Location-Based Insights from the Social Web

Yohei Ikawa
IBM Research – Tokyo, IBM
Japan Ltd.
yikawa@jp.ibm.com

Maja Vukovic
T.J. Watson Research, IBM
Cooperation
maja@us.ibm.com

Jakob Rogstadius
University of Madeira
jakob.rogstadius@m-iti.org

Akiko Murakami
IBM Research – Tokyo, IBM
Japan Ltd.
akikom@jp.ibm.com

ABSTRACT

Citizens, news reporters, relief organizations, and governments are increasingly relying on the Social Web to report on and respond to disasters as they occur. The capability to rapidly react to important events, which can be identified from high-volume streams even when the sources are unknown, still requires precise localization of the events and verification of the reports. In this paper, we propose a framework for classifying location elements and a method for their extraction from Social Web data. We describe the framework in the context of existing Social Web systems used for disaster management. We present a new location-inferencing architecture and evaluate its performance with a data set from a real-world disaster.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications; J.4 [Computer Applications]: Social and Behavioral Sciences.

General Terms

Experimentation

Keywords

Microblog, geolocation analysis, text analysis.

1. INTRODUCTION

The Social Web is becoming a prevalent means for information sharing. Many people spend over 25% of their online time using social networking sites [1]. In the domain of disaster management, several systems, including Ushahidi [2], TweetTracker [3], CSIRO [4], Twitcident [5] and CrisisTracker [6] have been developed to help humanitarian agencies and disaster relief workers with situational reports distilled from large volumes of Social Web data. Beyond situational reports, disaster agencies may use such systems to operate in volatile environments for various tasks, such as recruiting volunteers, providing emergency contacts, or making decisions about distribution channels.

These systems utilize and provide location information at various accuracy levels and operate over different geographical scopes (e.g. a street, a suburb, a city, a country), and work with different Social Web sources (e.g. Twitter, Facebook, etc.). For example, Ushahidi relies on the users who are reporting the events to geo-tag them, while Twitcident looks at low-granularity data in smaller geographical areas (such as parts of a city). Though geo-tagging is supported by many Social Web systems, the challenge is that most sources do not automatically provide such location information.

Social Media messages contain different types of locations, such as place names appearing in the message, a location from which the message was sent, and so on. When we geo-locate a message, we should consider which location type is appropriate. For example, a message “Syria forms new government, retains defense minister in new cabinet.” should be associated with “Syria”, a place name appearing in the message, whereas a message “Shaking!” (following an earthquake) should be associated with the location where the message came from.

Location data varies not only in its type, but also in how the associated data is aggregated and analyzed, resulting in many complexities that must be addressed by an effective location extraction engine. For example, the location information can be used to filter or segregate events. On the extraction side, there are various challenges in extracting a location and its type from a single message, so as to recognize and analyze the relationships between information on the Social Web and geo-locations from various viewpoints.

In this paper we analyze and classify the types of location-based information a disaster management tool may provide or consume. To put our study in context of disaster management tools, we integrated our location inferencing engine with the CrisisTracker system, which clusters Twitter messages (based on their textual similarity) to construct cohesive stories. The key challenge is how to enable CrisisTracker to infer each relevant location from multiple, similar messages. The main contributions of this paper are a location-use case classification framework and the architecture of the location-inferencing engine. We present an evaluation of 182 Twitter messages to show the performance of our location inferencing engine.

The next section describes the characteristics of locations and how they are represented in text, differentiating the user’s location from that of the event. Section 3 discusses applications of

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

location in the CrisisTracker system. Section 4 describes the location inferencing engine and the evaluation results. Section 5 puts our work in the context of the state of the art. Section 6 concludes and outlines future work items.

2. Inferring Locations from the Social Web

In this section we identify four types of locations: Locations in Text, Focused Locations, the User's Current Location and the User's Location Profile. We define each type and present its relevance to disaster management.

Locations in Text

Locations in Text is a location type for place names described in a target message (e.g. New York). In general, it contains Points of Interest (POIs), terms that are place names or strongly associated with specific locations (e.g. ABC hotel, Golden Gate Bridge). For example, a message "Syrian group says 3 intelligence officers killed. Syria's Assad faces growing rebel, foreign threat: LONDON (XYZ Press)" contains "Syria" and "London" as Locations in Text. Locations in Text help to understand the geographic distribution of the locations mentioned in a target message. For disaster management, we can locate the places relevant to the unexpected events or incidents, and create map views to understand the geographic characteristics.

However, it does not always hold that all of the locations appearing in the target message are relevant to the main topic of the message. In the above example, "London" is a place name, but it is not strongly relevant to the news the message describes. Therefore, we should consider "Focused Locations", our second type of location, to more appropriately locate references on a map.

Focused Locations

Focused Locations is a location type that represents the relevant locations of events or incidents described in a target message. Focused Locations are identified by selecting locations of interest from Locations in Text. It is possible that there is more than one Focused Location even for a single event. For example, a message "Russia Sending Air and Sea Defenses to Syria." contains two Locations in Text ("Russia" and "Syria"), and both of them are also Focused Locations. In addition, there may be no Focused Locations if a target message describes an incident involving an organization that has no specific locations (e.g. the United Nations, or other multinational organizations). By using Focused Locations, we can locate information on a map more appropriately.

User's Current Location

The User's Current Location is a location type that represents a location from which a message was sent. Since the Social Web, including microblog services like Twitter, is easily accessed with smartphones, people can send messages frequently even when they are outdoors. By identifying a User's Current Location, we know the location where the message originated even if the message contains no direct clues about its location. For example, the text of a message "I got caught in a traffic jam…" does not contain any clues about the location where the traffic jam occurred. We might obtain the User's Current Location from a geotag attached to the message, but most users choose not to attach geotags to their messages. However, in this example, we may be able to infer the User's Current Location by using past messages [7] about routine and repeated travel.

User's Location Profile

User's Location Profile is a location type that represents locations with close ties to a user. One of the key characteristics of the Social Web is that the author of each message is labeled. By analyzing the relevant locations of users, we can recognize areas where people are reacting to unexpected events or incidents. The primary location of a User's Location Profile is the location of the user's home. Some Twitter users disclose their home location at the level of a country or city in their user profiles. Although most users do not disclose their home location, algorithms have been proposed for inferencing home locations by using the users' past messages [8] or social graphs [9]. In addition to a user's current home location, previous home locations and frequently visited locations are also available in the User's Location Profile.

Table 1 summarizes the four location types and use cases. When we analyze geolocation information on the Social Web, we have to choose the location type appropriately depending on the purpose of the analysis. We can apply the location types not only for disaster management, but also for reputation analysis, location-based marketing, etc., since the concepts of the location types are common in the Social Web.

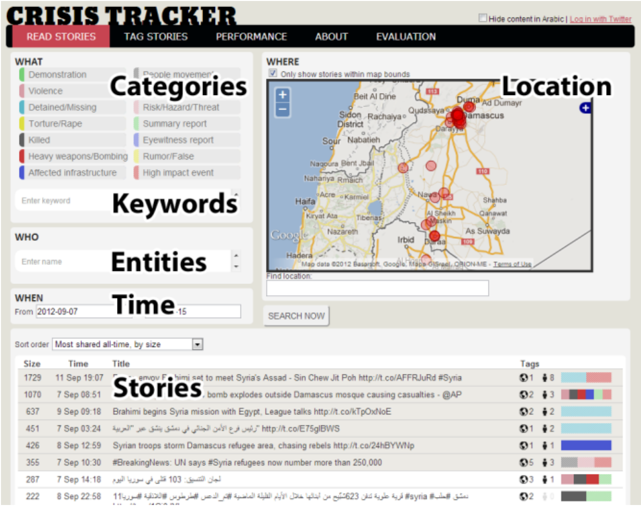
Table 1: Location types and use cases.

Location Type	Use Case for Disaster Management
Locations in Text	Recognize locations mentioned in a target text.
Focused Locations	Locate news events on a map by relevant sites.
User's Current Location	Detect the place where an event happens.
User's Location Profile	Detect areas where people are interested in incidents.

3. Location Utility in CrisisTracker

CrisisTracker is a Web-based system that automatically tracks sets of keywords on Twitter, and constructs stories by grouping related tweets based on their textual similarity. Beyond basic information such as timestamps, the system relies entirely on crowdsourcing to collect meta-data annotations for stories. Figure 1 shows the CrisisTracker interface. Users can filter stories by categories, keywords, mentioned named entities, and time. The map supports location-based filtering of the story list.

CrisisTracker also provides a tagging interface. A content curator can use CrisisTracker to infer the location of the story by reading the reports and following links to news articles, videos and image, and then tag the story on the map. In addition to location tagging, the user can categorize the story according to a set of instance-specific report categories, add named entities, merge it with similar stories, remove unrelated tweets from the story, or hide irrelevant (or misleading) stories.



Size	Time	Title	Tags
1729	11 Sep 19:07	envoy Brahimi set to meet Syria's Assad - Sin Chew Jit Poh http://t.co/AFFRJuRd #Syria	1 8
1070	7 Sep 08:51	, bomb explodes outside Damascus mosque causing casualties - @AP	2 3
637	9 Sep 09:18	Brahimi begins Syria mission with Egypt, League talks http://t.co/kTpOxNoE	2 2
451	7 Sep 03:24	رئيس فرع الأمن الجنائي في دمشق ينشق عبر "العربية" http://t.co/E75glBWS	1 2
426	8 Sep 12:59	Syrian troops storm Damascus refugee area, chasing rebels http://t.co/24hBYWNp	1 1
355	7 Sep 10:30	#BreakingNews: UN says #Syria refugees now number more than 250,000	5 3
287	7 Sep 14:18	لجان التنسيق: 103 قتلى في سوريا اليوم	3 1
222	8 Sep 22:58	دمشق #حلب #syria قرية طوية تدفن 623تشيع من ابنائها خلال الأيام القليلة الماضية #دير_الدص #طرطوس #اللاذقية #سوريا11	2 0

Figure 1: CrisisTracker: Location based filtering. [6]

4. Implementation and Evaluation

In this paper, we present the prototype architecture for detecting Locations in Text in Section 2, which is the base functionality for detecting other location types, as the first step of implementing the Location Inferencing Framework. We evaluated the prototype system by using real Twitter messages that mention the Syrian civil war.

4.1 Architecture

The system architecture for detecting Locations in Text is shown in Figure 2. The inputs are the messages and the outputs are the locations. To associate the detected locations with a map, the output includes coordinates. We use GeoNames [10] as the location database of this system. There are two components in the system: Location Name Recognition and Toponym Resolution.

4.1.1 Location Name Recognition

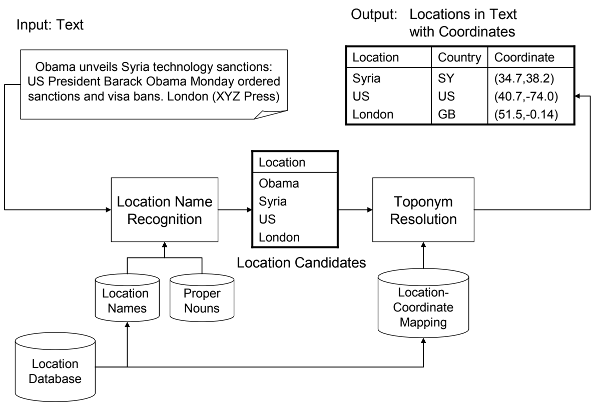
Location Name Recognition is a process that detects location candidates in the input text. To achieve high accuracy, it is not sufficient to extract expressions listed in a location names dictionary, since the locations can be proper nouns. In some cases, we can identify a reference as a location or non-location by simple linguistic rules. For example, an expression "Mr. Paris" is not for a city but is a person's name. We can use the mature technology for Named Entity Recognition to detect proper nouns such as locations, personal names, and organizations in the input text [11].

However, it is generally a challenging task to resolve whether or not an ambiguous term is a location name. In Figure 3, for example, "Obama" appears in the input text. If we knew it was a human being, then it is obvious that "Obama" is the name of the US president. However it is difficult for a system to identify it as a person's name from linguistic features alone, since "Obama" is also the name of a Japanese city. In this case, we can predict that most Obama references are to the person because the Japanese city is not a major city. Our system extracts terms that are possibly location names as location candidates, and resolves whether or not they are location names in the Toponym Resolution component.

4.1.2 Toponym Resolution

Toponym Resolution is a process that associates location candidates with location instances and assigns coordinates. Some location names represent many location instances. For example, London is not only the capital of England, but a city in Ontario, Canada. Therefore we have to identify a location instance to associate it with an actual geolocation. We also resolve location candidates as locations since some of location candidates are unlikely as location names (such as "Obama" in the example).

To associate a location name with an actual location instance, we calculate a confidence score for each possible location instance using the Location Popularity and Region Context. The score for the Location Popularity is based on the population of the location. We used the population data provided by GeoNames to calculate Location Popularity. The score for Region Context is based on areas that are focused on by the context of the message. The score of Region Context is higher when a location instance is in the country referenced in the target message. Then a confidence score for each location instance can be calculated by multiplying the Location Popularity and Region Context scores. After these calculations, the location instance with the highest confidence score is selected as the result of Toponym Resolution. If the highest confidence score is lower than a threshold, the location is evaluated as a non-location term.


Input: Text
Output: Locations in Text with Coordinates
Location Name Recognition
Location Candidates
Toponym Resolution
Location Names
Proper Nouns
Location-Coordinate Mapping
Location Database

Obama unveils Syria technology sanctions: US President Barack Obama Monday ordered sanctions and visa bans. London (XYZ Press)

Location
Obama
Syria
US
London

Location	Country	Coordinate
Syria	SY	(34.7,38.2)
US	US	(40.7,-74.0)
London	GB	(51.5,-0.14)

Figure 2: Architecture for detecting Locations in Text.

4.2 Evaluation Setup and Results

We evaluated the accuracy of the prototype for recognizing Locations in Text. For the evaluation, we sampled 182 real tweets mentioning the Syrian civil war as collected by CrisisTracker [6]. In this experiment, we used a subset of the GeoNames database: world cities with populations above 15,000 and all Syrian locations. To create a gold standard data set, all of the place names were manually extracted from the messages. We input each message into the system and evaluated the precision and recall for both Location Name Recognition and Toponym Resolution.

The evaluation results are shown in Table 2. The #appearance and #unique columns represent the total number of locations and the number of locations after removal of duplicate elements in the evaluation dataset respectively. The evaluation results are aggregated by location levels.

The results indicate that the system performs well for major place names and reasonably well for villages, even for Twitter messages (which are generally considered to be poorly formed text). One of the reasons for the high performance is that messages in our dataset are better-formed than average tweets, since the selected tweets were cluster centroids in stories from

CrisisTracker. Many of these centroids are excerpts from news articles or otherwise written to clearly describe breaking news, whereas other less popular versions of the story may be less clear. In contrast, the recall for the village names is worse than for the other location types. The poor recall of the villages is because some of the small villages are not included in the GeoNames database (often because of non-standard transliterations from Arabic to Latin characters).

In the process of examining tweets manually, we also found that remarkably few tweets contained any mention of the exact location where an event took place. For example, a tweet may mention a bombing near a police station in the capital, but without the specific police station. This means that a geo-inferencing algorithm, no matter how accurate, can only associate such a tweet with the entire city. However, in most cases precise location information was present in a linked-to resource, such as in the body of a news article or in the title or description of a YouTube video. Extending the proposed architecture to also include the external sources would likely lead to major improvements in performance, but this is future work.

Table 2: Location levels and evaluation results.

	Country	State	City/ Town	Village	Total
#appearance	250	39	41	12	342
#unique	20	7	11	8	46
Precision	0.996	1.000	1.000	0.917	0.994
Recall	0.992	1.000	0.927	0.750	0.977

5. Related Work

The concept of location types for a document was introduced in [12]. They mentioned that a document may have two types of geographical information associated with a source and a target, and they also attempted to identify focused locations from place names appearing in the target document. In this paper, we expanded the concept of location types by adding two types, User's Current Location and User's Location Profile to apply to Social Web messages, and presented its use cases in the context of disaster management.

One of the key issues of toponym resolution is how to decide the Region Context for each message. For instance, [13] introduced a model that distinguished a global lexicon known to all audiences and an audience-specific local lexicon, and proposed generic methods for inferring local lexicons for toponym resolution. In our work, we do not have to consider the locality beyond a message since the evaluation data is for an arbitrary audience. However, we should consider the locality for a more practical system.

6. Conclusion

As the volume of Social Web messages increases, the certainty for an individual event and its associated location decreases. In this paper we introduced four location types and presented use cases in disaster management. We presented an architecture for a location inferencing engine that addresses the challenges of locating the user and the event being reported given a set of Social Web messages. We evaluated the prototype implementation and demonstrate that it performed well for major place names and reasonably well for a specified location level. Future work should include extending the proposed system to also include external sources would likely lead to great improvements in performance.

7. REFERENCES

[1] "Social Networks/Blogs Now Account for One in Every Four and a Half Minutes Online | Nielsen Wire," Online: http://blog.nielsen.com/nielsenwire/global/social-media-accounts-for-22-percent-of-time-online

[2] Ushahidi, Online: http://ushahidi.org

[3] S. Kumar, G. Barbier, M. A. Abbasi, H. Liu. "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief," Demonstration Paper at 5th International AAAI Conference on Weblogs and Social Media, Barcelona, 2011.

[4] F. Abel, C. Hauff, G. J. Houben, R. Stronkman, K. Tao, "Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams," in Proc. International Conference on Hypertext and Social Media, Milwaukee, 2012, pp. 285-294.

[5] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power. "Using Social Media to Enhance Emergency Situation Awareness," Intelligent Systems, IEEE , vol. 27, no. 6, 2012, pp. 52-59.

[6] J. Rogstadius, M. Vukovic, C. Teixeria, V. Kostakos, E. Karpanos, J. Laredo. "CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness". IBM R&D Journal. (to appear).

[7] Y. Ikawa, M. Enoki, M. Tatsubori. "Location Inference Using Microblog Messages," in Proc. international conference companion on World Wide Web, 2012, pp. 687-690.

[8] Z. Cheng, J. Caverlee, K. Lee. "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users," in Proc. Conference on Information and Knowledge Management, 2010, pp. 759-768.

[9] C. A. D. Jr., G. L. Pappa, D. R. R. de Oliveira, F. de Lima Arcanjo. "Inferring the location of twitter messages based on user relationships," Transaction in GIS, 15(6), 2011, pp. 735-751.

[10] GeoNames, http://www.geonames.org/

[11] B. Sundheim. "Overview of results of the MUC-6 evaluation," in Proc. message understanding conference, 1995, pp. 13-32.

[12] E. Amitay, N. Har'El, R. Sivan, A. Soffer. "Web-a-where: geotagging web content," in Proc. ACM SIGIR, 2004, pp. 273-280.

[13] M. D. Lieberman, H. Samet, J. Sankaranarayanan. "Geo-tagging with local lexicons to build indexes for textually-specified spatial data," in Proc. International Conference on Data Engineering (ICDE), 2010, pp. 201-212.