Real-Time Crisis Mapping using Language Distribution

Justin Sampson*, Fred Morstatter*, Reza Zafarani †, and Huan Liu*
*Computer Science and Engineering, Arizona State University, Tempe, Arizona
†Department of EECS, Syracuse University, Syracuse, New York
*{justin.sampson, fred.morstatter, huan.liu}@asu.edu, †reza@zafarani.net

***Abstract*—With the increase in GPS-enabled devices, social media sites, such as Twitter, are quickly becoming a prime outlet for timely geo-spatial data. Such data can be leveraged to aid in emergency response planning and recovery operations. Unfortunately, the information overload poses significant difficulty to the quick discovery and identification of emergency situation areas. The system tackles this challenge by providing real-time mapping of influence areas based on automatic analysis of the flow of discussion using language distributions. The workflow is then further enhanced through the addition of keyword surprise mapping which projects the general divergence map onto specific task-level keywords for precise and focused response.**

***Keywords*-Information Overload; Crisis Mapping; Data Visualization; Social Media**

I. Introduction

The use of sensors to detect, plan, and respond to emergency and disaster situations has grown dramatically in the past decade. Such sensors allow us to track earthquake activity, predict the extent of hurricane-based flooding, and save lives by enabling swift and accurate situational awareness. Recently, the idea of using a network of individuals as "social sensors" has gained significant traction and shown promise. With the advent of social media, in-particularly micro-blogging sites such as Twitter, the volume and range of data has skyrocketed. Additionally, the culture of social media promotes hasty dissemination of information. This was illustrated by Sakaki et al. where Twitter generated data was detected and used as an early warning system that was sometimes able to alert users to an incoming earthquake faster than time it takes for the earthquake to travel at their average movement rate of 3-7 km/s [1]. In some areas it has become common to post new emergency information to social media sites even before contacting emergency response agencies.

Many characteristics of social sensors make them extremely favorable in disaster response, however, there are also difficulties involved in harnessing this data source. Firstly, social media data is undoubtedly "big data". Twitter alone generates over five hundred million tweets per day[1]. First response coordination efforts do not have the time or ability to search through individual messages resulting in what is now referred to as "information overload". Therefore, we need new methods to find pertinent, accurate, and actionable information within the flood of data. Secondly, while much can be accomplished with social media data, it is imperative that first response resources must be allocated in such a way that is not wasteful. Unfortunately, it is difficult to gauge the authoritativeness and accuracy of individual sensors. In order to avoid these pitfalls, we propose a solution that leverages surprising changes in the language probability distributions with the intuition that a shift in the normal language distribution of an area more strongly indicates unusually occurring events. This is in contrast to standard analysis which leverages keyword occurrence rates.

[1]https://about.twitter.com/company

II. Divergence System Architecture

The process of transforming streamed geospatially tagged social media data into a useful and easily manageable representation follows a number of steps. First, the sparseness inherent to GPS-generated location data needs to be resolved through the use of geo-coded regions, or bins. Second, statistical analysis of each geo-coded region using a combination of bootstrapping [2] and Jensen-Shannon divergence[3] allows us to create a macro level representation of the tracked area to further hone the information retrieval process. Third, a metric of surprise is defined which translates the statistical language representation into a keyword weighting system allowing first response planning at the micro level without prior or complete knowledge of the most important event keywords for tracking.

In order to gather actionable and timely spatio-temporal data, we target the Twitter stream focusing solely on data with GPS device coordinates. While the volume of data available from this source is relatively small when compared to the complete Twitter stream, the resulting nine million tweets per day still represents a significant challenge to data analysis for the purpose of timely emergency response.

Since the system operates on data captured and stored from the Twitter stream, it is capable of rebuilding the start state for any area on the globe from historical data as well as other stored data sources. This process only takes a few minutes and upon completion it can be iteratively updated in real-time from live sources. At that point, all calculations and analysis are performed on-the-fly and response time is

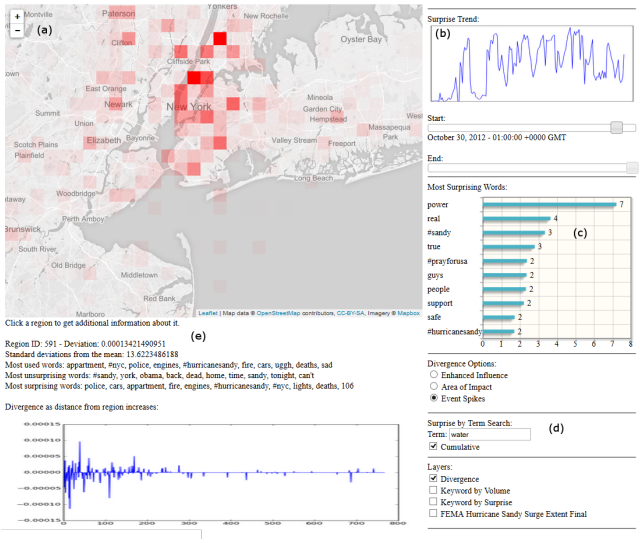

Figure 1. An overview of the system. Seen at (a) the multi-layered geographic visualization can be used to inspect macro-level divergence across a region. At (b) the surprise trend line allows for discovery of segments of time with high cumulative surprise rates. (c) shows the surprising words over the selected period of time as a focus for task-level searching. (d) provides a search term interface to add keywords to be mapped by surprise. (e) is a micro-level display useful for inspecting the cause of divergence in a given region.

subject to volume of data specified by the search parameters. Two tuning parameters are also available: resolution, the total number of regions, and length of time between chunks. The high resolution of GPS-generated coordinates means that users who are making posts from within the same building are likely to generate very different location data. Instead of forcing a specific region size, the system maintains its flexibility through this resolution parameter where the number and size of each geo-coded region is determined from the combination of the desired region to map over and the resolution specified. After generating these regions, each streamed tweet can be quickly assigned to a specific region.

The system workflow acts as a progression between two layers of abstraction. The macro-level divergence layer focuses on the task of identifying regions that possess interesting language distribution characteristics. The micro-level surprise layer targets task-specific keyword mapping to allow for targeted emergency response.

A. Macro-level Divergence Mapping

The highest level view, the divergence layer, leverages the language distribution of the given layer when compared to previous language distributions. It has been shown by Ryoo and Moon that the language of geographical regions can be used to identify the source of text to within 10 kilometers for 60% of users [4]. Intuitively, it is expected that the language representation of a region will shift as topics change. However, these changes tend be small while large language divergence represent surprising, sudden, or extreme fluctuation in the normal language of a region. To compare a given region during a specific period of time we generate a probability distribution for this period, $\mathbf{r}$, as well as an expected distribution generated from previous periods at the same time of day using a bootstrapping [2] method to create an expected distribution, $\mathbf{r}_b$, using fifty randomly chosen samples. The divergence between the two distributions is then calculated using Jensen-Shannon divergence [3]. These divergence values are normalized against the size of the vocabulary. This reduces noise caused in extremely small population areas where the small volume of data would result in large divergence values regardless of actual events. The resulting equation is as follows:

$$\mathbf{r} \in R, NormDiv(\mathbf{r}) = \frac{|V_\mathbf{r}|}{|V_R|}\left(H(\frac{1}{2}(\mathbf{r}+\mathbf{r}_b)) - \frac{1}{2}(H(\mathbf{r})+H(\mathbf{r}_b))\right),$$

where $R \in \mathbb{R}^{kxl}$ is the geo-coded matrix containing the language probability distribution for each region $\mathbf{r}$ with coordinates k and l, V is the vocabulary. H is the Shannon entropy [3] using the distribution from the test period of the region as $\mathbf{r}$ and the expected distribution created during the bootstrapping step as $\mathbf{r}_b$.

B. Micro-level Surprise and Volume Mapping

Though the high level view allows emergency response agencies to discover points of interest, a task which divergence is ideally suited for, an additional metric is necessary in order perform queries at the micro level. Since the divergence level for a specific region is related directly to the change between the distributions, directly examining the probability distributions can give additional insights into the reason for the divergence. To determine the difference between a normal distribution we take the average of previous distributions for the region as a vector and subtract out the distribution vector of the desired test period to create a new vector. In this way, we can understand the impact of the difference between each word probability. Under this scheme, words that are used commonly within a language distribution should appear at roughly the same rate between each distribution and contribute little to the discovery of interesting keywords. After taking the difference between each distribution, the common words with these characteristics should become small values around zero which is ideal. Words that normally have a high probability in previous distributions but fail to be represented in the test set represent a shift in communication topic that also provides little information as to the current situation of an area. This type of word will continue to be hold the largest positive values. The last, and most useful, type of words are those that were not previously represented but show a sudden and large increase in probability. Since these words have not previously been well represented, they are the most

informative source of information to the current situation of an area. As a result, this type of word will take the largest negative values.

The use of surprise, which we define as the range of values discovered in the above process, as a metric has several advantage. First, further mitigation of information overload can be achieved by calculating the cumulative surprise over a period of time in order to gain a high-level view of the keywords that explain spikes in the divergence. The combination of the divergence map along with the surprise metric are effective for quick identification of regions and keywords of interest. Second, once keywords have been identified, surprise in each geo-coded region can be mapped based on individual keywords in order to further "drill down" to the micro-level impact details of each keyword over an event period. In this way, the first response planning flow becomes simplified and streamlined as automatic discovery of important keywords is accomplished entirely by the system. Third, crowd wisdom is leveraged inherently by the surprise metric as, by definition, the most surprising words will be those that show large and constant growth.

Volume mapping, where the keywords are mapped to their corresponding regions based on keyword frequency instead of surprise, is included for the purpose of direct comparison to standard frequency-based filtering methods. Surprise, as a function of language distribution probability, remains linked to the volume of a given keyword within the geo-coded region with some important differences. Volume mapped keywords suffer from population bias that can be difficult to account for while the surprise metric is inherently normalized to the language distribution of a given geo-coded region. Additionally, frequently occurring words that do not provide significant information, such as references to city names or holidays, dominate frequency-based methods. This forces additional time and effort to be expended to separate useful keywords from those that are simply common.

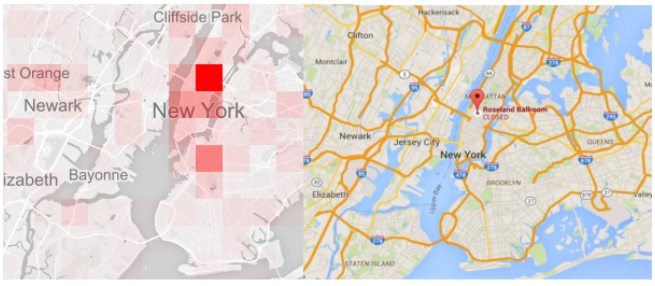

Figure 2. A small event discovered by the divergence map. On the left, the divergence map along with the surprising keywords that explain the map are shown. On the right is the actual location of the event within the region according to Google Maps.

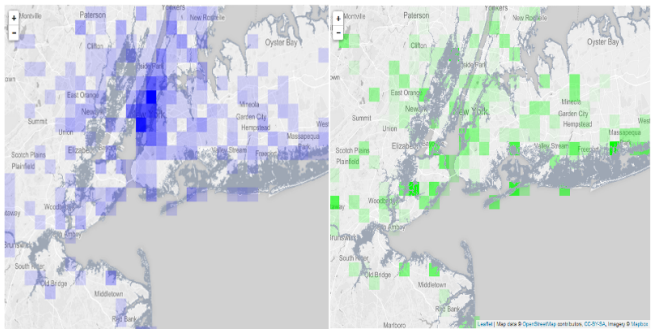

Figure 3. Using the geo-coding system, the impact of the keyword "water" can be mapped based on frequency of use as well as surprise. The colored mappings are displayed over the FEMA Final Surge Impact Assessment report which can be seen in gray. On the left in blue, the frequency map indicates regions with a high volume of conversation containing the specified keyword while on surprise-based map in green can be seen on the right.

III. Case Study: Hurricane Sandy

Though the system is designed to operate on streamed data, we focus on a test dataset containing 14,528,732 tweets with GPS device coordinates originating anywhere in the world. Each tweet is introduced in the order that they were received from the stream and encompass dates between October 18, 2012 at 21:00 GMT and October 31, 2012 at 23:00 GMT.

Initial landfall of Hurricane Sandy was made on October 29th at 18:00 GMT. Additional data from the stream before the landfall event is used by the system to create a baseline as well as to provide qualitative comparison between system results over the period of standard conversation preceding the disaster event with the conversation during and after. It is important to note that since the system gathers and stores all geographically significant tweets, it is not necessary to have foreknowledge of the event to be tracked in order to create the initial divergence baselines as they can be created and analyzed retroactively.

Beginning the initial analysis workflow of the Hurricane Sandy area surrounding Manhatten, New York, an immediate overview of the can be found in the Surprise Trend graph. Several hills are distinguishable that indicate shifts in the topic flow of the area. Taking a closer look at the surprise at the macro level is accomplished by selecting appropriate start dates on the appropriate sliders. Conversation topics before the hurricane include mundane conversation words such as "college", "hotel", and "visiting", however, initial mentions of the hurricane and the weather grow significantly in the days before landfall. Emerging words and topics at the divergent points display interesting surprising words such as "roseland", "ballroom", and "#gracejones" which refer to the Grace Jones show at the Roseland Ballroom on October 27th, 2012. This event is mapped within minutes of the starting time of the concert as well as the correct location

as shown in Figure 2. Other interesting events discovered in this way including a football game held at Kenneth P. Lavalle Stadium shortly before landfall. Many of these event locations become evident based on their divergence within as few as thirty tweets.

When selecting regions of time after landfall the top words become increasingly weighted towards the disaster event. Standard conversation terms drop out of the overall view and are replaced by such as "storm", "power", "hurricane", "support", and "help". These terms become useful as they highlight searches to be targeted for specific forms of aid. For example, shortly after the storm hit, the Con Edison power plant suffered an explosion that resulted in large blackout regions. Within minutes the divergence map discovers the event, and using the keywords "power", "explosion", and "blackout" as suggested by the system a map is generated of the surprise area for immediate aid. While this is a small event, the same work flow can be applied to flooding as a result of the hurricane. Figure 3 illustrates this by showing the volume-based mapping and surprise-based mapping of the discovered keyword "water" over the course of the heaviest impact of the hurricane on top of the Storm Surge Impact Analysis released by FEMA[2] on February 14th, 2013, several months later. As can be seen, the volume-based method displays population bias as well as a tendency to over-predict impact regions while the surprise-based method provides another view leveraging the difference in the language.

IV. Related Work

The use of social sensors as a supplemental data source for crisis detection and response can be useful when standard data gathering methods are not available. Schnebele et al. used social media data to fill the data gaps during the 2013 Calgary flooding when satellite imaging was obstructed by cloud cover [5]. This method performs well but relies on targeted keywords which may not necessarily be well known at the time of the event. Middleton et al. approached the problem by gaining additional information from gazetteers to target at risk areas [6]. They succeeded in showing that social media data performs well in comparison to "gold-standard post-event impact assessment".

In order to use divergence as a measure of topic shift, the standard language of a given geographic location should be consistent. Han et al. show that is possible to infer the location of a non-geographically tagged tweet based on models built from exclusively geographically tagged tweets. Additionally, they discovered that general language models that include multiple languages are increasingly effective as the inclusion of extra languages into the model provide additional discriminating factors [7]. Morstatter et al. approached the geo-location task by using language features to determine if a tweet originates from within a region of crisis [8]. The consistent success of language-based location inference provides us with the groundwork required to apply our method.

[2]http://arcg.is/McDSPO

V. Conclusion

The dual system architecture employed gives emergency response planners a new and effective tool to combat information overload and extract useful geo-spatial information from social media. At the highest level, divergence serves as a general filter providing detailed information on suddenly emerging topics as well as their area of impact. Additional low level inspection and planning tasks can then make use of the surprise-based search functionality to target specific types of response based on discovered situational needs. The system provides a high level of flexibility and is able to detect events ranging from local concerts and sports gatherings to state-wide flooding and power outages. This provides a highly streamlined process of information discovery and extraction which aims to shift the workload from the shoulders of the emergency responder to that of a dynamic intelligent system.

VI. Acknowledgments

This work is sponsored, in part, by Office of Naval Research grant N000141410095.

References

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 851–860.

[2] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, 1982, vol. 38.

[3] J. Lin, "Divergence Measures Based on the Shannon Entropy," *Information Theory, IEEE Transactions on*, vol. 37, no. 1, pp. 145–151, 1991.

[4] K. Ryoo and S. Moon, "Inferring Twitter User Locations with 10km Accuracy," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. International World Wide Web Conferences Steering Committee, 2014, pp. 643–648.

[5] E. Schnebele, G. Cervone, S. Kumar, and N. Waters, "Real Time Estimation of the Calgary Floods using Limited Remote Sensing Data," *Water*, vol. 6, no. 2, pp. 381–398, 2014.

[6] S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time Crisis Mapping of Natural Disasters using Social Media," *Intelligent Systems, IEEE*, vol. 29, no. 2, pp. 9–17, 2014.

[7] B. Han, P. Cook, and T. Baldwin, "Text-based Twitter User Geolocation Prediction," *Journal of Artificial Intelligence Research*, pp. 451–500, 2014.

[8] F. Morstatter, N. Lubold, H. Pon-Barry, J. Pfeffer, and H. Liu, "Finding Eyewitness Tweets During Crises," in *ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014.