Template-Based Information Extraction System for Detection of Events on Twitter

Mr. Steven Toes and Dr. Majdi Owda

School of Computing, Mathematics & Digital Technology
The Manchester Metropolitan University, Chester Street,
Manchester, M1 5GD, UK

Telephone: (+44) 0161 247 1520

E-mail: steven.toes@stu.mmu.ac.uk, m.owda@mmu.ac.uk

Abstract

With the number of prosecutions for crimes on online social networks rising each year combined with the volume, velocity and variety of the data produced on these online social networks, it is almost impossible to effectively discover instances of crime or criminal behavior manually. Past work has been conducted in the area of event detection using Twitter, including TEDAS (Li, et al., 2012) and Jasmine (Watanabe, et al., 2011). This paper proposes a novel solution allowing for real-time monitoring of geo-tagged posts on Twitter using information extraction techniques to discover relevant information. This approach has had promising results, returning a number of relevant results from within the data set used in the evaluation.

1 INTRODUCTION

Online social networks (OSNs), such as Twitter and Facebook, offer users the ability to instantly publish information of a wide variety. Much of this data can have locations attached to it, showing almost exactly where they were published using a GPS-equipped smartphone. These OSNs are typically loosely moderated by service operators, with the majority of moderation that occurs being a reaction to offensive or inappropriate content being reported by other users. In part this may be due to the service operators wishing to promote free speech, but also due to the impracticality of monitoring the huge amount of data produced on a daily basis for such content.

With the increasing use of a variety of social networks, from Facebook to Twitter, there has been an increase in crime that is facilitated by, or committed on, them. In 2008, according to statistics from a Freedom of Information request (BBC, 2012), there were 556 reports and 46 people were charged with crimes where an online social network was a factor. By 2012 this had risen to 4,908 reports and 653 people charged (BBC, 2012).

The primary challenges for automated collection and processing of posts on social networks are the volume, velocity and variety of the data produced. In particular, the volume and velocity of the data makes it impossible to manually examine – for example, Twitter users create over 400 million posts per day (Farber, 2012). As such, an automated approach is required to effectively examine the data.

There have been many different approaches to detection of events on Twitter proposed. TEDAS (Li, et al., 2012) is designed to detect events from shootings to vehicular accidents. These events are classified by type and presented on a map, representing each event as a marker with an icon based on its classification. Jasmine (Watanabe, et al., 2011) is another event detection system, with the intention of detecting more local events with support for geo-tagging status updates based on their contents. Events are detected by identifying groups of related Tweets with the same theme that are both generated within a short time and geographic area. In addition to events, Twitter has also been used as a method of discovering breaking news as posted by users (Sankaranarayanan, et al., 2009). Like other work, these are visualized on a map.

All of these systems use automatic grouping of posts, based on the posts being published in a short timeframe in a geographic area to identify events. They offer no ability to explicitly define rules that can be used to monitor posts for specific events in real-time.

In the paper *GeoIntelligence: Data Mining Locational Social Media Content for Profiling and Information Gathering* (Hannay & Baatard, 2011), a system for collecting and mapping of Tweets containing geo-locational data is demonstrated. The system allows for searching of associated meta-data such as the text of a Tweet, location or username. However, the interface is rudimentary and provides no broader overview of search results or the ability to store searches for repeated use nor does it appear to offer real-time monitoring. Additionally, they did not offer the ability for the prototype to be evaluated when contacted.

The act of collecting Tweets in itself is something that can pose challenges depending upon the collection methodology chosen. Because of this, most researchers devise their own (often similar) methodologies. A standardised architecture was proposed for the collection and search of Twitter posts (Oussalah, et al., 2011) in a pre-defined geographic area that is denoted by the upper and lower latitude-longitude coordinates of a rectangular bounding box. It, like other work mentioned, allows for searching of collected Tweets based on a variety of criteria such as the textual content of a Tweet. Their proposed system makes use of Twitter's Streaming API. This theoretically reduces the impact of rate limiting, but in practice still only returns a tiny fraction of publically available Tweets. It does, unlike others, offer a broader overview of the returned Tweets, with a table being located to the right of the map view. However, the results display appears to offer no ability to further refine the results.

The ability to geo-tag posts on online social networks also poses privacy risks. While a single post may be considered innocuous and harmful, when a person's posts are taken as a whole it becomes very easy to infer information based on the clustering of points when their posts are visualized on a map (Friedland & Sommer, 2010). For example, the location of their home - and when they are most likely to not be present at it. Through a combination of this location information and the use of the API offered by FourSquare (a social service allowing users to 'check in' to locations), it was proven possible to profile a user and often accurately determine their future location based on previously visited locations (Li, et al., 2012).

This paper proposes a novel solution for real-time event monitoring of geo-tagged posts on Twitter. This novel methodology utilizes template-based information extraction in order to specify a crime event, the results of which are visualized on a map.

2 INFORMATION EXTRACTION

Information Extraction (IE) is the process of extracting structured information from large and unstructured sources, such as blogs, social networks or newspapers (Appelt, 1999). IE is used in a wide variety of tasks including the construction of academic citation databases such as CiteSeer (Lawrence, et al., 1999), the detection of potentially illegal activity on financial discussion boards (Knott & Owda, 2012), and the summarizing of patient records for healthcare delivery systems (The University of Sheffield, 2011).

Information Extraction can be categorised in to two types of approaches (Appelt, 1999). The first approach consists of producing rules that will be used when running the extraction process. While this is labour intensive, it is best used when the party tasked with rule creation is familiar with the target domain and allows for better creation of rules to target a specific event. The second approach is more automatic, consisting of training a system to generate rules through providing it with training data that allows for learning. While less labour intensive than the previous approach, this will not necessarily allow for as fine-grained targeting. For the purposes of the proposed system, the former approach was chosen.

3 ARCHITECTURE

Figure 1 shows the architecture of the novel solution and how each component interacts within the system as a whole. The sections below will address how each component functions.

A user selects a Template designed for a specific domain of interest – for example, harassment of other Twitter users. The Gatherer Engine queries Twitter for Tweets within, for the purposes of the research, a 30-mile radius of Manchester, UK. The results of these queries are processed and stored in the database for analysis. The Reasoning Engine uses the Template to populate rules that define what to look for within Tweets stored in the database. The Query Builder executes the rule-set on the database, and returns the results to the user for analysis.

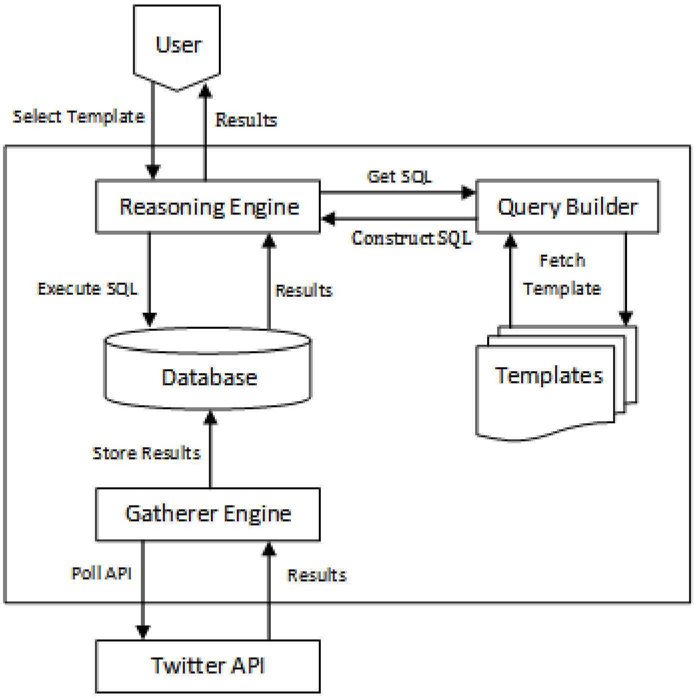

Figure 1: Prototype Architecture

3.1 Gatherer Engine

The gatherer engine component is responsible for collecting and parsing Tweets containing location information. To accomplish this it regularly queries Twitter's API. Due to API restrictions imposed by Twitter, the number of Tweets that can be gathered per hour is limited meaning that some Tweets are likely to be missed – this is an unavoidable consequence of using the API.

3.2 Database

The system is backed by a database that stores and manages the data set of Tweets and Users. It also stores the templates that can be executed against the data set.

3.3 Templates

The templates contain rule-sets defined by users, typically intended to target a particular domain of interest. In order for a template to be successful, they should be well-researched and contain rules relevant to the chosen domain in order to increase the likelihood of the results matching the chosen domain.

The rules use regular expressions to specify matching terms, in order to allow flexibility. A template can consist of any number of these expressions, allowing for both fine-grained and broader matches.

Figure 2 shows the template selection dialog, where users can run a chosen template. In addition, beneath this, are further options that allow refinement of search parameters such as location and time – allowing for templates to be reused by changing these parameters.

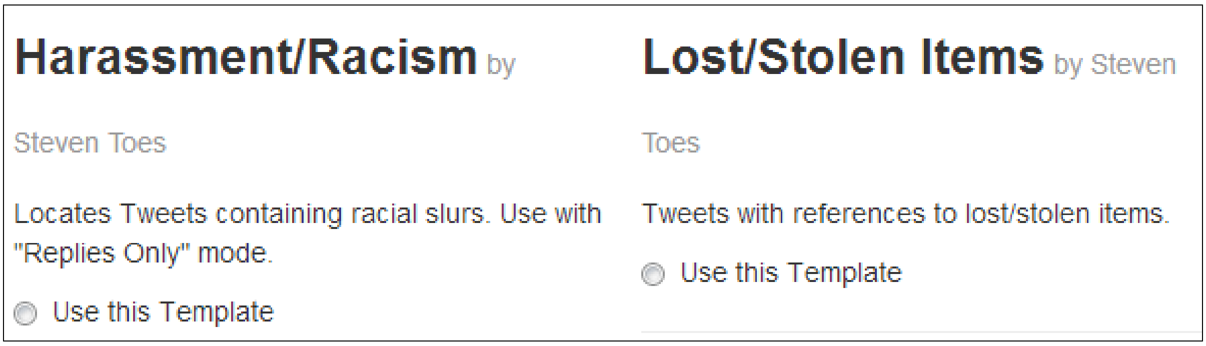

Figure 2: Template Selection

Following execution, the user is shown the results of their template – seen below in Figure 3. Each marker represents a tweet, and a broader overview can be obtained from the tabular display to the right of the map. This allows the user to quickly identify tweets of interest.

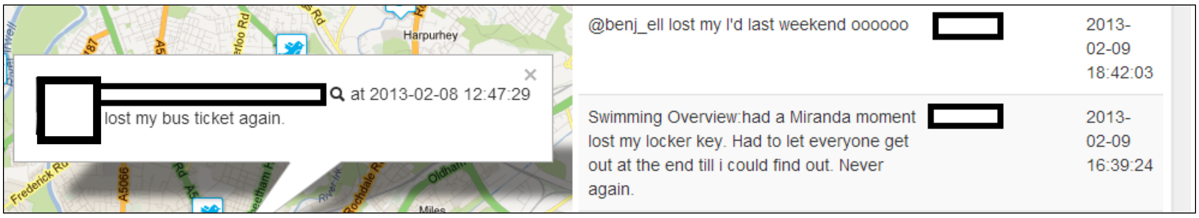

Figure 3: Template Results

3.4 Query Builder

This component is responsible for interpreting the rules set out in Templates in to SQL queries to be executed on the database, and is the main component that sits between the user and the data set.

3.5 Reasoning Engine

The Reasoning Engine interprets the rules from Templates, and passes them to the Query Builder in order to be executed.

4 Discussion

Figure 4 shows the number of unique Tweets collected over the 33 day collection period. An average of 7,000 Tweets was collected per day from within a 30-mile radius of Manchester, United Kingdom.

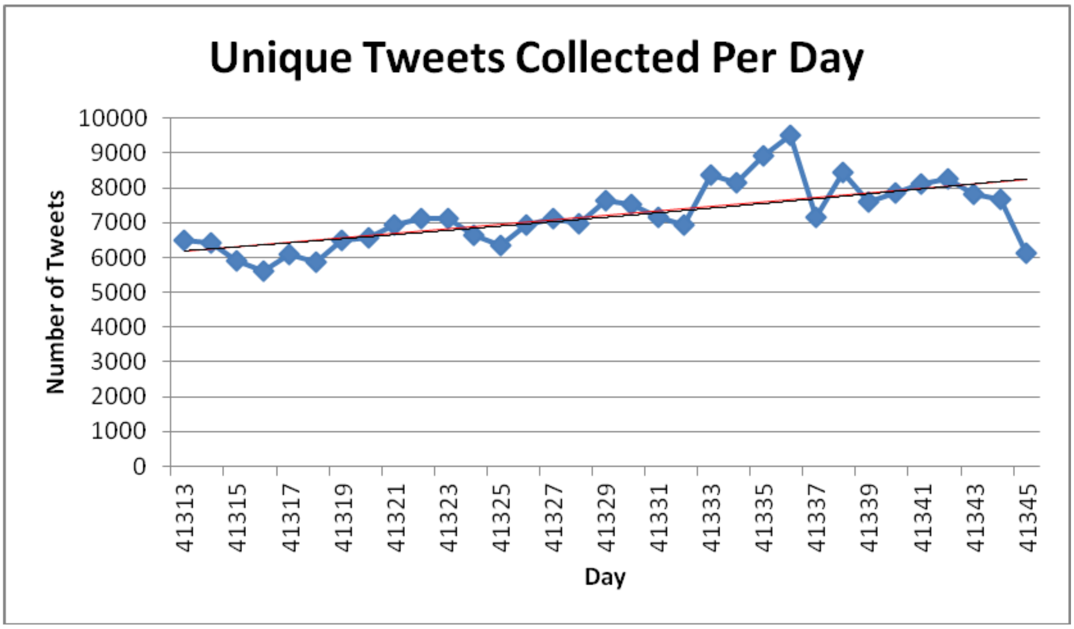

Figure 4: Tweet Collection

Statistics from a template designed to capture posts relating to the loss or theft of personal belongings is below, in Figure 5. This was executed on a data set of approximately 237,000 Tweets.

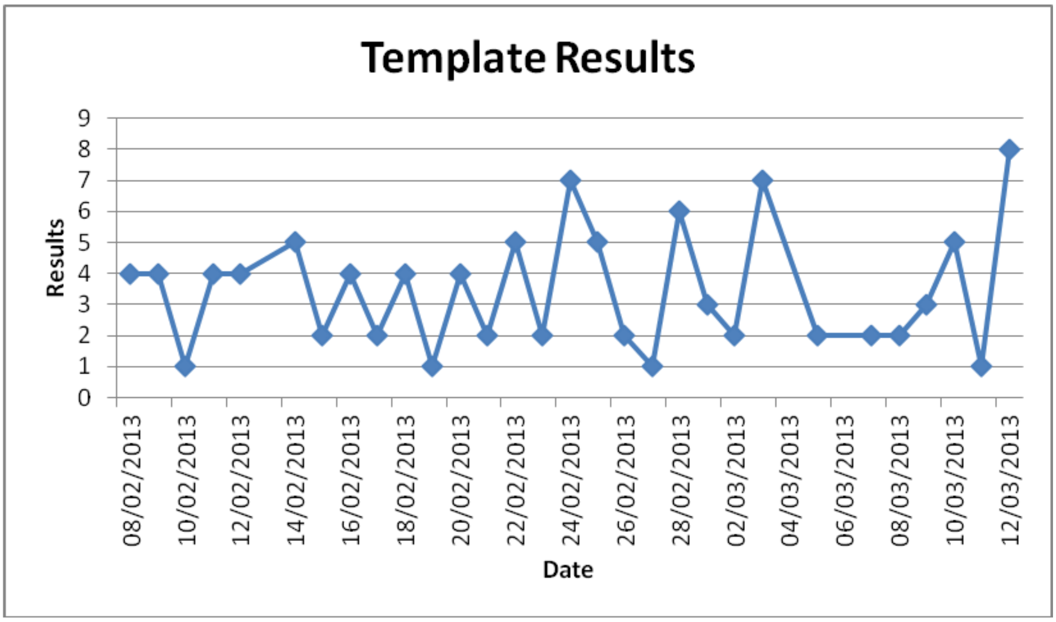

Figure 5: Template Results Graph

The prototype solves the issue of dealing with the volume of data with the ability to create templates that can be used to analyse large volumes of data. However, it does not fully address the variety of data, as it is concerned solely with geo-tagged Tweets and does not deal with any other information attached to these such as pictures or video.

5 Conclusion

This paper highlights the creation of a novel system for the real-time detection of events on Twitter, through the use of template-based Information Extraction techniques. The system returned promising results relating to the loss or theft of personal property, each containing a location – something that could be of great use to an investigator to take forward.

6 REFERENCES

Appelt, D. E., 1999. Introduction to Information Extraction. *Ai Communications,* 12(3), pp. 161-172.

BBC, 2012. *Huge rise in social media 'crimes'.* [Online]
Available at: http://www.bbc.co.uk/news/uk-20851797
[Accessed 20 March 2013].

Farber, D., 2012. *Twitter hits 400 million tweets per day, mostly mobile.* [Online]
Available at: http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/
[Accessed 10 April 2013].

Friedland, G. & Sommer, R., 2010. *Cybercasing the Joint: On the Privacy Implications of Geotagging.* Washington, D.C., s.n.

Hannay, P. & Baatard, G., 2011. *GeoIntelligence: Data Mining Locational Social Media Content for Profiling and Information Gathering.* Perth, s.n.

Knott, E. & Owda, M., 2012. *The detection of potentially illegal activity on financial discussion boards using information extraction.* London, UK, s.n.

Lawrence, S., Giles, C. L. & Bollacker, K., 1999. Digital libraries and autonomous citation indexing. *Computer,* 32(6), pp. 67-71.

Li, R., Lei, K. H., Khadiwala, R. & Chang, K. C.-C., 2012. *TEDAS: a Twitter Based Event Detection and Analysis System.* Washington, DC, USA, IEEE, pp. 1273-1276.

Li, W., Eickhoff, C. & de Vries, A. P., 2012. *Want a coffee?: predicting users' trails.* Portland, Oregon, ACM, pp. 1171-1172.

Oussalah, M., Bhat, F., Challis, K. & C., S., 2011. A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems,* Volume 37, pp. 105-120.

Sankaranarayanan, J. et al., 2009. *Twitterstand: news in tweets.* Seattle, Washington, USA, ACM, pp. 42-51.

The University of Sheffield, 2011. *GATE.ac.uk - ie/index.html.* [Online]
Available at: http://gate.ac.uk/ie/
[Accessed 3 March 2013].

Watanabe, K., Ochi, M., Okabe, M. & Onai, R., 2011. *Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs.* Glasgow, Scotland, ACM, pp. 2541-2544.