A Semi-automated Display for Geotagged Text

Vincent A. Schmidt
Air Force Research Laboratory
Dayton, Ohio USA

Jane M. Binner
Sheffield Management School
University of Sheffield, UK

Abstract

The changing dynamic of crisis management suggests that we should be leveraging social media and accessible geotagged text data to assist with making emergency evacuations more effective and increasing the efficiency of emergency first responders. This paper presents a preliminary visualization tool for automatically clustering geotagged text data, and visualizing such data contextually, graphically, and geographically. Such a tool could be used to allow emergency management personnel to quickly assess the scope and location of a current crisis, and to quickly summarize the state of affairs. Discussion herein includes details about the clustering algorithm, design and implementation of the visualization, and ideas for improving the utility for use in a variety of circumstances. [1]

Keywords: visualization, geotagged text, social media

1 Introduction

The dynamics of crises are changing. In particular new communications and social networking technologies mean that unprecedented opportunities for real-time communication in evacuation at street level are emerging. Public information for warning of impending crises need no longer be transmitted in a 'top-down' fashion and rapidly we have moved to a situation where public information is just one signal in an information market place. Victims and evacuees can now communicate with each other before, during and after an emergency and gain real-time access to information. As a result, they may indulge in individual and collective strategic behavior (by generating alternative information and rumors perhaps). Such strategic decisions by evacuees can be analyzed to design safer and faster evacuations. In order to do so public information and planning needs to become interactive, dynamic and responsive. For example, it is plausible for emergency management agencies to collect information during a crisis (e.g. messages posted on social networking sites) and to use this, not only in managing evacuations, but also to intervene by posting messages, selectively targeting trusted sources and modifying electronic signage. This would modify incentive structures in evacuation through acting on real-time information.

One way to promote this capability is to design visualizations for emergency management personnel that are capable of displaying relevant data about the contents and metadata within real-time social media messages. Such systems should summarize salient semantic points in such a way that first responders could re-inject recommendations back to social media and emergency management communications systems in near real time.

The uses for such visualization reach beyond field use by first-responders, though. Similar systems could be used to monitor general community "health," acting as an additional input into crisis prediction and recognition scenarios. Social scientists and analysts could also use such systems to investigate causal activities and understand their impacts.

This paper provides the basis of a preliminary study to examine the utility of visualizing geotagged text data, similar to that available through social media, as a tool for analysts, policy makers, first responders and crisis management personnel, and social scientists. After summarizing certain background information, we describe the contents of a typical dataset and discuss how it could be prepared and demonstrate how it is used in straightforward geospatial visualizations. We conclude with comments about the types of visual-

[1] Paper cleared for public release.

ization we believe might be most appropriate, and indicate interesting areas for future work.

2 Background

A number of synoptic and small-scale studies have already noted the use of individual electronic communication including the Californian wild fires of 2007 [1] and the 2007 Sheffield floods [2]. The role of agents and computational models in crisis management has been considered by Chen and Xiao [3] who consider that real-time information can give feedback resulting in the adjustment of plans by the emergency services. Innovatively, Nakajima et al [4] have considered the use of ubiquitous devices such as GPS and mobile phones to build a multi-agent evacuation strategy for the city of Kyoto, whilst Ushahidi [5] maintains up to date reports submitted by the public and makes available latest incident reports on their website to assist victims of the Haiti earthquake. There is, of course, a large body of existing work on mathematical modeling of evacuations [6–8], but not much work has been published on the role of feedback loops and social networks in evacuation. Our project is one of the first to systematically combine emergency planning and visualization simulation of crisis behavior taking into account exchange of information through social networking, and considering the resulting aspects of strategic decision making.

Given the difficulty of conducting real-world experiments of crisis behavior it is hard to make valid inferences as to the effects of social network technologies without advanced computational modeling and systematic validation and testing. Similarly without computational modeling it is difficult to identify the ways in which responders may intervene in such networks. Furthermore, there is a need to involve policy makers and responders in the validation of these models. Governmental organizations are already interested in social networking and communication technologies for resilience. For example, Twitter has been noted in earthquake prediction (http://twitter.com/quakeprediction), is used by the Los Angeles Fire Department (http://www.govtech.com/gt/579338) and the Cabinet Office is considering how mobile technologies could be used in warning and informing the general public.

An impressive demonstration of the importance of ideas of collective behavior in connection with the use of social networking technology, is the recent 'balloon hunt' by the Defense Advanced Research Projects Agency (DARPA), a Pentagon agency in the US (see [9] for an account in the recent press). Further details of new research into developing tools to model collective behaviour from the theory of complex adaptive networks involving the inoculation of networks with information, and advance agent-based models of emergency planning allowing for emergent communication channels are available in [10]. This project will address these new challenges through a systematic computational modeling approach.

We desire to analyze the behavior of populations in a crisis and evacuation, focusing on the effect of receiving, spreading and acting on information on the behavior of agents/evacuees. Based on these data visualizations an understanding will be developed of how the behavior of agents/evacuees can be modified and controlled through the use of real time intervention in social networking and communications technologies. The main outputs of the project will be the speedier identification of intervention strategies for more effective crisis management, thereby making evacuations safer and faster. Ultimately, we will make recommendations to stakeholders as to the efficiency of different communication channels and control strategies arising from our simulations. This will provide a sound basis for policy makers and responders to strategize about intervention in communication and social networking technologies.

3 Dataset Preparation

The dataset used in this preliminary concept is drawn from the 2011 VAST Challenge, related to the 2011 IEEE Conference on Visual Analytics Science and Technology (IEEE VAST). The mini-challenge 1 dataset (Geospatial and Microblogging — Characterization of an Epidemic Spread) is interesting to us because the data consists of a million uniquely identified records containing originator id, timestamp, geospatial information, and message text. This information is typical of social networking data obtainable from SMS text messages, Twitter, and other commonly used sources.

The VAST (mini-challenge 1) problem is to use the text records, coupled with information about a fictional city (its population, industry, hospitals, weather, and transportation modes), to determine the source of an epidemic disease and how it is spread throughout the community.

Instead of using the data to solve the VAST challenge, our immediate interest is examining how this type of information can be effectively displayed to a specific end-user or analyst. To that end, 10000 records were randomly selected from the original dataset, provided as a comma-separated value (.CSV) file, for evaluation. Figure 1 shows the header line of our resulting data file, and includes a small representative selection of records. ID is a unique identifier for a specific user, Created_at is the record timestamp, Location is a Lat/Lon pair indicating the geographic source of the message, and the body of the message is last.

Casual examination of this small set of examples promotes the correct conclusion that the messages included in the dataset do not all relate directly to the question posed by the challenge problem. Of course, this is typical of such data. It is also easy to see that the messages may contain numerous typographical and punctuation irregularities, acronyms, and various shorthand symbology. Note that the record content is not limited to the English language (and for VAST, there are many foreign language messages in the corpus).

There are several geotagged datasets available for research use, and they are generally fairly large. In reality, this type of data is practically infinite in size, since it can often be streamed in real time directly from its source. Trying to graph or display all of this data simultaneously is obviously burdensome to the system and overwhelming to the user. Therefore, the display must be designed to support the user's work needs.

Visualizing a large quantity of messages effectively is best accomplished by dividing the dataset into smaller and manageable message groups. We accomplish this by sorting the records by timestamp, then grouping the results according to messages that are temporally "close" to one-another. We believe that "conversations" can be found by looking for messages that happen in close temporal succession, and these conversations are a reasonable way to begin to select message subsets. (Admittedly, there are many other ways to group this data.) Groups of messages in a conversation can be further subdivided geographically, yielding a finer level of detail for the analyst.

The method used to divide, then subdivide, the collection of messages is based heavily on the automated clustering algorithm developed by Schmidt [11] for preprocessing neural network datasets. It is important to note that the automated algorithm does not require any user intervention to generate reasonable clusters, and is not based on fixed-length or fixed-time strategies for dividing the data. This is best shown by example.

Figure 2 depicts the basic technique using 50 randomly generated uniform data points in (0,1000). Figure 2(A) shows the original data. The algorithm sorts the data (2(B)), then finds the difference between consecutive data points. The differences for this dataset are shown in 2(C). The algorithm slices the data at the positions where the difference is greater than twice the standard deviation of the differenced data. (The standard deviation is shown as a horizontal red line in this Figure). The observant reader will notice the "steps" in the sorted data of Figure 2(B) are located directly above the most prominent differences indicated in Figure 2(C). These are the same places the data cluster boundaries are created.

Clustering our geospatial data is a multistep process. The first step is to find clusters based on the distance from some arbitrary point. Our system currently selects the center of the map, and messages are sorted and clustered based on their distance from the center. Then, each cluster is examined and, possibly, subdivided again, depending on the (angular) location of the messages in relation to the center of the given cluster. The same algorithm is reused for this subdivision, but the "distance" measure is the relative angles separating message locations.

Although this approach yields generally useful geospatially clustered data, it is unlikely that using geospatial indicators is the best approach to clustering cyber data such as social network messages, tweets, SMS text messages, and similar types of information. To address this shortcoming, we are working on adding semantic clustering algorithms as an addition to (or alternative for) the existing technique.

```
ID,Created_at,Location,text

3,5/18/2011 13:26,42.22717 93.33772,

      this convention filled with technology could be better don't plan on leaving
      anytime soon

57,5/12/2011 21:08,42.23363 93.34164,

      There's no point of trying if no one else i s...

73,5/16/2011 5:28,42.28818 93.33605,

      sick of hearing bout the Oil Spill now like

127,5/10/2011 16:53,42.28051 93.34164,

      I hope #inception is as good as everyone s ays  ten bux is ten bux...

238,5/9/2011 14:13,42.21771 93.33845,

      ...I agree  the most demanding task in our  time is come to terms with space;
      so let time be our personal unit of confusion. ..

464,5/4/2011 20:04,42.28523 93.44908,

      Isnt it weird that after all the hoopla abo ut the TSA there was another
      'near disaster' averted? Seems like it happens too  much.

592,5/17/2011 3:27,42.28818 93.28493,

      Being sick is exhausting or I'm exhausted b /c I'm sick. Either way
      a bed would be nice right now.
```

Figure 1: Selected Dataset Records

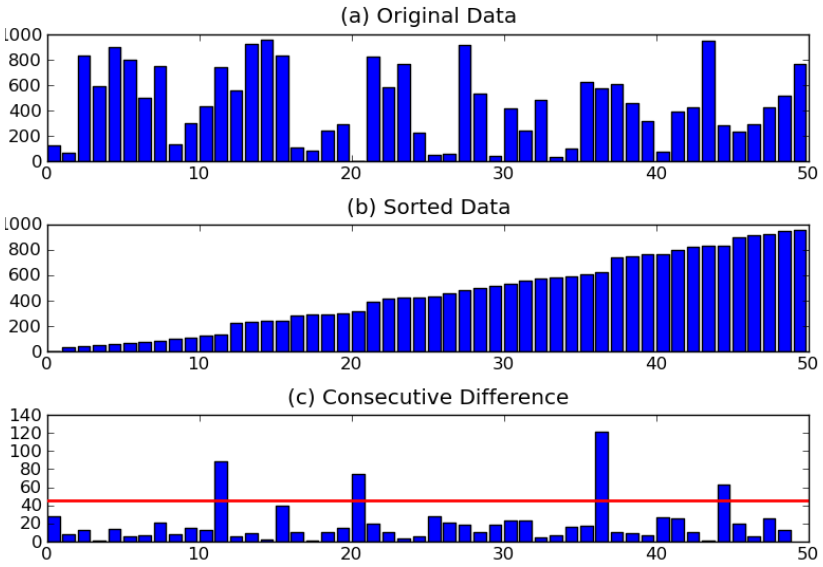

Figure 2: Automated Clustering Example

4 Visualizing the Data

Our primary purpose for visualizing geotagged textual data is to provide first responders and other analysts a mechanism for quickly identifying and responding to trends in social media data that indicate the status of a catastrophe or crisis. When the visualized data has no real form, and looks like "noise" from a variety of sources, it is most likely that no crisis is underway, so no immediate action is required. If, however, the data tends to converge to include specific topics of interest, or many messages are suddenly from (or about) a certain geographic location, a catastrophe or crisis requiring specific action may be under way. Good visualization of such data is expected to enhance an analyst's ability to quickly offer relevant guidance to the appropriate authorities and catastrophe response personnel.

The visualization we are currently considering explicitly divides the display into three sections: message relationships, geospatial information, and message clusters. Figure 3, derived from the VAST data, is typical of this display.

The top portion of the figure represents the message relationships. It is a graph, partitioned into (automatically) geographically clustered sets of messages. Clusters are labeled alphabetically, starting with the letter "A," and messages within clusters are numbered from 0. In this figure, message nodes are linked in monotonically increasing time, such that A0 is timestamped before A1, which is timestamped before A2, etc. This arrangement of links is certainly not ideal, however. It would probably be more valuable to link the message nodes based on the semantic similarities of the contents of the message bodies. Work to revise the message graphs in this manner is in progress.

The center of the figure contains a map of the geographic area of interest. A line connects the message graph to the mean location of all messages within that cluster. The actual position each message is indicated by the corresponding message label on the map, and the mean location of the message cluster is indicated by the placement of the cluster label on the map. We can optionally draw a line from each message node on the graph to its location on the map, but we found this additional information clutters the display and makes it tedious to read.

On the bottom of the figure is a histogram representing (on the X axis) the number of message clusters (actually 406 in this case, not explicitly indicated on the display), and (on the Y axis) the number of messages in each cluster. The red mark on the histogram indicates the message cluster currently being graphed and shown on the map.

The user can change the currently displayed cluster by clicking the mouse on the histogram, or by using the cursor keys to move left or right by a single cluster. Hovering over a graph node at the top of the figure shows a tooltip containing the text of the selected message.

Other visualization are being added to the display as the semantic analysis and related development continues.

5 Conclusions & Future Work

There are many ways to display geotagged text data. However, the effectiveness of the visualizations depends heavily on the objectives of the end-users. Users with different work requirements will often need different types of interfaces, even if they use identical data sources.

The visualizations we demonstrate in this effort are entirely exploratory in nature. We perceive a variety of uses for this basic type of interface:

- *Fully automated operation* might watch one or more social media streams in real time, indicating an alarm condition if a certain subset of words becomes frequent, or if a particular location is referenced in a certain way. This would require the addition of flexible filters to the existing design. Such a mode would be valuable to first responders as an additional watchdog for catastrophic events.

- *Query design* would allow a user to type a textual query, and the message graphs would be reorganized depending on the semantic contents of the query. A more structured interface would have to be added to make this utility viable. This mode promotes a more visual geospatial search operation.

- *Exploratory visualization* occurs when the records are displayed using self-organizing algorithms. Deciding the "interestingness" of data is the grail of data mining, but a growing

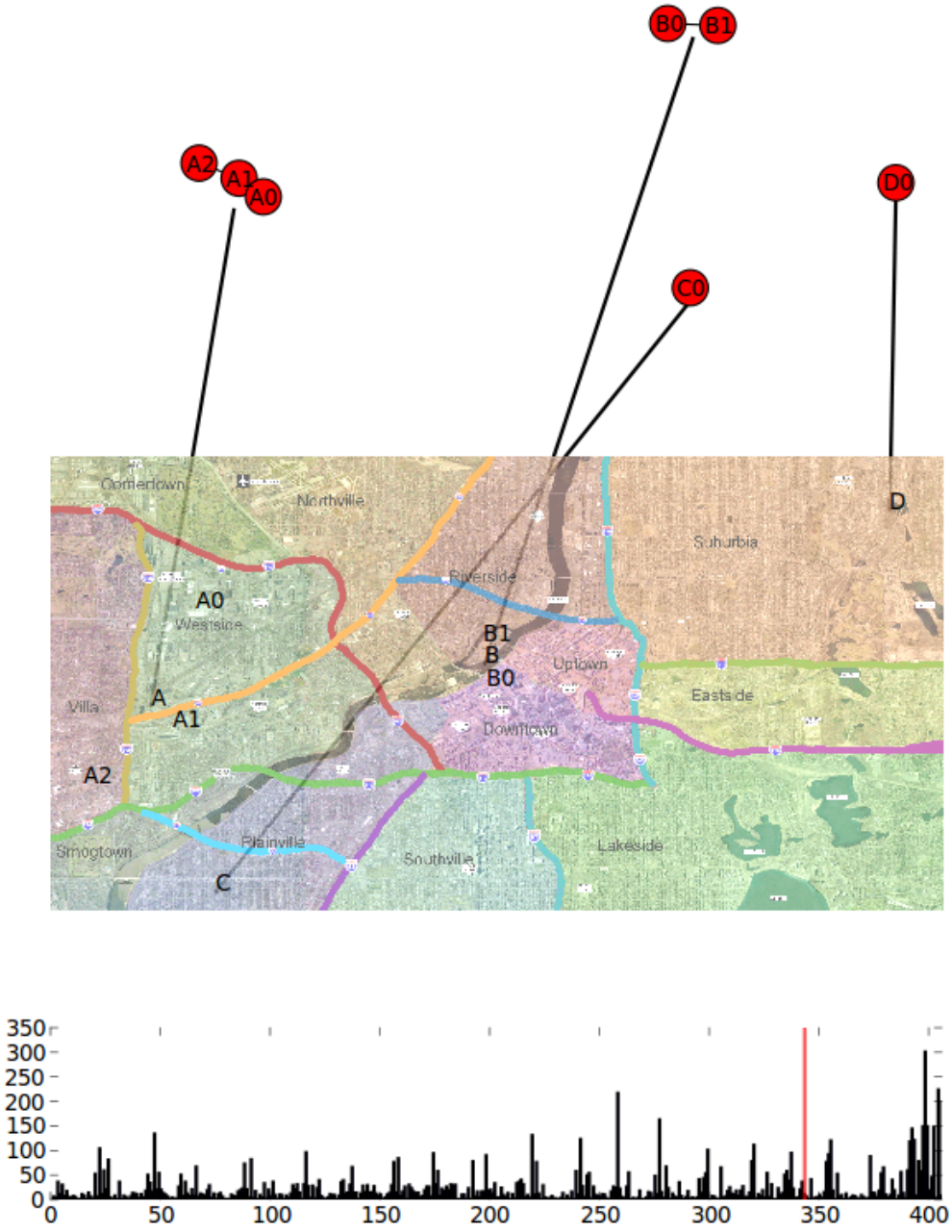

Figure 3: Visualizing Geotagged Data

collection of algorithms and visual interfaces allow analysts and scientists to leverage these systems as tools more easily, especially when the automation can be trusted to identify and display certain trends without explicit interaction. For now, the utility we demonstrate requires direct control by a user.

Our short-term plans include the incorporation of a variety of semantic algorithms, permitting the message graphs to be connected in more meaningful ways. This enhancement would also allow the graph clusters to be tagged to support semantic search operations.

Adding a search tool or a watch-list of interesting terms would enable the utility to be used to display the results of simple searches. One simple extension to this concept is to collect geotagged results from commercial search engines, then use

the visualizations we describe here to display those results.

Although our initial implementation is designed for traditional computer screens, we have access to a sizable collection of advanced 3-d technologies. It would be interesting to revise the visualizations such that these technologies could be used when available. It would also be useful to experiment with adding these visualization concepts to small displays, such as cell phones and the latest generation of tablet computers.

Such a move is in keeping with other similar developments such as the new national alert system which is set to begin in New York City to alert the public to emergencies via cell phones. This new Personal Localized Alert Network (PLAN) will enable presidential and local emergency messages as well as Amber Alerts to appear on cell phones equipped with special chips and software. The Federal Communications Commission and the Federal Emergency Management Agency confirm that the system will also warn about terrorist attacks and natural disasters.

There is clearly much work to be done. The goal is far more important than the mere display of message data on a graph or map. The ultimate objective is to create a reliable tool that allows first responders and others to leverage social media to protect the public at large. The testimony of the value of such a tool occurs when those who use the prototypes designed for their work areas are able to claim these devices and visualization are directly responsible for lives being saved.

6 Acknowledgements

The authors would like to acknowledge the generous support of EPSRC Grant reference number EP/I005765/1 for the funding provided and to the members of the EPSRC Dfuse team, particularly John Preston, University of East London UK and Maria Angela Ferrario, University of Lancaster for research support.

References

[1] J. Sutton, L. Palen, and I. Shklovski. Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires. In *Proceedings of the 5th International ISCRAM Conference*.

[2] V. Lanfranchi and N. Ireson. User requirements for a collective intelligence emergency response system. In *Proceedings of 23rd BCS HCI Group conference (HCI 2009)*.

[3] Chen and Xiao. Real-time traffic management under emergency evacuation based on dynamic traffic assignment. In *Proceedings of the IEEE International Conference on Automation and Logistics (ICAL 2008)*.

[4] Y. Nakajima, S. Yamane, H. Hattori, and T. Ishida. Evacuation guide system based on massively multi-agent systems.

[5] Ushahidi. http://haiti.ushahidi.com/main, 2009.

[6] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 2000.

[7] C. Burstedde, K. Klauck, A. Schadschneider, and J. Zittartz. Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A 295*, 2001.

[8] A. Ferscha and K. Zia. Lifebelt: Crowd evacuation based on vibro-tactile guidance. *IEEE Persvasive Computing Issue*, 2010.

[9] M. Hesse. Mit team wins social networking balloon hunt. *Washington Post*, 2009.

[10] J. Preston, L. Branicki, MA Ferrario, and M Kolokitha. Multiple attacks on transport infrastructure: an inter-disciplinary exploration of the impact of social networking technologies upon real time information sharing, response and recovery. *Journal of Homeland Security (forthcoming)*, 2011.

[11] Vincent A. Schmidt. *An Aggregate Connectionist Approach for Discovering Association Rules*. PhD thesis, Wright State University, Dayton, OH, May 2002.