VAST 2011 Mini-Challenge: UNC Charlotte Viral Tracker

Alex Midgett[*] Jackie Guest[†] Kalpathi Subramanian[‡]

Charlotte Visualization Center
Department of Computer Science
The University of NorthCarolina at Charlotte

Abstract

We built *UNCC Viral Tracker*, a visual analytic tool to characterize the spread of an epidemic, as part the VAST 2011 Minichallenge. Our goal with this application was to provide an intuitive interface that keeps relevant data in view while eliminating as much noise as possible. We achieved this using a combination of data preprocessing and customized interactive visual analysis tools. We describe how our tools and methods drove the visual analytic process.

1 Introduction

Given the nature of the tasks in this challenge, we focused the design of our application on providing users with a clear overview first and foremost. All of the provided data is utilized in the display and arranged in an intuitive manner, so that users can quickly determine which factors may be in play and form hypotheses. Data is presented in geographical and temporal format, with the ability to browse data in a familiar style. Additional features allow users to examine data in detail, and quickly utilize what information they uncover via real-time search and filter tools directly linked to the display.

2 Methods

2.1 Text Processing

The goal of our text preprocessing pipeline, as illustrated in Fig. 1, is to build a ranked, time ordered subset from the original Microblog file. A simple keyword search for relevant words from the problem domain produces an unranked "sick" file. Words such as "flu" are added one at a time, and if they result in significant growth in the "sick" file they are used for final analysis. The Python Natural Language Toolkit(NLTK)[1] is integrated into a custom script that produces 30 random concordances of words relevant to the problem domain. Visual analysis is used with the resulting concordances to build the grammar search scripts. The grammar extraction process ranks the blog entries and the ranked "sick file" is used in the viral tracker application, as seen on the right side of Fig. 1. We also produce a ranked event file using the same process.

2.2 Viral Tracker

Viral Tracker was built in C++ on top of SDL and OpenGL libraries. It supports loading multiple data sets which can be overlayed or individually filtered. Micro-blog and weather data have their own respective displays, each linked to a time slider(Fig. 2(a)). Users can temporally scan the data quickly via the slider(with icons to turn on/off various features), or use the play/pause buttons to perform more precise analysis. Beyond this there are 3 primary features, (1) The *Adaptive Grid Overlay*, which computes an average threshold and highlights sectors where Micro-blog posts exceed this value. It is adaptive in the sense that the grid automatically recomputes based on the chosen grid resolution and time slice being viewed. This can help users gain clarity on dense datasets and more easily identify trends. (2) Next is a *Ranked Filter* which is tied to the preprocessed data, allowing users to selectively refine visible data by rank. This was very helpful for noise reduction while still allowing users to verify which data they were omitting, (3) A *Search Feature* was also implemented, and linked to the datasets and the plotter. This tool can find and color code data points that contain the given search term. Examining blog data by clicking data points of interest allows users to hypothesize which terms may be useful in identifying a trend. In this challenge this feature helped to find and distinguish two mutually exclusive trends within a matter of seconds.

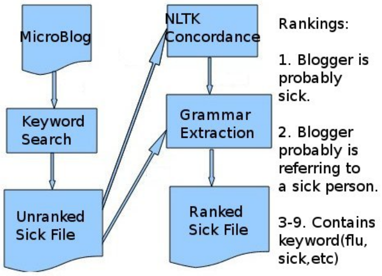

MicroBlog
Keyword Search
Unranked Sick File
NLTK Concordance
Grammar Extraction
Ranked Sick File
Rankings:
1. Blogger is probably sick.
2. Blogger probably is referring to a sick person.
3-9. Contains keyword(flu, sick,etc)

Figure 1: Text Processing Pipeline.

3 Hypothesis Construction

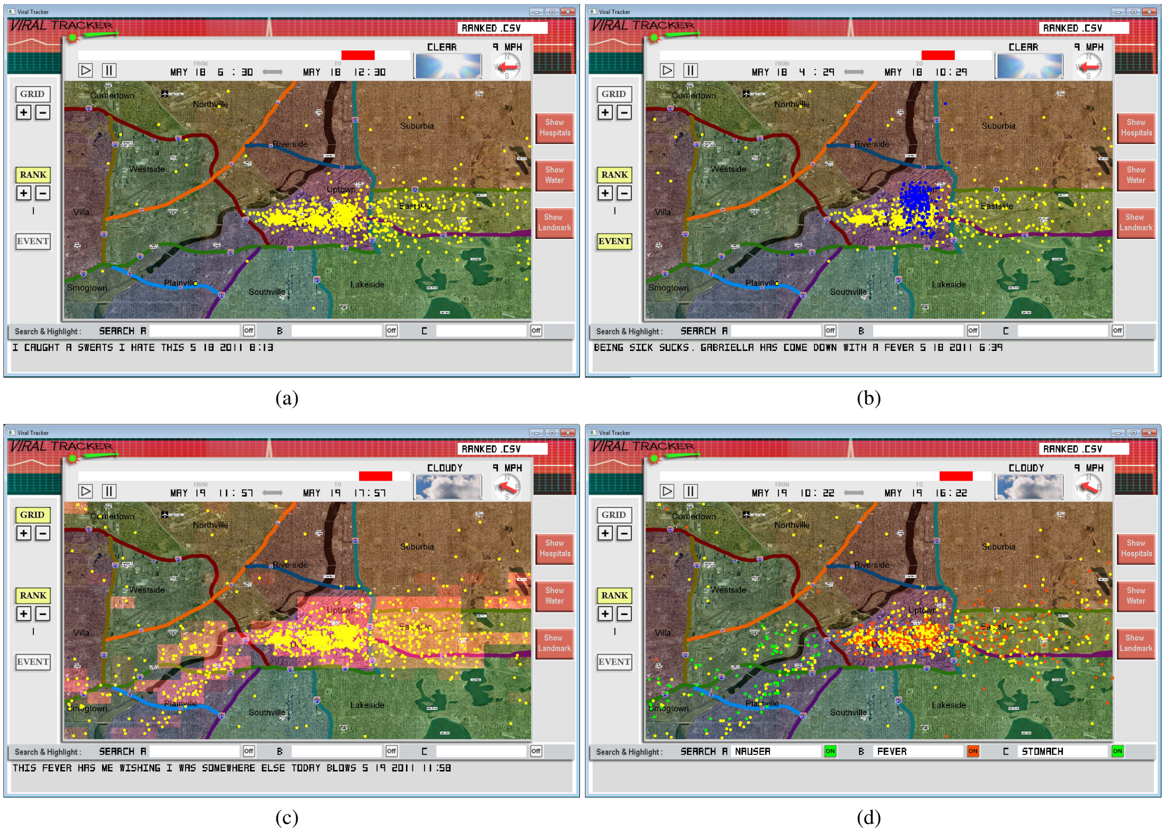
Hypothesis: Our analysis detected two distinct outbreaks, an airborne outbreak occurring in the Downtown district on May 18th and a waterborne contaminant that manifested itself on the 19th and makes its way out of the area by the 20th. The airborne outbreak is largely contained by the 20th while the waterborne contaminant may still be making its way down the Vast River and is worth noting to emergency personnel.

3.1 Visual Analysis and Reasoning Using Viral Tracker

The outbreak began on May, 18th at approximately 8AM in the Downtown region (Fig. 2(a)). We estimated Ground Zero within a short range of the Vastopolis Dome. The affected area extends eastward towards Interstate 278 and Eastside. This conclusion came after scanning a time lapsed plot of the micro-blog subset related to illness. After seeing signs of the initial event (the three clusters in Fig. 2(a)), we performed a minute by minute viewing of the time frame in question using the automated player. The analysis was refined by filtering the data by rank, focusing on the rank indicating the bloggers themselves were actually sick. This eliminated noise, revealing clear clusters of ill individuals which were confirmed by sampling blog data within the clusters.

Symptoms consistent with a flu or cold spread locally and very rapidly, particularly within densely populated areas. By super-

[*] e-mail: amidget3@uncc.edu
[†] e-mail: jguest@uncc.edu
[‡] e-mail:krs@uncc.edu



I CAUGHT A SWEATS I HATE THIS 5 18 2011 8:13

(a)

BEING SICK SUCKS. GABRIELLA HAS COME DOWN WITH A FEVER 5 18 2011 6:39

(b)

THIS FEVER HAS ME WISHING I WAS SOMEWHERE ELSE TODAY BLOWS 5 19 2011 11:58

(c)

SEARCH A NAUSEA ON B FEVER ON C STOMACH ON

(d)

Figure 2: (a)Initial Outbreak (Downtown Area): May 18, 6.30am-12.30pm. Virus spreads eastwards along Interstate 278, (b)Convention Center Overlay: May 18, 4.29am-10.29am. Blue dots indicate blog posts from the Convention Center, (c) May 19, 11.57am-5.57pm. Overlaid grid to better illustrate trends in viral spread, (d) May 19, 10.22am-4.22pm. In order to highlight the waterborne contaminant and its spread, specialized filters searching for 'nausea', 'fever' and 'stomach' highlight and distinguish the two types of viral spreads(green versus red dots(blogs)). Green dots are flowing along the river.

imposing the subset containing event activity, we could see a technology convention occurring slightly after the initial outbreak, shown by the blue dots in Fig. 2(b). The low correlation caused us to favor the airborne theory although we did not rule out person-to-person transmission. The affected area extended down Interstate 278, the most direct route to the Uptown/Downtown districts.

By that evening, (the time we estimated people returning home) we were able to see the virus popping up across the entire map (not shown). Even though the virus had left the Downtown region, its spread from outside areas was nowhere near as rapid, which we reasoned to mean people carried it home. We still did not rule out human contact due to lower population density as a possible cause for this.

On the 19th we were able to detect the virus trending largely in two directions. We further isolated this by overlaying the adaptive grid, as seen in Fig. 2(c). By using this we clearly defined a trend along the interstates through the east side, and another along the banks of the Vast River, moving southwest, and originating from the Downtown area.

The trend along the interstates was expected, however the trend along the river was a new development, needing further analysis. We randomly sampled a few points from the river to identify common symptoms. This quickly turned up terms such as 'stomach', 'pain', 'nausea', 'diarrhea', and the like, whereas other regions had flu/fever symptoms. We verified this using the search feature to highlight flu symptoms in one color and common waterborne symptoms in another. Indeed, it revealed a distinctly separate virus making its way down the river, as seen in Fig. 2(d), with the green dots corresponding to blogs matching 'nausea' or 'stomach', while the red dots match 'fever'. Additional confirmation of the waterborne contaminant was that its progression down the river was consistent with the flow(southward and out of Vastopolis). It left in a visual and predictable fashion, by the morning of the 20th, with just a few lingering traces. At this point the airborne virus also was in recession, with the infected in hospitals, confirmed by the samples indicating the virus as some type of flu-like strain.

4 Conclusion

Our *Viral Tracker* application was able to clearly and accurately define the vectors and timing for each of the two distinct outbreaks. The tool's simplicity, ease of use and interactivity were the strong features (as per feedback from the reviewers) that enabled the visual analysis and reasoning. We also felt that our methods to detect and distinguish the strains were effective and intuitive. However, we failed to detect the traffic accident that incited the outbreak. More rigorous analysis of event subsets would have likely revealed the incident.

References

[1] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, 2009.