent multiple phishing emails requesting donations for the victims of the earthquake.

An example of how globalization and technology have enhanced this evolution can be shown by the way organized crime operates. Organized crime has adopted more structured models in their operations and uses technology to expand their activities beyond national levels. The severity of this situation can be proved by the development by the Obama Administration, of the "Strategy to Combat Transnational Organized Crime (2011 Strategy)". The strategy describes the threat of transnational organized crime networks toward US national security by defining this threat, outlining five policy objectives and presenting six priority actions that need to be taken (Bjelopera and Finklea, 2012). Furthermore, terrorist groups have adopted the use of technology to carry more sophisticated attacks on their targets. They use also advanced technologies to ideologically affect large number of people (e.g. use of Web pages and blogs that

contain messages of hatred) and plan their operations (e.g. exchange of emails between different terrorist organizations). Especially during and after a large natural disaster, societies are more vulnerable toward terrorist groups. In 1984, the New People's Army increased the frequency of its attacks in the Philippines after the two severe typhoons, Nitang and Undang. In 1976, a 7.5-magnitude earthquake hit Guatemala City. The terrorist attacks increased after the earthquake. The same happened in Thailand and Sri Lanka post the 2004 tsunami (Berrebi and Ostwald, 2011).

McQuade (2006) has created a matrix of complexity of crimes against complexity of used technologies. Specifically, he distinguished four categories of crime realized versus used technology. The first category involves a simple crime, such as a building check by a police officer who uses a flashlight. The second category involves the issuing of neighborhood block watch stickers by a police officer for the purpose of conducting crime prevention seminars in the community. The third category is related to simple policing using, however, complex tools, such as a well-equipped police car. The fourth category is related to complex policing using complex tools. An example in this case is the prevention of cellular phone-based fraud by using electronic surveillance equipment and geographical information systems (McQuade, 2006). Other criminal activities based on the use of advanced information and communication technologies are the online publication of hate speech and defamatory information, the publication and exchange of child pornography images, violation of intellectual property rights, financial theft and the spreading of computer viruses. The continuously evolving complexity of crime requires the use of advanced information and communication processing tools and especially advanced techniques and algorithms that will allow the efficient extraction of useful data related to possible criminal activities.

In recent years, social media have become very popular. Even though the term "Social Media" is vague, nowadays it is used to describe a group of internet-based applications that build on the ideological and technological foundations of Web 2.0, and allow the exchange of data generated by users (Kaplan and Haenlein, 2010). Social media can be extremely useful in crisis situations. Disaster responders can use social media to reconnect families and analyze critical information (Armour, 2010). Palen *et al.* (2009) specify that the gradual addition of information through social media channels can help in the creation of an overall image about a disaster. During the 09/11 attacks, cell phones (the then new technology) were used to communicate messages for the few minutes they worked after the planes crashed. In 2004, during the Indian Ocean tsunami, people made use of SMS, even though the cell phone service was down. In 2005, during Hurricane Katrina, social media's role was crucial in locating missing persons and coordinating disaster management services (Nelson *et al.*, 2010).

Through social media monitoring, potential criminal activity especially during and post large crisis incidents can be detected. Social network analysis could be used to identify a criminal network and profiles could be matched across social media platforms and onto police records. Facial recognition may be used to pair profile pictures or pictures sent in by the public or appearances in videos to police records. This builds up a picture of the complete online profile and networks. This allows LEAs to target specific members of the network or identify weak links. Messages and pictures with geo-location data posted by not only criminals but also members of the public in social networks are compared with information stored in databases that contain historical data. Text mining techniques can be used to

extract more metadata from the posted messages and pictures, such as names of other criminals or keywords related to criminal activities. Missing persons or items can also be identified through posted images.

Chen *et al.* (2004) have identified four major categories of crime data mining techniques, which are entity extraction, association, prediction and pattern visualization. Each category represents a set of techniques for use in certain types of crime analysis. For example, artificial neural networks can be used for the extraction and prediction of crime entities. Clustering techniques are used in crime association and prediction. Social network analysis can facilitate crime association and pattern visualization (Chen *et al.*, 2004). Chen *et al.* (2003) have suggested a number of scenarios using data mining techniques for police investigations, such as the use of artificial neural networks for extraction of entities from police narrative reports; the use of an algorithmic approach based on the calculation of Euclidean distances for the identification of identity deceptions by criminals; the tracing of identities of criminals from posted messages on the Web using learning algorithms, such as support vector machines (SVMs); and the use of social network analysis for uncovering structural patterns from criminal networks (Chen *et al.*, 2003).

Crowdsourcing data are also of great significance in crisis situations. Crowdsourcing allows efficient information flows between emergency management specialists and the public. Crowdsourcing covers a wide range of activities at different forms. Brabham (2008) defines crowdsourcing as an online, distributed problem-solving and production model. Howe (2006) defines crowdsourcing as the act of a company or institution to take a function once performed by employees and outsource it to an undefined and usually large group of people in the form of an open call. Crowdsourcing allows a crowd of users to cooperate and develop an artifact that can benefit the whole community. As indicated by Howe (2006), the word crowdsourcing is used for a wide group of activities that take on different forms. The adaptability of crowdsourcing allows it to be an effective and powerful practice but makes it difficult to define and categorize.

There are a number of tools, such as Pathfinder, Sense.us and Many Eyes, that are used for analysis of crowdsourcing information. Formal concept analysis (FCA) is also a very important mathematical technique that can be used for the analysis of large data sets. FCA can be used for the analysis of data generated by social media and which are related to criminal incidents. For example, in an extremely harsh crisis situation, like the 09/11 attacks or the Haiti earthquake, large data sets can be created by the broadcasting of online messages through social media, like Twitter or Facebook. In this case, FCA can be used to extract these data, classify them and find relations between them based on their attributes.

This paper focuses on the use of social media and crowdsourcing data scanning and mining techniques through the ATHENA Web-based system for the prevention of criminal activities (e.g. looting) during and post crisis situations. The paper provides an overview of the main techniques used for social media and crowdsourcing scanning and then presents the ATHENA crisis management system, with focus on the social media and crowdsourcing data scanning and mining techniques used by the system. A number of conclusions are drawn about the importance of both the social media and crowdsourcing scanning techniques and of the ATHENA system.

2. Examples of social media and crowdsourcing data scanning and mining techniques

The content of social media and crowdsourcing data sources is dynamic both in quantity and quality. There must therefore exist an extensive evaluation of the existing tools used for the analysis of the generated information. Examples of such tools are link analysis, sentiment analysis/opinion mining and crowdsourcing data analysis. These methods are used in the prevention of different types of crime (Kontostathis *et al.*, 2009; Agarwal *et al.*, 2013).

2.1 Link analysis in social media

Social network sites, such as MySpace and Facebook, allow users to list interests and link to friends based on trust levels. Link analysis algorithms are used to examine the structure of interconnected links. If two pages are linked, this means that their content is related. If many pages point to a specific link, this means that its content is important. These two assumptions have been used for page ranking, finding site homes pages and document classification.

There are a number of algorithms which are used for exploring possible relationships between Web documents. The hyperlink-induced topic search (HITS) algorithm is used to find related pages when a topic is defined by a single page. It characterizes a page as an "authority" if it is pointed by another page and a page as a "hub" if it links to other pages. HITS starts with a root set of text-based search engine results in relation to a query about some topic, expands the root set to a base set with the in-links and out-links, eliminates the links between pages with the same domain to define a graph, runs the above equations until convergence and returns a set of documents with high h(p) weights and another set with high a(p) weights (Yang, 2002).

Another algorithm which is used for Web document classification is Google's PageRank. This algorithm is more efficient than HITS, as the time used by the algorithm to assign a specific importance or weight to a specific Web document is low. Furthermore, it is less susceptible to link spam (Haveliwala, 2003). PageRank is used in the Google search engine and assigns to a page, a score proportional to the number of times a random user visits that page (Richardson and Domingos, 2002). The basic idea is that if a page u points to a page v, then the author of u is providing some weight to page v (Haveliwala, 2003).

The companion algorithm is another example of a link analysis algorithm. The algorithm uses a starting uniform resource locator (URL) and includes four steps:

(1) *Construction of a vicinity graph for the specific URL*: Given the specific URL, a directed graph of nodes that are nearby to the specific link in the Web graph is constructed.

(2) *Duplicate elimination*: After the graph has been constructed, there is combination of near-duplicates. Two nodes are near-duplicates if they each have more than ten links and they have at least 95 per cent of their links in common.

(3) *Assignment of edge weights*: A weight is assigned to each edge. An edge which is found between two nodes on the same host has weight 0. If there are k edges from documents on a first host to a single document on a second host, an edge is given the authority weight of 1/k. If there are l edges from a single

document on a first host to a set of documents on a second, then the edge is given a weight of 1/l.

(4) *Computation of hub and authority scores*: Hub and authority scores are computed using the *imp* algorithm, which is an extension to the HITS algorithm (Dean and Hezninger, 1999).

Rose and Chandran (2012) presented the normalized Web distinction-based query classification. This approach classified the user's queries into an intermediate set of categories. Features were extracted using feature selection algorithms. Normalized Web distance was used for the mapping of categories of features into the target categories. Experiments for the normalized Web distinction query classification were conducted and proved that the precision with this approach is 10 per cent higher than existing approaches.

Alamelu Mangai *et al.* (2012) have presented a novel feature selection framework for automatic Web page classification. The performance of the framework was based on the elimination of noisy data. The first step is the conversion of the Web page to a text document and the mining of the textual features. The best features are then used for the classification. The selection of these features results to dimensionality reduction and also the reduction of the time and resources required for the classification. An entropy measure called ward entropy was used to estimate the most relevant features for Web page classification. Artificial neural networks were used for the training of the classifier with selected features.

2.2 Opinion mining/Sentiment analysis

Opinion mining (or sentiment analysis) is the field that analyzes people's opinions, sentiments and attitudes toward specific entities, such as products, services, individuals and topics (Liu, 2012). Sentiment analysis involves identification of sentiment expressions, polarity and strength of the expressions and their relationship to the subject. The most important issue in sentiment analysis is the identification of how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Therefore, sentiment analysis involves identification of sentiment expressions, polarity and strength of the expressions and their relationship to the subject (Nasukawa and Yi, 2003).

A number of tasks are associated to sentiment analysis. These are sentiment classification, subjectivity classification, opinion summarization and opinion retrieval. Sentiment classification is based on the idea that a document expresses an opinion on an entity from an author and the sentiment of the author toward this entity is measured. Subjective classification involves the detection of whether a sentence is subjective or not. Opinion summarization is related to the extraction of the main features of an entity shared across different documents and the sentiments related to these features. Opinion retrieval tries to retrieve documents which express an opinion about a given query. In this case, two scores are required to be computed for each document: the relevance score against the query and the opinion score about the query (Serrano-Guerrero *et al.*, 2015).

Sentiment analysis follows machine-learning and lexicon-based approaches. Machine-learning approaches can be distinguished to supervised and un-supervised learning techniques. Supervised learning includes decision tree classifiers, linear classifiers, rule-based classifiers and probabilistic classifiers. Linear classifiers include

SVMs and neural networks. Probabilistic classifiers include naïve Bayes classifier, Bayesian networks and maximum entropy. The lexicon-based approach of sentiment analysis includes the dictionary-based and the corpus-based approach, which is classified as statistical and semantic (Serrano-Guerrero *et al.*, 2015).

Feldman (2013) has defined different techniques for sentiment analysis and these are document-level, sentence-level, aspect-based, comparative sentiment analysis and sentiment lexicon acquisition. The document-level supervised approach assumes that there is a finite set of classes into which the document is classified and training data are available for each class. Given the training data, the system learns a classification model through the use of one or more classification algorithms, such as SVMs, naïve Bayes or logistic regression. Document-level unsupervised approaches are based on the determination of the semantic orientation of specific phrases within a document. Prior to any determination of the polarity of the sentences, it must be clarified whether the sentences are subjective or objective. Only subjective sentences are further analyzed. Aspect-based sentiment analysis refers to the identification of sentiment expressions in a text and the aspects to which they refer to. Comparative sentiment analysis is the identification of the sentences that contain comparative opinions and extract the preferred entities in each opinion. The sentiment lexicon is the most crucial resource. There are three approaches for developing a lexicon and these are the manual approach, the dictionary-based approach and the corpus-based approach. An example of a corpus-based approach is the identification of adjectives with consistent polarity. A set of linguistic connectors are used to identify the connection between adjectives with not-known polarity to those with known (Feldman, 2013).

2.3 Crowdsourced data analysis

The Social Analysis and Intelligence Group provides behavioral analysis, sentiment analysis and clustering of data related to civil unrest (Gibbs, 2015). Shah *et al.* (2011) have presented the CROWDSAFE system, which is used for crowdsourcing of crime data. CROWDSAFE uses density-based methods to produce heat-maps for crime incidents for each month. Furthermore, based on user data, the system computes safe routes for crowdsourced users. Crime clusters are also used based on the K-means and the density-based spatial clustering of applications with noise algorithms. Crime clusters help the police to plan patrol routes and identify patrol boundaries (Shah *et al.*, 2011). Techniques that are used for mining crowdsourcing data are classification, clustering, semi-supervised learning, sampling and association rule mining. In classification, a classifier is used to extract features from given data sets. In social networking sites (e.g. Twitter), classification is based on the use of tags. Semi-supervised learning uses labeled data to acquire useful knowledge. Sampling is related to the verification of hypotheses based on a sample set of information from a whole data set. Association rule mining is related to the finding of relationships among seemingly unrelated data (Xintong *et al.*, 2014).

3. The ATHENA Project

The ATHENA Project is a European Union project that aims to develop a crisis communication and management system that allows the public to take part in the coordination of search and rescue operations during a crisis situation using social media. The ATHENA Project aims to develop a set of guidelines for the police and LEAs

for the use of social media in crisis situations and also a suite of software tools that automates the search and rescue processes. Although the ATHENA Project is focused more on the efficient use of media communications during crisis situations and the resolution of problems arising from crisis incidents, it can also provide means for preventing crime and looting during large disasters.

The ATHENA Project includes nine work packages (WPs). These WPs cover areas such as crisis communication requirements and ethics, and human factors and best practices. The core element of the ATHENA Project is the Command and Control Centre Intelligence Dashboard (CCCID) that will provide crisis summary information, direct communication between community members and first responders and a content management system to monitor the dedicated crisis social media pages, headline and alerts. The ATHENA system uses data mining algorithms, such as FCA and sentiment analysis, to analyze data collected from crisis-dedicated social media (Domdouzis *et al.*, 2014).

The vision of the ATHENA Project is to achieve an effective crisis response by giving a voice to the citizens during a crisis incident. The ATHENA system includes six main components. These are the Crisis Mobile, the Crisis Information Processing Centre (CIPC), the CCCID, the Social Media Manager, Interoperability (Crisis Management Language, Decentralized Information Processing Framework) and the ATHENA Cloud Secure Information Centre. Figure 1 shows the different elements of the ATHENA system.

A significant element of the ATHENA system is the CIPC. The CIPC includes acquisition and pre-processing tools as well as aggregation and analysis tools. The acquisition and pre-processing tools include the social media scanner, the citizen report streaming/recording center, the speech recognition system, the filter system and the crisis taxonomy system. The aggregation and analysis tools include the classification/ clearance system, the FCA summarizing system, the data fusion system, the credibility scoring system and the sentiment analysis tool. Specific tools and techniques that are used by the CIPC to analyze information are listed below.

4. Social media crowdsourcing data collection tools of the ATHENA system

The ATHENA system uses the following tools for collecting crowdsourcing data. These tools are presented in detail in the following text.

4.1 ATHENA mobile app, crisis map and news list

The ATHENA crisis map is an integral element of the ATHENA system. It shows validated messages of users during crisis situations. The messages on the map have the form of pins. By clicking on a pin, the user of the CCCID can validate a report by checking its content, the credibility, the clearance level and the priority level. The users can use the ATHENA mobile app to post messages to the crisis map. Each pin represents a report. Only validated (by the CCCID user) reports will be displayed on the crisis map. There are different categories, which are displayed in Figure 2.

The mobile application will be one application that supports two types of user: trusted user and citizen user. There are two tiers of trusted user: Tier 1 (top tier) and Tier 2 (lower tier). The type of user will determine what the user can see on the map. Citizen users do not need to login. Tier 1 trusted users include first responders and
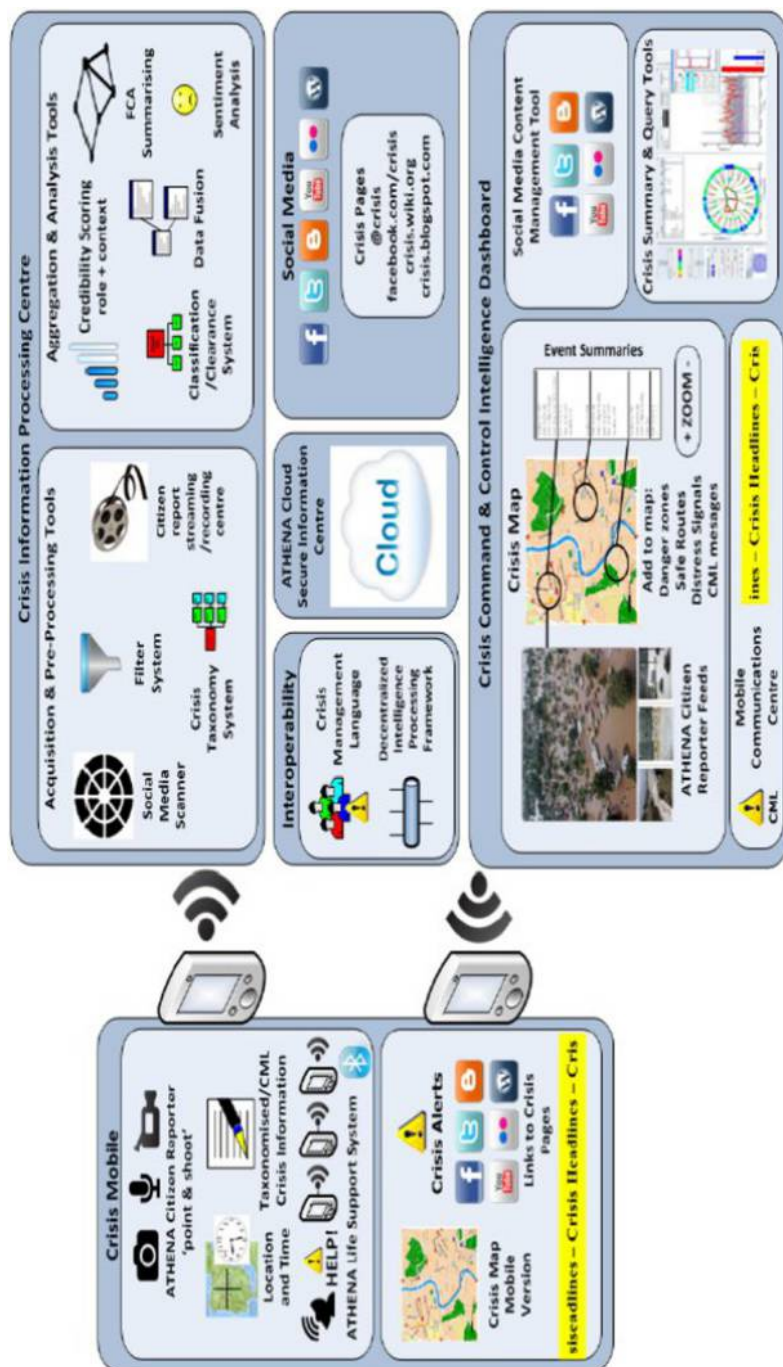
Figure 1.
The ATHENA
system

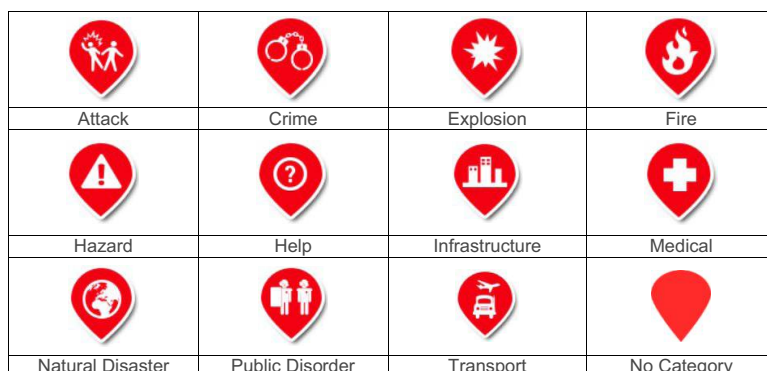| Attack	Crime	Explosion	Fire
Hazard	Help	Infrastructure	Medical
Natural Disaster	Public Disorder	Transport	No Category

Figure 2.
Different pin
categories

bronze (operational), silver (tactical) and gold (strategic) command. Tier 2 trusted users include utilities controllers, official volunteers and credible community voices. Reports that are validated for all by the CCCID user will be displayed to all users. Validated at Tier 1 reports will only be displayed only to Tier 1 trusted users, whereas validated at Tier 2 reports will be displayed to Tier 1 and Tier 2 users. Un-validated reports are only displayed to Tier 1 users. Figure 3 shows the iOS and Android versions of the mobile app.

By clicking on the top left icon of the ATHENA mobile app, the user can open the side menu of the mobile app. This is shown in Figure 4.

The mobile app allows the display of information in the form of pins that correspond to crisis information in relation to specific places. This is shown in Figure 5.

The crisis map used by the CCCID displays information related to crisis received by citizens. This information is displayed on the CCCID crisis map in the form of pins. By clicking each pin, the relevant information is displayed. A news list located next to the crisis map includes a list of the reports IDs with their titles. The CCCID crisis map is shown in Figures 6 and 7.

The CCCID operator can validate the incoming reports and the setting of clearance levels, which will determine which ATHENA mobile app users are able to see them. Users of the ATHENA System are divided into two categories: citizen users and trusted users. Trusted users can be further divided into two categories. Tier 1 users include first-responders and the command team, and Tier 2 users include utilities controllers, official volunteers and credible community voices. The reports can be validated or un-validated (rejected). The clearance levels are 3 = all able to see, 2 = Tier 1 and Tier 2 users only and 1 = Tier 1 users only. The CCCID operators and Tier 1 users can see all reports. This includes also reports which are "rejected"/un-validated by the CCCID operator.

4.2 Social presence of the ATHENA system
The ATHENA system uses the Twitter and Facebook social media to extract collective information about a crisis. A tweet can include information such as text, image, video, links and hash-tags. Twitter data can also be extracted through the Twitter search API. The search API is focused on relevance. There is a significant amount of metadata that comes with each tweet. These metadata include geo-location data, author name and

Figure 3.
The iOS and Android versions of the ATHENA mobile app
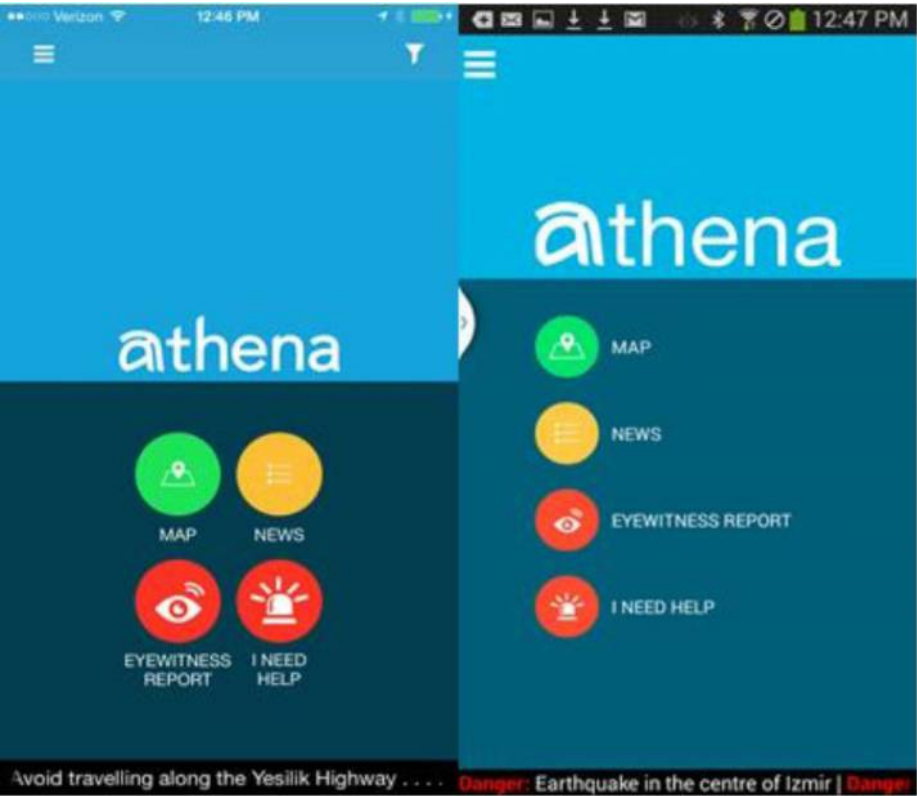
pre-set location, timestamp, number of re-tweets, number of favorites, list of hash-tags, list of links and other users that are possibly mentioned in the tweet.

Users of Facebook can restrict their profiles to their friends only; therefore, the viewing of their posts is not always possible. For this reason, the analysis of Facebook data focuses on the comments made on dedicated ATHENA crisis pages. Figures 8 and 9 show the dedicated ATHENA social media pages.

5. ATHENA social media and crowdsourcing data scanning and mining techniques

The ATHENA system crowdsourced data analysis tools use a number of data analysis techniques. These are presented in detail further.

5.1 Web-crawling using the SAS information retrieval studio

The SAS Information Retrieval Studio (IRS) is used to set-up crawlers for websites in order for the ATHENA system to mine data of interest. Another functionality offered by the SAS IRS is data filtering. An example of filtering is to ignore re-tweets. This is facilitated by the Twitter API, which contains a field that shows the number of times a tweet has been re-tweeted. Additional functionalities of the SAS IRS are categorization, concept and context extraction. The latter stages of the SAS IRS include sentiment analysis and data export to
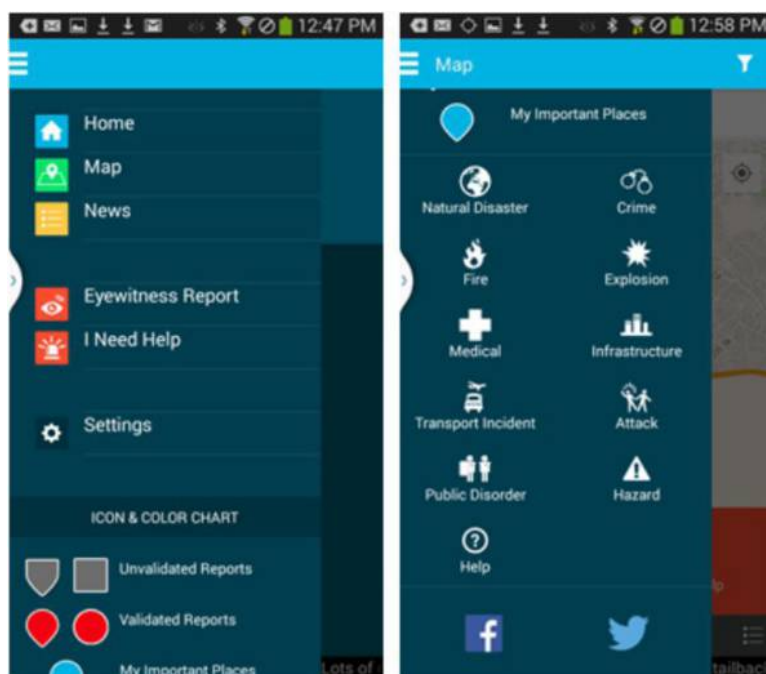
Figure 4.
Side menu of the
ATHENA mobile
app

SQL databases, csv, text files and individual XML documents. All these functionalities offered by SAS IRS consist of the SAS IRS pipeline server. The output of each query is controlled by the SAS IRS pipeline server, which receives documents from the crawlers. The pipeline server processes the document through the stages of filtering, categorization and sentiment analysis and then onto the export process.

SAS IRS provides a number of crawlers, such as the Web Crawler, the File Crawler, the Feed Crawler, the Facebook Crawler, the Flickr Crawler, the Google Crawler, the CSV file explorer, the Twitter Crawler and the YouTube Crawler. Figure 10 is an example of the IRS interface for initiating a Twitter crawl. By using this interface, the user can input search terms and configure the various other options before running the query.

As part of the ATHENA Project, the Twitter crawler has been programmatically updated so that it returns the geo-location of tweets if they are included as part of the tweet, and the Facebook crawler has been updated so the public timeline and the individual pages can be crawled.

5.2 Crisis data summarization using formal concept analysis
The purpose of the integration of FCA within the ATHENA system is the provision of short summaries of crisis situations. Specifically, FCA is used to develop short summaries of social media postings and the monitoring of crisis dynamics (Gibson *et al.*, 2014).

Concepts are the basic units of thought developed in dynamic processes within social and cultural environments. Concepts can only live in relationships with other concepts. FCA is based on two stages. The first stage is the representation of the data as a very basic data type, called the formal context. A formal context is defined as a set structure
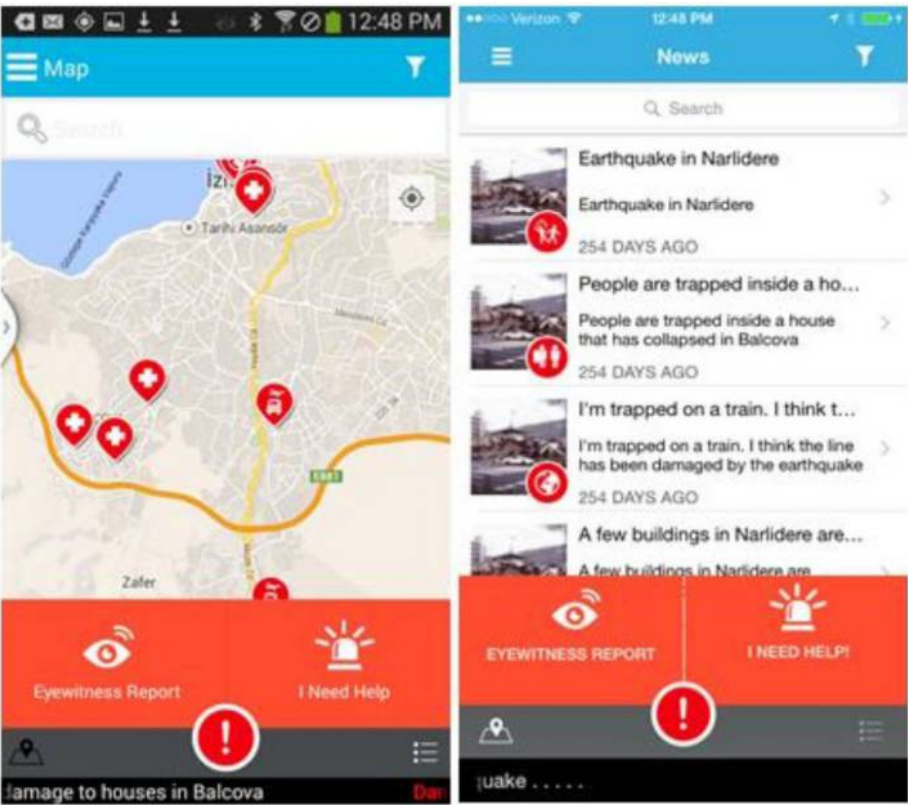
Figure 5.
Mobile crisis map and list of reports

K: = (G, M, I), where G and M are sets and I is the binary relationship between them. The elements of G and M are called (formal) objects and (formal) attributes, respectively, while the expression (g, m) ∈ I (Wille, 2005). Each formal context is transformed into a mathematical structure called concept lattice. The concept lattice shows the relationships among concepts. A concept is determined by its extent and its intent. The extent is the set of all objects that belong to a concept. The intent is the set of all attributes shared by the objects in a concept (Belavkin, 2014).

FCA is an unsupervised learning technique for conceptual clustering. It is a graph-theoretic approach to categorization based on mathematical order and lattice theory (Wille, 2005). Given a set of objects and a set of attributes and a matrix which shows which attributes characterize which object, FCA will first construct all objects. The objects are the extent of the concept and the attributes are its intents. The concepts will then be organized into a lattice (Wermelinger *et al.*, 2014).

5.3 Sentiment analysis and credibility scoring based on natural-language processing
Within the ATHENA CIPC, the sentiment score of each tweet is evaluated based on the text of the tweet. The credibility score can be classified as positive, negative, neutral or "don't know", or the score can be used directly as percentage. This results to a

Figure 6.
ATHENA CCCID
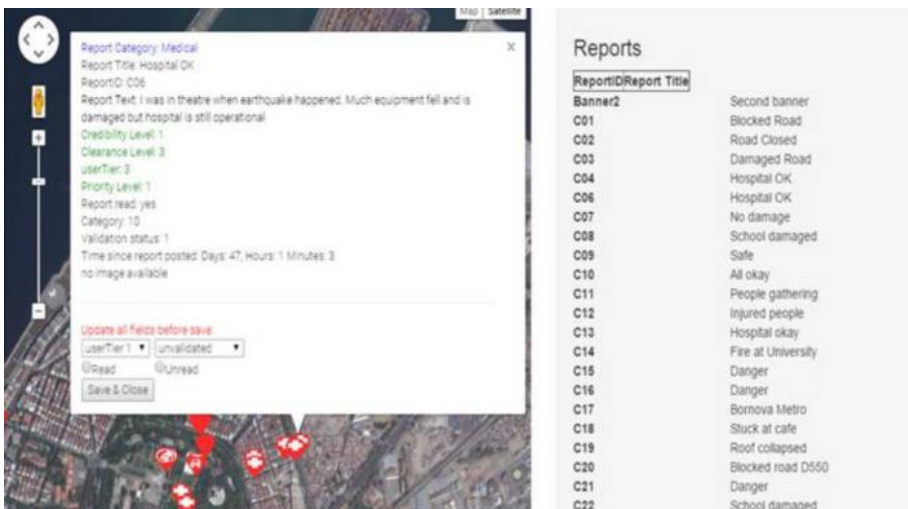crisis map with news
list

Figure 7.
ATHENA CCCID
crisis map displaying
of crisis reports

hierarchical classification of the tweets. In this case, it is possible to detect which categories or concepts are related to positive or negative concepts (Gibson *et al.*, 2014). In the ATHENA project, natural-language processing (NLP) techniques are used for sentiment analysis and credibility scoring. Using NLP and machine learning techniques, a suite of classifiers can be developed to classify tweets based on several criteria such as subjectivity, personal or impersonal style and linguistic register (formal or informal style). The ATHENA system uses categorization, concept and contextual

Figure 8.
Interface of SAS
Information Retrieval
Studio (Twitter
Crawler)

Figure 9.
ATHENA
crisis-dedicated
Facebook page

extraction and the development of crisis concepts. Categorization is the process of analyzing a document's content and associating it to a category.

A taxonomy has been developed that presents these categories and their associated sub-categories. These categories are Attack, Crash, Hazard, Health, Natural Disaster,

Figure 10.
ATHENA
crisis-dedicated
Twitter page
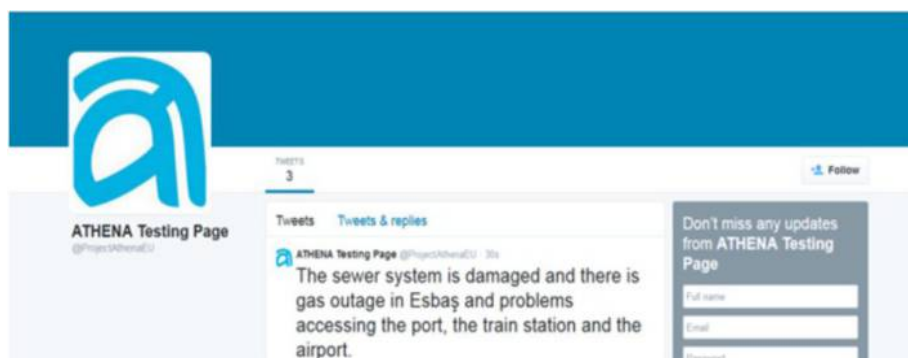
Public Order Incidents and Other. Examples of sub-categories are Bomb, Hostage, Killing, Knife, Lone wolf, Shooting and Suicide Bomb that are associated to the category "Attack". Documents (social media posts and mobile reports) can belong to multiple categories. Each category is described by a series of rules that use Boolean statements. This is a controlled vocabulary related to each category simultaneously evaluating word distance and order.

6. Conclusions

Social media and crowdsourcing contribute significantly in the resolution of problems caused during and after a large crisis, as they are able to provide large amount of crisis-related data in a short period. The appropriate collection and intelligent use of data is very significant for the resolution of crisis-related consequences such as looting, gang violence and sexual assaults. The integrated use of advanced data mining technologies in coordination with crisis-dedicated social media opens a new way of exploring crisis-related data through data classification and aggregation. Specifically, the use of specific techniques and algorithms optimizes the operation of social media and crowdsourced data analysis, as it allows the police and LEAs to identify useful data patterns or even hidden data in this mass of information. The ATHENA system is an example of a Web-based system that uses the power of social media and crowdsourced data mining techniques to synchronize the operations of the police and LEAs during a large crisis (e.g. large natural disaster) and also provide the necessary guidance to search and rescue operations. By collecting and analyzing this real-time and dynamic mass of information, the police and the LEAs can be proactive in the handling of consequences developed as part of a large crisis. Vuori (2012) has underlined the need for using the social media for internal communication and knowledge transfer but also mainly for crowdsourcing.

The system also acts as a tool for the evaluation of the performance of policing operations, and it creates a sense of community among citizens, as it allows online collaboration and easier access to police and LEA services during a crisis. It also allows the development of safe routes and zones around different areas for citizen protection based on validated crisis-related mined data, and it classifies areas based on their current safety conditions. As a result, the system provides efficient organization of the policing procedures, and this results to better decision making during crises, as the police has access to credible crisis-related data. The important feature of the ATHENA

system is its flexibility to exploit the power of a number of data mining algorithms and connect them to a variety of social media and crowdsourced data sources. This provides the capability to the operators of the ATHENA system to use the system under different crisis scenarios and in different requirements. The adaptability of the ATHENA system to the different requirements of each crisis scenario is based on its technical architecture. The technical architecture of the ATHENA system follows the principles of a service-oriented approach suggested by Domdouzis *et al.* (2016) that provides flexibility in the structure of the system by modifying its structure dynamically depending on the requirements of each scenario. The ATHENA system is also sustainable according to the definition of information system sustainability given by Standing and Jackson (2007), as it considers social factors, such as the diversity of opinions through the social media in the crisis decision-making.

Future steps include the integration of the ATHENA system with cloud technologies that will allow massive data storage and better analysis of crime-related data for the purpose of crime prediction. Also, the use of audio and video streaming technologies will allow the improvement of the quality of information received during a crisis.

References

Agarwal, J., Nagpal, R. and Sehgal, R. (2013), "Crime analysis using K-means clustering", *International Journal of Computer Applications (0975-8887)*, Vol. 83 No. 4, pp. 1-4.

Alamelu Mangai, J., Santhosh Kumar, V. and Appavu Alias Balamurugan, S. (2012), "A novel feature selection framework for automatic web page classification", *International Journal of Automation and Computing*, Vol. 9 No. 4, pp. 442-448.

Armour, G. (2010), "Communities communicating with formal and informal systems: being more resilient in times of need", *Bulletin of the American Society for Information Science & Technology*, Vol. 36 No. 5, pp. 34-38.

Belavkin, R.V. (2014), *Lecture 7: Formal Concept Analysis*, BIS4410 – Knowledge Management Strategies at Middlesex University London, available at: www.eis.mdx.ac.uk/staffpages/rvb/teaching/BIS4410/hand07.pdf (accessed 19 May 2016).

Berg, L.A. (2010), "Crime, politics and violence in post-earthquake Haiti", available at: www.usip.org/sites/default/files/PB%2058%20-%20Crime%20Politics%20and%20Violence%20in%20Post-Earthquake%20Haiti.pdf (accessed 10 August 2015).

Berrebi, C. and Ostwald, J. (2011), "Earthquakes, hurricanes, and terrorism - do natural disasters incite terror? Rand labor and population", available at: www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR876.pdf (accessed 10 August 2015).

Bjelopera, J.P. and Finklea, K.M. (2012), "Organized crime: an evolving challenge for US law enforcement", CRS Report for Service, pp. 137-177, available at: www.fas.org/sgp/crs/misc/R41547.pdf (accessed 10 August 2015).

Brabham, D.C. (2008), "Crowdsourcing as a model for problem solving – an introduction and cases", *Convergence: The International Journal of Research into New Media Technologies*, Vol. 14 No. 1, pp. 75-90.

Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J.J., Wang, G., Zheng, R. and Atabakhsh, H. (2003), "Crime data mining: an overview and case studies", *Proceedings of the 2003 Annual National Conference on Digital Government Research*, Boston, MA, 18-21 May, pp. 1-5.

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y. and Chau, M. (2004), "Crime data mining: a general framework and some examples", *Computer: IEEE Computer Society*, Vol. 37 No. 4, pp. 50-56.

Dean, J. and Hezninger, M.R. (1999), "Finding related pages in the World Wide Web", *Computer Networks*, Vol. 31 No. 11, pp. 1467-1479.

Feldman, R. (2013), "Techniques and applications for sentiment analysis", *Communications of the ACM*, Vol. 56 No. 4, pp. 82-89.

Gibbs, M. (2015), "Big data and intelligence: minimizing or preventing violence against citizens and the destruction of property", available at: www.networkworld.com/article/2923045/security0/big-data-and-intelligence-minimizing-or-preventing-violence-against-citizens-and-the-destruction-of.html (accessed 11 August 2015).

Gibson, H., Andrews, S., Domdouzis, K., Hirsh, L. and Akhgar, B. (2014), "Combining big social media data and FCA for crisis response", *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing (UCC)*, London, 8-11 December, pp. 202-207.

Haveliwala, T.H. (2003), "Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search", *IEEE Transactions on Knowledge and Data Engineering*, pp. 784-796.

Howe, J. (2006), "Crowdsourcing: a definition", available at: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html (accessed 19 July 2015).

Kaplan, A.M. and Haenlein, M. (2010), "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, Vol. 53 No. 1, pp. 59-68.

Kleemans, E.R., Soudijn, M.R.J. and Weenink, A.W. (2012), "Organized crime, situational crime prevention and routine activity theory", *Trends in Organized Crime*, Vol. 15 No. 2, pp. 87-92.

Kontostathis, A., Edwards, L. and Leatherman, A. (2009), *Text Mining and Cybercrime In Text Mining: Application and Theory*, John Wiley & Sons, Chichester, West Sussex.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, CA.

McQuade, S. (2006), "Technology-enabled crime, policing, and security", *Journal of Technology Studies*, Vol. 32 No. 1, p. 32.

Nasukawa, T. and Yi, J. (2003), "Sentiment analysis: capturing favorability using natural language processing", *Proceedings of the 2nd international conference on Knowledge capture (K-CAP '03)*, FL, 23-26 October.

Nykodym, N., Taylor, R. and Vilela, J. (2005), "Criminal profiling and insider cybercrime", *Digital Investigation*, Vol. 2 No. 4, pp. 261-267.

Palen, L., Vieweg, S., Liu, S.B. and Hughes, A.L. (2009), "Crisis in a networked world: features of computer-mediated communication in the April 16, 2007, VA Tech event", *Social Science Computer Review: Special Issue on e-Social Science*, Vol. 27 No. 5, pp. 467-480.

Richardson, M. and Domingos, P. (2002), "The intelligent surfer: probabilistic combination of link and content information in pagerank", *Advances in Neural Information Processing Systems*, Vol. 14, MIT Press, Cambridge, MA.

Rose, S.L. and Chandran, K.R. (2012), "Normalized web distance based web query classification", *Journal of Computer Science*, Vol. 8 No. 5, pp. 804-808.

Serrano-Guerrero, J., Olivas, J.A., Romero, F.P. and Herrera-Viedma, E. (2015), "Sentiment analysis: a review and comparative analysis of web services", *Information Sciences*, Vol. 311 No. 2, pp. 18-38.

Shah, S., Bao, F., Lu, C.T. and Chen, I.R. (2011), "Crowdsafe: crowd sourcing of crime incidents and safe routing on mobile devices", *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, *Chicago, IL*, 1-4 November, pp. 521-524.

Standing, C. and P. Jackson (2007), "An approach to sustainability for information systems", *Emerald Journal of Systems and Information Technology (JSIT)*, Vol. 9 No. 2, pp. 167-176.

382

Vuori, M. (2012), "Exploring uses of social media in a global corporation", *Emerald Journal of Systems and Information Technology (JSIT)*, Vol. 14 No. 2, pp. 155-170.

Wille, R. (2005), "Formal concept analysis as mathematical theory of concepts and concept hierarchies", in B., Ganter *et al.* (Ed.), *Formal Concept Analysis*, Vol. 3626, pp. 1-33.

Xintong, G., Hongzhi, W., Song, Y. and Hong, G. (2014), "Brief survey of crowdsourcing for data mining", *Expert Systems with Applications*, Vol. 41, pp. 7987-7994.

Yang, K. (2002), "Combining link- and text-based retrieval methods for web IR", *Proceedings of the 10th Text Retrieval Conference (TREC2001), Gaithersburg, MD*, 15-18 November, pp. 609-618.

Further reading

Aoki, T., Fukumoto, Y., Yasuda, S., Sakata, Y., Ito, K., Takahashi, J., Miyata, S., Tsuji, I. and Shimokawa, H. (2012), "The great east Japan earthquake disaster and cardiovascular diseases", *European Heart Journal*, Vol. 33 No. 22, pp. 2796-2803.

Doan, A., Ramakrishnan, R. and Alon, Y. Halevy (2011), "Crowdsourcing systems on the world-wide web", *Communications of the ACM*, Vol. 54 No. 4, pp. 86-96.

Lovelyn Rose, S. and Chandran, K.R. (2012), "Normalized web distance based web query classification", *Journal of Computer Science*, Vol. 8 No. 5, pp. 804-808.

Nelson, A., Sigal, I. and Zambrano, D. (2015), "Media, information systems and communities: lessons from Haiti", available at: http://knightfoundation.org/media/uploads/publication_pdfs/KF_Haiti_Report_English.pdf (accessed 15 July 2015).

Sahana Software Foundation (2014), "Vesuvius", available at: http://sahanafoundation.org/products/vesuvius (accessed 15 July 2015).

Wermelinger, M., Yu, Y. and Strohmaier, M. (2009), "Using formal concept analysis to construct and visualise social hierarchies of software developers", *International Conference on Software Engineering, Vancouver (ICSE'09)*, British Columbia.

Yates, D. and Paquette, S. (2011), "Emergency knowledge management and social media technologies: a case study of the 2010 Haitian earthquake", *International Journal of Information Management*, Vol. 31 No. 1, pp. 6-13.

Corresponding author

Konstantinos Domdouzis can be contacted at: K.Domdouzis@shu.ac.uk