epSpread - Storyboarding for Visual Analytics

VAST 2011 Mini Challenge 1 Award: Outstanding Analysis Using Custom Tools

Llyr ap Cenydd[*] Rick Walker[†] Serban Pop[‡] Helen Miles[§] Chris Hughes[¶] William Teahan[‖]
Jonathan C. Roberts[**]

School of Computer Science
Bangor University

Abstract

We present epSpread, an analysis and storyboarding tool for geolocated microblogging data. Individual time points and ranges are analysed through queries, heatmaps, word clouds and streamgraphs. The underlying narrative is shown on a storyboard-style timeline for discussion, refinement and presentation. The tool was used to analyse data from the VAST Challenge 2011 Mini-Challenge 1, tracking the spread of an epidemic using microblogging data. In this article we describe how the tool was used to identify the origin and track the spread of the epidemic.

Keywords: Visual Analytics, Coordinated Multiple Views, VAST 2011, information visualization, epidemic visualization

1 Introduction

One of the mini challenges (MC1) for the VAST Challenge 2011 was to track an epidemic over a city from microblogging data. The dataset consisted of short messages together with user id, time sent and user location, and information on the city of Vastopolis: a map, population by district and time of day, weather conditions, locations of hospitals and other significant locations and also observed symptoms of the epidemic. The task was to discover the source of the epidemic and show how it spread across the city, and to suggest possible emergency response options.

Our analysis was supported by tools developed over the competition period. We took two avenues of investigation: distribution (spatial and temporal) and microblogging content. Following these paths led to the development of a number of visualization and analysis tools which were then combined to allow us to ask, and answer, more complex questions in an interactive manner. Finally, our software tool, epSpread, allows us to create storyboards of visualizations to use in explaining hypotheses to other group members. In this article, we discuss our analytic process and give some details of the visualizations we used in arriving at our conclusions.

2 Analytical Process

Dividing our analysis into two broad streams — distribution and content — enabled progress to be made simultaneously on each by splitting the team to explore avenues in each. Some small prototypes were built and brief analyses conducted, with results fed back to the rest of the group through regular meetings and via Google Wave to share data and screenshots.

As our understanding of the scenario grew, so too did the complexity of the questions raised at meetings. This quickly reached the point where it become important to be able to quickly and interactively answer at least simple questions during meetings, which in turn required more serious tool development. Further, the more complex and interesting questions were largely concerned with interaction between streams: for example, to answer the question 'Which people at the stadium later tweeted that they had fever?' requires input from both distribution (spatial location) and content (message details).

Delaying integration of prototypes until required helped focus tool-building effort on areas required for analysis, and this question-led iterative process continued until our event narrative was complete. Further, to aid in explaining hypotheses to the rest of the group, we developed a visual storyboard system that combines visualizations over time and allows quick presentation of the evidence for a given narrative. Our tool was developed in Processing [2] using a number of additional libraries and an SQLite database backend.

3 Visualizations

epSpread uses a variety of visual representations. Here, we discuss our use of stream graphs to show patterns in usage, word clouds to examine content, maps to show distribution, and all three in combination with querying to perform more complex analysis. A screenshot of epSpread is shown in Figure 1.

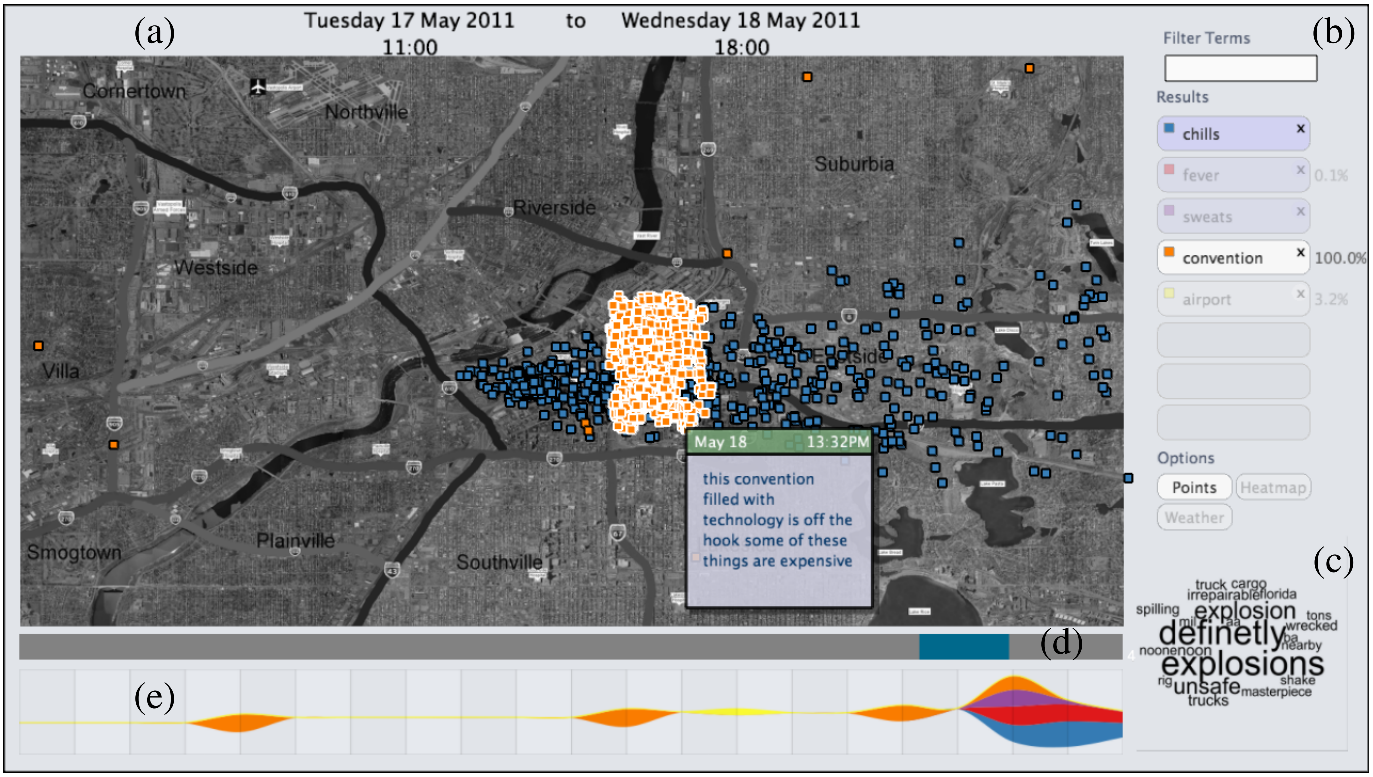
Early questions on the distribution stream were concerned with identifying events through number of messages: can the dates of large events be determined by message counts per day? While the overall trend is largely flat until the last four days, splitting messages by district of origin clearly indicated events such as conventions. Stacked area graphs and, later, streamgraphs [1] were used to show this visually.

We used a relative entropy-based probabilistic measure for ranking word unigrams that have a probability that differs from the probability of the same unigram in a reference corpus (estimating the probability naively using the frequency count for the unigram divided by the number of unigram tokens). Messages sent in the first three days were used as our reference corpus. A word cloud [3] depicts the output from this measure over a selected time period, and allowed us to identify important words (and hence events) over selected time periods.

A simple querying interface allows tweets to be shown directly on the map (and on a streamgraph below). Multiple queries can be overlaid for comparison, mapped to different colours. We also implemented a heatmap display of tweets, to address two problems with directly visualising tweets: overplotting (many tweets in a small region will mean some are not visible) and population density. Put simply, a hundred tweets with fever from a sparsely populated district are far more significant than the same number from a densely populated one. Further, the population of each district varies significantly between working hours and the rest of day. To represent these changes accurately, we weighted tweets in the

[*] e-mail: llyr.ap.cenydd@bangor.ac.uk
[†] e-mail: rick.walker@bangor.ac.uk
[‡] e-mail: serban@bangor.ac.uk
[§] e-mail: helen.c.miles@bangor.ac.uk
[¶] e-mail: c.j.hughes@bangor.ac.uk
[‖] e-mail: w.j.teahan@bangor.ac.uk
[**] e-mail: j.c.roberts@bangor.ac.uk

IEEE Symposium on Visual Analytics Science and Technology
October 23 - 28, Providence, RI, USA
978-1-4673-0014-8/11/$26.00 ©2011 IEEE



Tuesday 17 May 2011 11:00 to Wednesday 18 May 2011 18:00

Filter Terms

Results

chills

fever 0.1%

sweats

convention 100.0%

airport 3.2%

Options

Points Heatmap

Weather

May 18 13:32PM

this convention filled with technology is off the hook some of these things are expensive

Figure 1: Some features of epSpread: (a) geographical view of two sets of tweets (b) querying interface with cross-query results as percentages (c) word cloud for selected time period (d) time range slider (e) streamgraph visualising query results over time. This figure shows the result of a cross-query between people at the convention in Downtown on the 18th May and those who reported suffering from chills, fever or sweats. This query was performed entirely through selection on the available views.

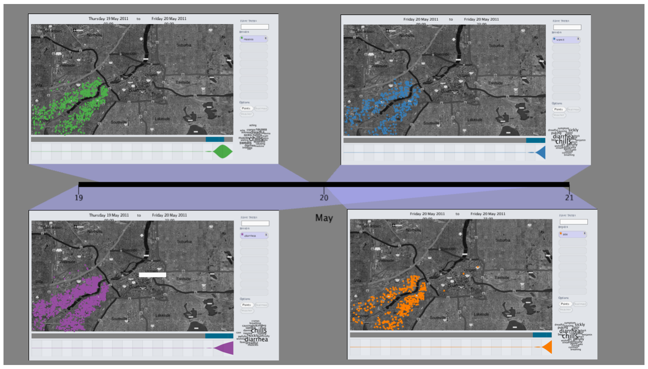


Figure 2: Storyboard for the spread of the epidemic. Spread shown clockwise from top left: nausea, vomiting, diarrhea and abdominal pain. While the spread pattern is the same, the temporal pattern differs and this is shown by linking each story panel to the timeline.

heatmap by population of district at the time the tweet was sent, using a kernel-based smoothing algorithm.

4 Geographic Cross-Querying

Some of the most important queries involve not just content but also location. By allowing cross-filtering between queries, we were able to answer questions such as ‘how many people with fever also had nausea?’ By taking this one step further and allowing filtering by selection on the map, more complex questions such as ‘how many people who went to see the baseball game later tweeted about fever?’ can be answered quickly and easily.

5 Storyboarding

Each complete visualization (map, streamgraph and word cloud) can be resized and shown on a timeline. This means that, rather than try to explain the whole scenario with one overly-cluttered view, many single events (some involving complicated queries) can all be visible (and editable) in context within the tool. This proved particularly useful in narratives where temporal ordering was important, including our final answer. One example of this is shown in Figure 2.

6 Conclusions

Our development methodology on epSpread - multiple prototypes in multiple directions, and integration of the most successful aspects into an interactive tool - played an important part in the quality of our analysis. In particular, supporting complex queries entirely through the interface and the storyboarding approach were very effective in forming our narrative of events, and could be equally effective for other datasets. Further implementation details and code are available at: `http://cvev.bangor.ac.uk/VAST2011/`

References

[1] L. Byron and M. Wattenberg. Stacked graphs–geometry & aesthetics. — *IEEE Transactions on Visualization and Computer Graphics*, pages 1245–1252, 2008.

[2] C. Reas and B. Fry. *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press, September 2007.

[3] F. Viégas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, pages 1137–1144, 2009.