CO-244

GEO-TWITTER ANALYTICS: APPLICATIONS IN CRISIS MANAGEMENT

MACEACHREN A., ROBINSON A.C., JAISWAL A., PEZANOWSKI S., SAVELYEV A., BLANFORD J., MITRA P.

Pennsylvania State University, UNIVERSITY PARK, UNITED STATES

ABSTRACT

In this paper, we introduce a web-enabled geovisual analytics approach to leveraging Twitter in support of crisis management. The approach is implemented in a map-based, interactive web application that enables information foraging and sensemaking using "tweet" indexing and display based on place, time, and concept characteristics. In this paper, we outline the motivation for the research, review selected background briefly, describe the web application we have designed and implemented, and discuss our planned next steps.

INTRODUCTION

Social media are becoming increasingly geographic. Following from this trend, maps that depict a wide range of data from geo-located social media are becoming fairly common. Most social media mapping research and most web maps of social media data thus far focus on the tasks of capturing, locating, and displaying data extracted from social media. This task itself is a challenge; for Twitter alone, the number of tweets reached 110 million per day in January, 2011 and that number is escalating rapidly (source: Forbes.com blog: http://tiny.cc/s8pbc). However, extracting meaning from these large, messy, but potentially important data sources is an even greater challenge that is just beginning to be addressed. One application domain for which the need is particularly important is crisis management; the potential usefulness of social media in this domain is reflected in the increased attention given to social media by major crisis management organizations (e.g., the Red Cross Emergency Social Data Summit held in August, 2010) and by those in the crisis management research community, e.g., (Palen et al., 2010; Starbird and Palen, 2011).

In this paper, we focus on efforts to address the challenge of social media information mapping and analysis more systematically, with an emphasis on potential applications for crisis management. More specifically, we apply a visual analytics perspective to develop and implement visually-enabled information foraging and sensemaking tools for leveraging data made publically available in social media. In this paper, we focus on the rapidly growing resource of Twitter tweets, both those that include geographic location and those for which location can be inferred through specific hashtags or by automated entity extraction methods.

BACKGROUND

In the work reported here, we have drawn upon a large volume of research that is relevant to addressing the challenges of geographically-enabled information foraging and sensemaking with text artifacts. Two primary streams within that research are: (a) geographic information retrieval / geotagging research focused on identifying text artifacts containing geographic information and extracting and geocoding that information and (b) geo/information visualization research focused on web-based, multiview interfaces for text artifact foraging and sensemaking. In this short paper, we highlight only a few selected research and technology advances that are directly relevant to geographically-enabled use of Twitter data.

In relation to technology, an increasing number of web applications have been created to map Tweets. Many of these have focused primarily on anchoring tweets to locations on a map (e.g., http://compepi.cs.uiowa.edu/~alessio/twitter-monitor-swine-flu/) or to depicting aggregate counts of Tweets from or about a place (e.g., http://www.planeteye.com/maps?set=map.4179). There have, however, been several innovative ideas on how to represent changing information from Tweets over time on maps (e.g., changing geographic distribution of chatter about the Oscars during the TV broadcast: http://www.neoformix.com/2009/OscarTwitterMap.html).

In relation to research, there is a very wide range of relevant efforts related to both of the research streams noted above. Here, we highlight just a few examples with emphasis on those with a cartographic and/or GIS component.

In relation to crisis management, the target application for our research, there is a fairly long history of use of information technologies (IT) and that history includes a wide range of mapping and GIS applications. But, as Zook, et al (2010) point out, until recently, most of the attention was on development of IT to support official government organizations supplemented by NGOs carrying out formal command, control,

and information dissemination operations. The crisis management community has recently begun to recognize the potential value of web 2.0 technology and related social media as a means to leverage volunteered information input, e.g., (Sakaki et al., 2010). At the same time, multiple challenges have been recognized related to volume, relevance, and quality of information, e.g., (Mendoza et al., 2010) together with the complexities of coordinating citizen activities (Starbird and Palen, 2011).

Work by Tomaszewski and colleagues (Tomaszewski, 2008; Tomaszewski and MacEachren, 2010) provides a conceptual approach for modeling geo-historical context (derived by applying entity extraction and relevance ranking methods to documents retrieved from news sources) and using representations of that context to underpin decision-making about humanitarian relief. In subsequent work building on the earlier context modeling ideas, Tomaszewski, et al (Tomaszewski et al., in press) design and implement tools supporting an iterative sensemaking process that leverages news reports to build a comprehensive understanding of human migration patterns in developing countries as input to infectious disease modeling. Complementary to this research, recent analysis by Vieweg, et al (Vieweg et al., 2010) provides insights about the relative frequency of place references and situation updates in tweets as a function of kind of crisis event. As these authors note, their work has the potential to inform strategies for automatic extraction of crisis-relevant information from tweets.

From a cartographic perspective (with attention to mapping tweets, but without a focus on crisis management) Field and O'Brien (Field and O'Brien, 2010) critique typical current map mashups as being relatively haphazard combinations of information. They contend that most work on mashups thus far has focused on how to capture data and render it on an overlay with little attention to cartographic design and map usability. An emphasis of their paper is on the challenges of designing maps of Tweets that communicate effectively and are usable. They provide insights on visual representation of multiple coincident points and on developing a real-time collaborative way to develop a common conceptual understanding of a problem in support of data capture. In relation to the latter, they used a class field exercise to demonstrate the potential of georeferenced Tweets to function as an effective collaboratory environment through which consensus c be built on how to carry out a complex real-world data collection task.

APPLICATION DESIGN AND IMPLEMENTATION

In this subsection, we outline the strategy for and describe the implementation of a geovisual analytics application focused on place-time-attribute based information foraging in the Twitterverse and visually-enabled sensemaking in support of crisis management with the information derived. This strategy integrates computational methods for capturing, storing, and indexing tweets with visual query and analysis methods. A primary contention that underlies our work is that a visual analytics approach is particularly relevant for dealing with Twitter because of the scale of data involved, the limited details in each tweet combined with potential to generate insights by combining many small fragments of information, plus the challenges in identifying both places and topics in short, cryptic tweets. No computational methods will be completely successful in tweet retrieval and interpretation. Thus, a tight integration between computational and visual methods is needed; the computational methods are essential to handle the scale of data and the visual methods are essential to help users deal with uncertainty about information quality and relevance.

Below, we sketch the overall challenges to leveraging Twitter to support crisis management (or other applications for which monitoring place-based activities, events, and attitudes is relevant) and we present an initial prototype geo-twitter analytics application: SensePlace2. (for information about the predecessor SensePlace environment, see (Tomaszewski et al., in press).) Specifically, SensePlace2 supports overview and detail maps of tweets, place-time-attribute filtering of tweets, and analysis of changing issues and perspectives over time and across space as reflected in tweets. The environment has been designed to integrate multiple text sources (e.g., news, RSS, blog posts), but we focus here just on tweets.

In subsections below, we provide details on the computational methods used to collect and process tweets and on design of the interactive web-based interface that supports foraging for and sensemaking with tweets relevant to place, time, and concept constraints imposed by a user. In the subsequent section, we discuss next steps in the research.

Collection and Computational Processing of Tweets

Tweets contain precise and relatively accurate information about when they were posted. They can contain precise and accurate where (location) information as well if the tweet comes from a GPS enabled device for which the user has opted in to geolocation. However, the proportion of users with geolocation turned on is probably still in the single digits. While it is possible that additional users will turn that feature on in

a crisis, using Twitter as input to geographically-enabled crisis management requires less explicit geographic information to be extracted as well, e.g., the location stored in user profiles, place references extracted from tweet content, and hashtags that are associated with places. Specifically relevant to crisis management, the Tweak the Tweet effort is prompting use of a #loc or #location hashtag during crisis events (Starbird and Palen, 2010) and http://tiny.cc/fb60g.

Attribute information in tweets is challenging to extract due to the combination of the 140 character limit on tweets (which prompts extensive use of abbreviations) and the dramatic variety of tweet content spanning a range that includes: individuals documenting mundane daily activities in their lives, through professionals alerting followers about events and information (e.g., the Director of FEMA tweeting about the challenges of social media security), government and non-government organizations making regular announcements (e.g., UNGlobalPulse announcing events and new stories or CrisisMappers.org announcing maps or training webinars), and advertisers using Twitter for marketing purposes.

SensePlace2 uses a crawler to systematically query the Twitter API for tweets that contain any topics deemed to be of interest. The current implementation of the system uses a set of keywords and phrases that our research team has proposed over time and that is added to as new events happen around the world. Queries for each term are run every day and each can retrieve tweets and auxiliary metadata (e.g., creation time, tweet id, user id etc.) in JSON format, which is then written to a file on disk. After parsing, each tweet is stored in a PostGresSQL database. Once tweets are loaded into the database, separate distributed applications analyze tweets for named-entities such as locations, organizations, persons, hashtags, URLs etc. These named entities are then written to separate tables such as an auxiliary location table, organization table etc. Lastly, locations that are extracted are then georeferenced using GeoNames. Once entities are identified and organized, a Lucene text index is generated that supports relatively fast full text querying as well as more advanced retrieval of relevant tweets within a geographic region and date range.

Visual Interface

To support geovisual analysis, we designed a coordinated, multiple-view interface for SensePlace2. The purpose of the SensePlace2 interface is to support an understanding of spatial and temporal patterns of activities, events, and attitudes that can be identified through analysis of our growing geo-located Twitter database. A key goal for this interface is to support an analyst's ability to explore, characterize, and compare the space-time geography associated with topics and authors in Tweets. This includes the ability to describe the geographic content associated with tweets as well as the locations where Tweets were reported by users that have enabled that feature. In many instances, these geographies will be quite different (for example, those who talked about the Haiti Earthquake versus those who are actually tweeting from Haiti in the aftermath). The default SensePlace2 interface includes a query window, map, time-plot / control, relevance-ranked list of tweets, and task list. The primary display views (map, time-plot/control, and tweet list) are dynamically coordinated. Each view is introduced below; then cross-view linking is discussed.

Query window: Users can enter single or multi-term queries and these can include place names. Each query retrieves a new set of information that is processed to populate display views.

Map: The map provides both overview, in the form of a gridded density surface representing all tweets that match the query, and detail in the form of point-based depiction of the most relevant 500 tweets. The density surface is generated for the globe and currently depicts frequency counts for tweets aggregated to 2 degree grid cells (grid resolution is flexible, but that flexibility has not yet been made accessible to users). A quantile classification scheme is applied, to allow comparison from one query to the next, and a sequential color scheme is used with dark=highest. It is likely that some locations for the top 500 tweets can have multiple tweets, thus those location are depicted with range-graded sizes for 1, 2-5, and >5 tweets from/about a place.

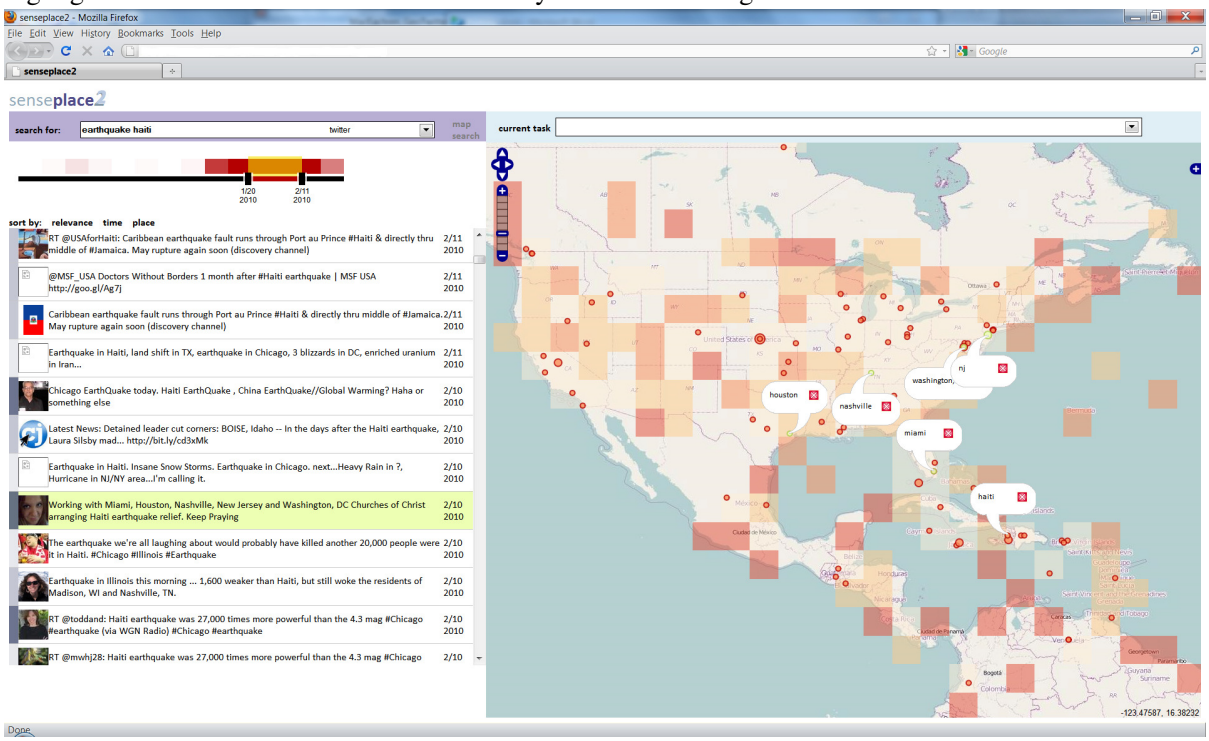
Time-plot/control: The time-plot doubles as a compact representation of the frequency distribution of tweets that match the query across the full time span of data in the database and a control to filter tweets by time. Specifically, both ends of the time range selector can be dragged and once a time range is set, that range can be shifted along the timeline by clicking on the time snap bar and dragging.

Tweet list: The 500 most relevant tweets are displayed in a scrolling list. The list can be sorted by relevance (the default), time, and place. Hierarchical sorting is also enabled (e.g., with time as the primary sort and relevance as the secondary to highlight recent and relevant tweets).

Task list: The task list view (not yet fully implemented) will allow users to label results of a query and store that result in a history. These stored queries will retain user set parameters that include any place and time filtering as well as decisions to promote at tweet to high relevance or hide an irrelevant tweet.

As noted above, the display views are dynamically linked. Clicking on a tweet location on the map moves the tweet to the top of the tweet list, highlights that tweet in the list, and highlights its time bin on the time-plot. If the location has multiple tweets, the sorting and highlighting is applied to the full set. When a tweet is selected in the tweet list, the map zooms and pans to bring it into focus.

The screen capture below illustrates the result of a short analysis session to explore flooding incidents. The session began with query on "earthquake Haiti". The tweet list was sorted by time to find those near the end of the time range of interest (approximately one month after the earthquake during the recovery phase). Interesting tweets were explored, with the map panning and zooming to include all places that a highlighted tweets was associated with (based on the computational processing outlined above). Places were also explored by pointing to them on the map; this moved the tweets linked to the place to the top of the list for easy reading. Thre role of various organizations in the relief effort starts to become apparent as does the variety of places in the U.S. that are active in relief efforts. In the view below, one tweet is highlighted that focuses on multi-state efforts by one church to organize relief efforts.



DISCUSSION AND FUTURE RESEARCH

SensePlace2 includes basic place-time-concept query capabilities along with visual interface tools that support exploration of query results. Our next steps will focus on three goals: add computational methods that support more complex, multipart queries and generate more reliable relevance ranking results, enhance the map and other visual interface components, and add support for a more complete analytical process. Each is outlined below briefly.

Computational extensions: A first step here will be to implement methods that cluster tweets and identify representative examples to stand for the cluster. We also plan to refine our geographic entity extraction methods to take the recent Tweet history of the individual posting the tweet into account as input to disambiguation of places mentioned. Further, we will implement methods to incorporate user input (e.g., indications that a tweet is not relevant) into automated revision of machine learning methods that determine tweet relevance. In addition, we will explore the utility of location-specific tweets (still the minority) to geographically contextualize related general tweets.

Visual interface enhancements: Key features to implement next in the visual interface are bounding box selection on the map to support geographic filtering, multi-scale gridded density surfaces (that allow users to drill-down until a cell has few enough tweets to make display of all relevant tweets in the cell practical), the ability to easily follow URLs contained in tweets, and text visualization tools to provide visual overviews of the concepts and concept structures extracted from tweets. We also plan cartographic extensions to better summarize and represent the spatial components of qualitative information extracted from tweets.

Supporting a sensemaking process: A primary focus here will be on developing and implementing strategies to support extended analytical sessions. Among the strategies planned are: support for user marking of tweets as particularly relevant or irrelevant, history mechanisms that enable users to back-track in their analysis and branch out in different directions, tools to merge results of one analysis stream with others. To support analysis of geographic components of evolving crisis situations, we will develop methods to explore relationships of tweets-from and tweets-about places.

Twitter and other social media are potential sources of information that can be analyzed to support a wide array of monitoring and decision-making activities. The crisis management community has recognized the potential value of social media as both a tool to monitor rapidly evolving situations and to disseminate information. So far, little has been done to expand the notion of geographical analysis of social media reports beyond simply placing geo-located reports on maps. Our work moves beyond self-reported location and identifies locations in the content itself to create space-time representations. The research reported here is an initial step in a longer term project to achieve the potential afforded by emerging social media information sources that refer to locations in a wide range of structured and unstructured formats, with a particular focus on support of place-based information foraging and sensemaking.

ACKNOWLEDGEMENT

This material is based upon work supported by the U.S. Department of Homeland Security under Award #: 2009-ST-061-CI0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security

REFERENCES

Field, K. and O'Brien, J. 2010: Cartoblography: Experiments in Using and Organising the Spatial Context of Micro-blogging. Transactions in GIS 14, 5-23.

Mendoza, M., Poblete, B. and Castillo, C. 2010: Twitter Under Crisis: Can we trust what we RT? , 1st Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA.

Palen, L., Anderson, K., Mark, G., Martin, J., Sicker, D., Palmer, M. and Grunwald, D. 2010: A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. British Computer Society, 1-12.

Sakaki, T., Okazaki, M. and Matsuo, Y. 2010: Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web: ACM, 851-860.

Starbird, K. and Palen, L. 2010: Tweak the Tweet: Leveraging Microblogging Proliferation with a Prescriptive Grammar to Support Citizen Reporting. Proceedings of the 7th International ISCRAM Conference (Short Paper), Seattle, WA.

Starbird, K. and Palen, L. 2011: Voluntweeters:" Self-Organizing by Digital Volunteers in Times of Crisis. Conference on Computer Human Interaction (CHI 2011), Vancouver, BC, Canada: ACM.

Tomaszewski, B. 2008: Producing Geo-historical Context from Implicit Sources: A Geovisual Analytics Approach. Cartographic Journal 45, 165-181.

Tomaszewski, B., Blanford, J., Ross, K., Pezanowski, S. and MacEachren, A. in press: Supporting Rapid Sense Making in Diverse Web Document Foraging Computers, Environment and Urban Systems.

Tomaszewski, B. and MacEachren, A.M. 2010: Geo-Historical Context Support for Information Foraging and Sensemaking: Conceptual Model, Implementation, and Assessment. IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2010), Salt Lake City, Utah, USA, 139-146.

Vieweg, S., Hughes, A., Starbird, K. and Palen, L. 2010: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Proceedings of the 28th international conference on Human factors in computing systems: ACM, 1079-1088.

Zook, M., Graham, M., Shelton, T. and Gorman, S. 2010: Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. World Medical & Health Policy 2, Article 2.