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

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

Published version

DOMDOUZIS, Konstantinos, AKHGAR, Babak, ANDREWS, Simon and GIBSON, Helen (2016). A social media and crowd-sourcing data mining system for crime prevention during and post-crisis situations. Journal of Systems and Information Technology, 18 (4), 364-382.

A SOCIAL MEDIA AND CROWD-SOURCING DATA MINING SYSTEM FOR CRIME PREVENTION DURING AND POST-CRISIS SITUATIONS

Domdouzis, K.[1], B. Akhgar[2], S. Andrews[3] and H. Gibson[4]

Centre of Excellence in Terrorism, Resilience, Intelligence & Organised Crime Research (CENTRIC)
Cultural, Communication and Computing Research Institute (C3RI)
Faculty of Arts, Computing, Engineering and Sciences (ACES)
Sheffield Hallam University
Sheffield, United Kingdom
1. K.Domdouzis@shu.ac.uk, 2. S.Andrews@shu.ac.uk, 3. B.Akhgar@shu.ac.uk, 4. H.Gibson@shu.ac.uk

ABSTRACT

A number of crisis situations, such as natural disasters have affected the planet over the last 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 system for the synchronization of the police and Law Enforcement Agencies (LEAs) for the prevention of criminal activities during and post a large crisis situation. The paper presents a review of the literature focusing on the necessity of using social media and crowd-sourcing data mining techniques in combination with advanced web technologies for resolving problems related to criminal activities caused during and after a crisis. The focus of the paper is 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. Its main strength is the combined use of a variety of data mining algorithms through a number of interfaces for the purpose of extracting useful social media information related to criminal activities during and after a large crisis. Conclusions are drawn on the significance of social media and crowd-sourcing data mining techniques for the resolution of problems related to large crisis situations with emphasis to the ATHENA system.

Keywords: Crisis, ATHENA, Social Media, Crowdsourcing, Sentiment, Analysis

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. In order 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 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-crises is very sensitive as the police and Law Enforcement Agencies (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 as the well as the presence of armed youth gangs that try to gain control of vulnerable areas. Displaced people that 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 (HNP), 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, 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 in order 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 towards 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 in order to carry more sophisticated attacks on their targets. They use also advanced technologies in order to ideologically affect large number of people (eg. use of webpages and blogs that contain messages of hatred) and plan their operations (eg. exchange of emails between different terrorist organizations). Especially during and after a large natural disaster, societies are more vulnerable towards terrorist groups. In 1984, the New People's Army increased the frequency of its attacks in 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 vs. 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 seminaries in the community. The third category is related to simple policing using however complex tools, such as well-equipped police car. The last 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 as well as 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 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 employed in order 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, 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 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 artefact 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 crowd-sourcing information. Formal Concept Analysis (FCA) is also a very important mathematical technique that can be used for the analysis of large datasets. Formal Concept Analysis 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 datasets 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 crowd-sourcing data scanning and mining techniques through the ATHENA web-based system for the prevention of criminal activities (eg. looting) during and post-crisis situations. The paper provides an overview of the main techniques used for social media and crowd-sourcing scanning and then presents the ATHENA Crisis Management System with focus on the social media and crowd-sourcing data scanning and mining techniques used by the system. A number of conclusions are drawn about the importance of both the social media and crowd-sourcing scanning techniques and of the ATHENA system.

2. EXAMPLES OF SOCIAL MEDIA AND CROWD-SOURCING DATA SCANNING AND MINING TECHNIQUES

The content of social media and crowd-sourcing 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 Crowd-sourcing 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 since 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 Google Search Engine and assigns to a page a score proportional to the number of times a random user visits that page (Richardson & 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 is constructed that are nearby to the specific link in the Web graph.

(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 10 links and they have at least 95% 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 weigh 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 & Henzinger, 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% higher than existing approaches.

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 analyses people's opinions, sentiments and attitudes towards 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 towards 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 Support Vector Machines (SVMs) and Neural Networks (NNs). 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 Support-Vector Machines, 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 Crowd-sourcing Data Analysis

The Social Analysis and Intelligence Group (SAIG) 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 crowd-sourcing 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 crowd-sourced users. Crime clusters are also used based on the K-Means and the Density-based spatial clustering of applications with noise (DBScan) algorithms. Crime clusters help the police to plan patrol routes and identify patrol boundaries (Shah et

al., 2011). Techniques that are used for mining crowd-sourcing data are classification, clustering, semi-supervised learning, sampling, association rule mining. In classification, a classifier is used to extract features from given datasets. In social networking sites (eg. 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 dataset. 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 Law-Enforcement Agencies (LEAs) for the use of social media in crisis situations and also, a suite of software tools that automate 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 work packages 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 Formal Concept Analysis (FCA) and Sentiment Analysis in order 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 Crisis Command and Control Intelligence Dashboard (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 Crisis Information Processing Centre (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 centre, the speech recognition system, the filter system and the crisis taxonomy system. The aggregation and analysis tools include the classification/clearance system, the Formal-Concept Analysis (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 in order to analyze information are listed below.

4. SOCIAL MEDIA CROWD-SOURCING DATA COLLECTION TOOLS OF THE ATHENA SYSTEM

The ATHENA System uses the following tools for collecting crowd-sourcing data. These tools are presented in detail as follows:

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 in order 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, 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 while validated at Tier 2 reports will only 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 Command and Control Centre Intelligence Dashboard (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 to 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, 1=tier 1 users only. The CCCID operators can see all reports as well as Tier 1 users. 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 in order 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 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.

Uses 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 Crowd-Sourcing Data Scanning and Mining Techniques

The ATHENA System Crowd-Sourcing Tools employ a number of data analysis techniques. These are presented in detail as follows.

5.1 Web-Crawling using the SAS Information Retrieval Studio

The SAS Information Retrieval Studio (IRS) is used in order 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 SQL databases, csv, text files as well as 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 (FCA)

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. Formal Concept Analysis (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 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).

Formal Concept Analysis 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 extend 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 (NLP)

Within the ATHENA Crisis Information Processing Centre (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 an 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 (ML) 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 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, 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 crowd-sourcing 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 of time. The use of specific techniques and algorithms optimizes the operation of social media and crowd-sourcing as it allows the police and Law Enforcement Agencies (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 crowd-sourcing data mining techniques in order to synchronize the operations of the police and Law Enforcement Agencies (LEAs) during a large crisis (eg. large natural disaster) and provide the necessary crisis-related information. Through its elements, the ATHENA system automates the process of collecting valid information from the social media. Furthermore, it integrates social media, crowd-sourcing and data mining techniques. The system also acts as a basis for the evaluation of the performance of policing operations, creates a sense of community among citizens through online collaboration, allows the development of safe routes/zones for citizens based on validated crisis-related mined data and classifies areas which need more policing than others. 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.

ACKNOWLEDGEMENTS

This work is co-funded by the European Union Seventh Framework Programme SEC Call 1 - FP7-SEC-2012.6.1-30. We acknowledge and thank Epidemico for their excellent work on the mobile application and Dr Laurence Hirsch for his work on the ATHENA CCCID Crisis Map.

REFERENCES

Agarwal J., Nagpal R., 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., 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

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

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*, pp. 34-38

Belavkin R.V. (2014) "Lecture 7: Formal Concept Analysis", Lecture Notes distributed in BIS4410 - Knowledge Management Strategies at Middlesex University London

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

Berrebi C., Ostwald J. (2011) "Earthquakes, Hurricanes, and Terrorism - Do Natural Disasters Incite Terror? Rand Labor and Population", available at: http://www.rand.org/content/dam/rand/pubs/working_papers/2011/RAND_WR876.pdf (accessed 10 Aug 2015)

Bjelopera J.P., Finklea K.M. (2012) "Organized Crime: An Evolving Challenge for U.S. Law Enforcement", CRS Report for Service, pp. 137-177, available at: https://www.fas.org/sgp/crs/misc/R41547.pdf (accessed: 10 Aug 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*, pp. 75–90

Chen H., Chung W., Qin Y., Chau M., Xu J.J., Wang G., Zheng R., Atabakhsh H. (2003) "Crime Data Mining: An Overview and Case Studies", in *Proceedings of the 2003 annual national conference on Digital government research*. Boston, MA, May 18 - 21. 2003, pp. 1-5

Chen H., Chung W., Xu J.J., Wang G., Qin Y., Chau M. (2004) "Crime Data Mining: A General Framework and Some Examples", *Computer. IEEE Computer Society*, pp. 50-56

Dean J., Hezninger M.R. (1999) "Finding related pages in the World Wide Web", *Computer Networks 31*, pp. 1467–1479

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

Feldman R. (2013) "Techniques and Applications for Sentiment Analysis", *Communications of the ACM*, pp. 82-89

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

Gibson H., Andrews S., Domdouzis K., Hirsh 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),* December 8-11, 2014, London, 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., Haenlein M. (2010) "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, Vol. 53 Issue 1, pp. 59-68

Kleemans E.R., Soudijn M.R.J., Weenink A.W. (2012) "Organized crime, situational crime prevention and routine activity theory", *Trends Organ Crim.*, Vol. 15, pp. 87–92

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

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

Lovelyn Rose S., Chandran K.R. (2012) "Normalized Web Distance Based Web Query Classification", *Journal of Computer Science*, pp. 804-808

McQuade S. (2006) "Technology-enabled crime, policing, and security", *Journal of Technology Studies*, Vol. 32 Issue 1, pp. 32

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

Nelson A., Sigal I., 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 Jul 2015)

Nykodym N., Taylor R., Vilela J. (2005) "Criminal profiling and insider cybercrime", *Digital Investigation*, pp. 261–267

Palen L., Vieweg S., Liu S. B., Hughes A. L. (2009) "Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007, Virginia Tech Event", *Social Science Computer Review*, pp. 467-480

Richardson M., Domingos P. (2002) "The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank", *Advances in Neural Information Processing Systems 14*, MIT Press

Rose S.L., Chandran K.R. (2012) "Normalized Web Distance Based Web Query Classification", *Journal of Computer Science*, pp. 804-808

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

Serrano-Guerrero J., Olivas J.A., Romero F.P., Herrera-Viedma E. (2015) "Sentiment analysis: A review and comparative analysis of web services", *Information Sciences 311*, pp. 18–38

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

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

Yang K. (2002) "Combining Link- and Text-based Retrieval Methods for Web IR", in *Proceedings of the 10th Text Retrieval Conference (TREC2001)*, pp. 609-618

Yates D., Paquette, S. (2011) "Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake", *International Journal of Information Management* 31, pp. 6-13

Wermelinger M., Yu Y., Strohmaier M. (2009) "Using Formal Concept Analysis to Construct and Visualise Social Hierarchies of Software Developers", in *International Conference on Software Engineering, Vancouver (ICSE'09)*, New Ideas and Emerging Results Track (4 page Poster Paper). Vancouver, Canada, 2009.

Wille R. (2005) "Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies" in B. Ganter et al. (Eds.): *Formal Concept Analysis*, Vol. 3626, pp. 1-33