Social-media Text Mining and Network Analysis to support Decision Support for Natural Crisis Management

Zielinski, Andrea
Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation
Karlsruhe, Germany
andrea.zielinski@iosb.fraunhofer.de

Middleton, Stuart E.
University of Southampton IT Innovation
Centre, Southampton, United Kingdom
sem@it-innovation.soton.ac.uk

Tokarchuk, Laurissa N and Wang, Xinyue
Queen Mary University of London,
School of Electronic Engineering and Computer Science,
London, United Kingdom
laurissa@eecs.qmul.ac.uk

ABSTRACT

A core issue in crisis management is to extract from the mass of incoming information what is important for situational awareness during mass emergencies. Based on a case study we develop a prototypical application, TweetComp1, which is integrated into the decision-support component of a Tsunami early warning system and demonstrates the applicability of our approach. This paper describes four novel approaches using focused twitter crawling, trustworthiness analysis, location analysis/geo-parsing, and multilingual tweet classification in the context of how they could be used for monitoring crises. The validity of our state-of-the art text mining and network analysis technologies will be verified in different experiments based on a human annotated gold standard corpus

Keywords

Decision support, social media, text mining, web mining, link analysis, VGI, location extraction

INTRODUCTION

Systems for Natural Crisis Management are increasingly looking to employ Web 2.0 and 3.0 technologies for future collaborative decision making, including the use of social networks. While social sensor data like Twitter is timely, it is not readily accessible and interpretable, since the texts are unstructured, noisy and multilingual.

Within the TRIDEC project[1], we have developed a dynamic spatial-temporal decision making component (DSS) and integrated it with a tsunami warning system that is able to continuously monitor the situation, and send an early official warning alert to the public. We are experimenting with how signs of abnormal activity in social networks can be monitored on a map in the command and control unit's graphical user interface (CCUI) in real-time so that the operator on duty can get an overall picture of the situation.

In this paper, we explore four novel approaches to support the automatic analysis of twitter data: focused crawling (FC), trustworthiness analysis (TA), multilingual tweet classification (MTC), and location analysis/geo-parsing (LOC), all embedded in a data fusion framework. These approaches make use of text mining, natural language processing (NLP), clustering and social network analysis.

The major design goals were a) Real-time analysis, as first alarm tweets are often published within a few seconds after a crisis event; b) High throughput of big data volumes in times of crises (e.g., Gupta et al. (2012) report 5,500 tweets/s for an earthquake in Virginia), and c) High accuracy, so that only information that is relevant for situation awareness is kept for manual inspection.

[1] www.tridec-online.eu

RELATED WORK

The use of Web 2.0 platforms for analysis in the natural disaster domain has been actively researched in the last 3-4 years. With the large amount of first-sight collaborative information available online, early situational awareness is now feasible through sophisticated processing of online social media sources. A 2008 earthquake in China was perhaps the first to be analysed in this way, with (Li and Rao, 2010) reporting a manual analysis of the Twitter response time & accuracy of the messages. Microblogs have been investigated further for Earthquake event detection in Japan (Sakaki, Okazaki, and Matsuo 2010).

The majority of analysis of situational information in Twitter relies on using the Twitter Streaming API to collect tweets in real-time by filtering on a few specialist keywords and hashtags. For example, Starbird and Palen, 2012) use the keys "egypt #egypt #jan25". The definition of keywords is subjective and can lead to incomplete data. (Bifet, Holmes and Pfahringer, 2011) introduce an adaptive Twitter collection mechanism, however they provide little insight into the effectiveness or noise introduced.

In the area of social network analysis in Web 2.0 platforms, Gupta and Kumaraguru (2012) investigated credibility mechanisms and Mendoza, Poblete and Castillo (2010) have looked at the trustworthiness of such messages during the 2010 Chilean earthquake. These initial investigations show the potential of such information in the disaster domain.

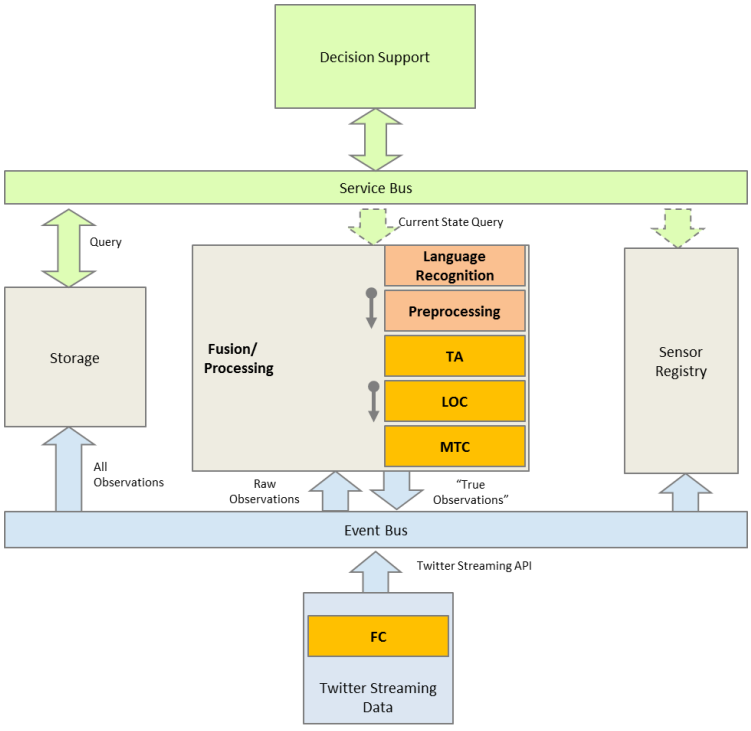

Figure 1: **Knowledge-base data fusion framework supporting a 'forest of classifiers' approach**

Most research on social-media based event classification has been conducted only on a single (i.e. the local) language. The usefulness of covering multiple languages has been explored by Collier (2011) in the context of a medical alerting system for outbreaks of infectious diseases. While some work exists on cross-lingual information extraction of news articles (Steinberger et al. 2009), to our knowledge, cross-lingual bootstrapping of classifiers for event detection on twitter has not been investigated before.

In the area of Geospatial Information Systems (GIS), techniques have been developed that take Web 2.0 information and make it available for display (L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli, 2011; Nunex-Redo, Diaz, Gil, Gonzalez and Huerta 2011). These systems use only geo-tagged tweets and thus produce useful, but limited, thematic reports of temporal Web 2.0 activity on specific disaster-related topics. Some work has been performed on natural language processing (NLP) of tweets, and the extraction of political regions found in ontologies and gazetteers such as geonames.org (MacEachren, Jaiswal, Robinson, Pezanowski, Savelyev, Mitra, Zhang and Blanford, 2011; Shi and Barker, 2011). These produce tweet spatial frequency maps but only operate using well-formed region names. This does not take into account the majority of twitter users who use local place names, street names and region names.

ARCHITECTURE

The TRIDEC data fusion framework (Middleton, 2011) supports a 'forest of classifiers' type approach, running several social media processing services controlled by an OGC Sensor Planning Services in parallel. Results are collated for publication of subsequent real-time updates (cf. Figure 1). The multi-lingual analysis TweetComp1 module is our first prototype version deployed for initial evaluation (cf. Figure 2). We separate data (MySQL tables) and metadata (OWLIM triples) allowing efficient data publication, and use scalable middleware (APACHE QPID) to communicate real-time reports to a decision support system (command and control UI).

DATA SETS

Our dataset consists of tweets about two events gathered by use of the Twitter Streaming API to track a list of keywords. The annotation task (i.e. the classification of the events) was jointly carried out by two annotators.

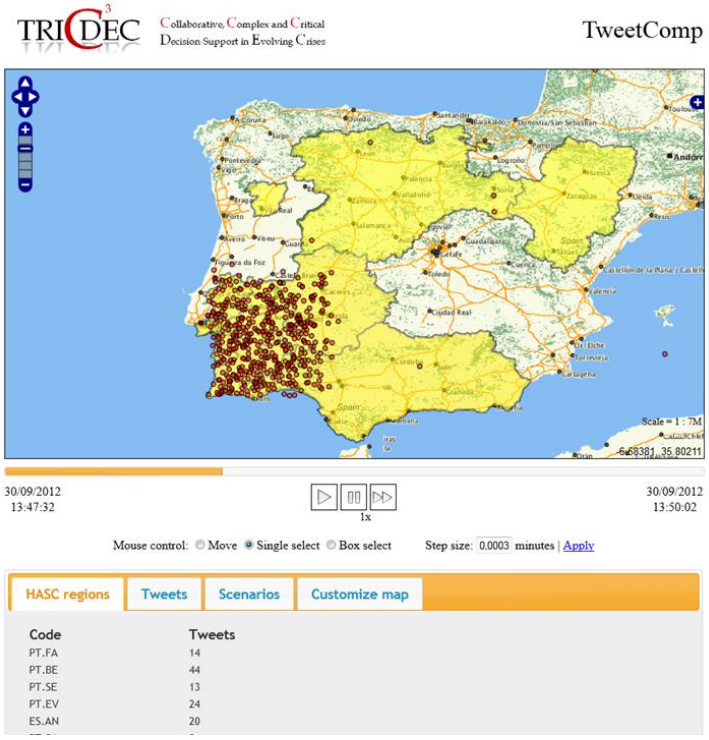

Figure 2: **GeoServer real-time visualization of tweet activity**

Dataset 1 - Tsunami in Philippine's 31/08/2012. [**12:47:32.0 UTC** + 7 hours]. Initial training was based on hotspots of seismic activity in the Mediterranean Sea during 2012 with tweets predominantly in *en, fr, el, it, tu.* Further analysis uses a real tsunami event , which illustrates the complexity of the MTC task and demonstrates the aspect of multilinguality in tweets. The monitoring process with a keyword list on earthquake-related terms produced a total of 558.126 unique tweets in languages such as *en, es, pt, fr, el, tl, ms, id, sl, it, ro, tu.*

Dataset 2 - Flooding in New York from storm sandy 29/10/2012. The 'super storm sandy' event has provided a lot (up to 20,000 per day for 'flood' keywords) of English language tweets relating to storm damage, flooding, power outages and subsequent inundation effects. This event has excellent media coverage, and there is a lot of ground truth data available from

NOAA satellite flood impact assessments through to lists of verified incidents by major media organizations.

SPECIFIC COMPONENTS

Focused Crawling (FC)

Goal: Motivated by the need to provide accurate and complete data, and thus unlike existing research which examines data collected from the use of a static set of pre-determined keywords, we provide a method for adaptively collecting from the Twitter stream to retrieve maximal amount of relevant information.

Algorithm approach: The challenge is to identify tags that, without the use of the original keywords, in the large, noisy and dynamic social network data world, generate content related to the event in question.

We perform the **following experiments** for refinement of adaptive crawling algorithm:

I. As a baseline, a dataset is collected using basic, easily generated terms (i.e. earthquake, shaking, etc.).
II. A more comprehensive dataset is collected by extending the collection process to include the N most popular hashtags, $H_t = \{h_1, h_2, \ldots h_m\}$. This is calculated using a sliding window approach. A keyword aggregation algorithm is used to update H at every time Ttf. The algorithm maintains two lists that record the frequency of each individual hashtag in both the entire collection period $FH = \{h_1,freq(h_1), h_2,freq(h_2), \ldots ,h_m,freq(h_n)\}$ and current timeframe , $FH_{tf} = \{h_1,freq(h_1), h_2,freq(h_2), \ldots , h_n,freq(h_o)\}$. $H_{t+1} = FH_{tf} + \{p \in FH: p \leq o - n\}$ is the combination of the two lists.

Expected outcomes: Focused crawling will produce a more complete dataset that more accurately captures all relevant aspects of the scenario. Validation of our approach will consist analyzing the information gain of the adaptive crawler both in terms of positive information (relevant information collected as a result of the inclusion of a related keyword not initially used for data collection) and negative information (noise introduced by the inclusion of a non-relevant search term)

Trustworthiness Analysis (TA)

Goal: In dealing with vast amounts of data generated by tweets a methodology for reducing or preferring certain users or classes of users is required. In situational awareness, efficient management of the situation will depend

on isolation of the relevant information. A tweet that belongs to a news agency is fully trusted. However, it may or may not be influential. Similarly, a tweet from a twitterer immersed in the event may be influential but not trustworthy. Influence and popularity of a user in a social network such as twitter depends on the type of audience engagement. In order to reduce or isolate user classes the influence measures will be calculated based on a variety of social network measures will be examined such as influence, activity, use of single topic, and language amongst the dependencies of influence, both [15] and [16] share the number of followers, Re-tweets, and mentions as building blocks of influence and popularity measurements. It was therefore decided to put the following three measures as the starting point for measurement of trust.

Algorithm approach: This algorithm analyses the users on a tweet by tweet basis and requires an initial setup period during which the core network graph will be constructed. After the construction of the network graph, core social network statistics (indegree, outdegree, betweenness) and a variety of influence measures will be calculated on a user by user basis. These statistics will subsequently be used to determine user classes using both hierarchical and partitioning clustering methods.

In order to evaluate these user classes standard metrics such as inter and intra-cluster distance will be calculated. Cluster meaning will be evaluated to determine the effectiveness of such a technique to map user cluster classes.

Expected outcomes: The outcome of this component is a set of user classes that will inform the other components. We plan to initially experiment using the New York 2012 flooding event by hand-annotated ground a ground truth set of users classes. These will then be compared to the ones created by our clustering methods.

Multilingual Tweet Classification (MTC)

Goal: For the analysis of tweets that carry information on situational awareness, many languages have to be covered, for which the number of training instances might not be large enough or not be available at all. For instance, in the Philippines Tsunami event, there was no initial training data for the languages (*Portuguese, Spanish, Tagalog, Malay*, and *Indonesian)*. To overcome the problem of data scarcity, Cross-lingual Text Classification (CLTC) algorithms can be used to gain as much information as possible from the monolingual data and transfer this knowledge to other languages.

Algorithm approach: We elaborate on the profile-based algorithm to CLTC using Machine Translation (MT) and human translation of the tweets. However, our set of monolingual classifiers can be trained beforehand, so that no translation effort will be required at the time of processing.

We perform the **following experiments** for classification of multilingual tweets:

I. As a baseline, classifiers are trained separately on the monolingual data. We adopt a vector space representation of tweets, comparing multi-nominal Naïve Bayes to SVM and Random Forest classifiers. It uses a variety of features, including stylistic features as well as lexical-semantic features.
II. We test if leveraging the training data by additional translated tweets (using Google Translate) results in improved monolingual classifiers. It has been shown by Bel et al. (2003) that although MT is not perfect, it is sufficient for the purpose of text categorization.
III. We leverage the data by means of human translation of the top 500-terms that appear in the monolingual classifier profiles.
IV. In order to avoid the language detection step, we also test a polylingual classifier. We use Latent Semantic Analysis to reduce the feature space.

For I. to III., we use ensemble learning and compare two different weightings to enhance the overall classification, i.e. a) giving priority to the main languages used in that area vs. b) favoring the monolingual classifiers with the best performance.

All experiments were performed using Weka. We used different sets for training (Earthquake Events in the Mediterranean Sea) and testing (Philippines Tsunami event).

Expected outcomes: It seems that our approach is most effective when few labeled tweets are available to learn an individual classifier and oversampling is required to handle rare classes which are to be predicted with high accuracy.

Location analysis/Geo-parsing (LOC)

Goal: The goal for location analysis is to extract from a tweet textual content inundation effects relating to coastal roads, places and regions thought to be at risk of any potential Tsunami or general flooding. We want to provide a set of real-time local situation assessments for the decision support system.

Algorithm approach: We first download gazetteer and volunteered geographic information (VGI) from sources

such as GoogleMaps, OpenStreetMap and Geonames. These local place names and street addresses are tokenized into sets of possible N-gram phrases using the NLTK toolkit. Tweets obtained from a Twitter crawler are then tokenized and the resulting N-gram phrase sets matched to the location N-gram phrase sets. Each match associates a specific tweet with a specific place, address or region.

Experiments: Regular statistical reports are generated over a rolling time window (configurable between 10 - 60 minutes) and lists the mentioned place names and street addresses in this period broadcast as geospatial 'hot spot' cluster summaries formatted as either KML (GoogleMaps) or shapefiles (GeoServer). These clusters can be clicked on to review the tweets themselves for human inspection of the results.

Expected outcomes: We plan to initially experiment using the well reported New York 2012 flooding event as our ground truth. This work will compare locations matched by our system over the flooding period with a ground truth consisting of (a) Google crisis map data deriving from NOAA aerial assessment of inundation and (b) verified flood incident lists compiled by the Guardian newspaper. We will look at the quality of the place/address matches, false positives and the effect of re-tweets and tweets from people outside the area.

Subsequent experiments will test how language-independent our approach is, using place names and addresses downloaded from coastal in Turkey and Portugal. Since gazetteer and VGI is written in the local language of each area of interest we hope that our approach will scale to at least Western regions.

CONCLUSION

We have given an overview of the functionality of our DSS for monitoring crises and detailed the integration of four modules from the fields of text mining and network analysis to classify tweets relevant for situational awareness. The multi-lingual analysis TweetComp1 module is our first prototype version deployed for initial evaluation (Zielinski, Bügel, 2012). Current ongoing work focuses on the scalability of the system, (i.e. testing if our algorithms can handle the volume of tweets in a typical crisis situation in real-time), the accuracy of the system, (i.e. evaluation and improvement of the overall accuracy of our algorithms by validating over ground truth data), and the usability as a component in a realistic monitoring system for decision makers who need to get an overall assessment of the situation. As next steps we aim to carry out a user study and incorporate the feedback of the operators on duty into our initial design of the visualization component that regularly updates info-graphics on a map.

REFERENCES

1. N. Bel, C. Koster, and M. Villegas. (2003) - Crosslingual text categorization. *Proceedings of European Conference on Digital Libraries (ECDL)*.
2. N. Collier. 2011. Towards cross-lingual alerting for bursty epidemic events. J Biomed Semantics.
3. A. Zielinski, U. Bügel (2012) - Multilingual Analysis of Twitter News in Support of Mass Emergency Events. Ijiscram.
4. R. Steinberger, B. Pouliquen, E. van der Goot (2009) - An introduction to the Europe Media Monitor family of application. Proceedings of SIGIR.
5. T.Sakaki, M. Okazaki, Y. Matsuo (2010) - Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of WWW.
6. J. Li, H.R. Rao (2010) - Twitter as a rapid response new service: An exploration in the context of the 2008 China earthquake. EJISDC.
7. M. Nunex-Redo, L. Diaz, J. Gil, D. Gonzalez, J. Huerta (2011) - Discovery and integration of Web 2.0 content into geospatial information infrastructures: A use case in wild fire monitoring. ARES.
8. L. Ferrari, A. Rosi, M. Mamei, F. Zambonelli (2011) - Extracting urban patterns from location-based social networks. LBSN.
9. A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford (2011) - SensePlace2: GeoTwitter analytics support for situation awareness. VAST.
10. G. Shi, K. Barker (2011) - Extraction of geospatial information on the web for GIS applications. IEEE ICCI'CC.
11. K. Starbird, L. Palen (2012) - (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. CSCW.
12. L. Bifet, G. Holmes, B. Pfahringer (2011) - MOA-TweetReader: real-time analysis in Twitter streaming data. DS.
13. M. Mendoza, B. Poblete, C. Castillo (2010) - Twitter under crisis: can we trust what we RT? SOMA.
14. A. Gupta, P. Kumaraguru (2012) - Credibility Ranking of Tweets during High Impact Events. Proceeding of the 1st Workshop on Privacy and Security in Online Social Media. ACM.
15. S. E. Middleton, Z. A. Sabeur (2011) - Knowledge-Based Service Architecture for Multi-risk Environmental Decision Support Applications. ISESS, Springer.