Visual Analytics of Terrorist Activities Related to Epidemics

VAST 2011 Grand Challenge Award: "Outstanding Comprehensive Submission"

Enrico Bertini Juri Buchmüller Fabian Fischer Stephan Huber Thomas Lindemeier
Fabian Maaß Florian Mansmann Thomas Ramm Michael Regenscheit Christian Rohrdantz
Christian Scheible Tobias Schreck Stephan Sellien Florian Stoffel Mark Tautzenberger
Matthias Zieker Daniel A. Keim

Data Analysis and Visualization Group*
University of Konstanz, Germany

1 Introduction

The task of the VAST 2011 Grand Challenge was to investigate potential terrorist activities and their relation to the spread of an epidemic. Three different data sets were provided as part of three Mini Challenges (MCs). MC 1 was about analyzing geo-tagged microblogging (Twitter) messages to characterize the spread of an epidemic. MC 2 required analyzing threats to a computer network using a situational awareness approach. In MC 3 possible criminal and terrorist activities were to be analyzed based on a collection of news articles. To solve the Grand Challenge, insight from each of the individual MCs had to be integrated appropriately.

2 Analytic Process

Tools and Methods

All MCs used the data exploration platform KNIME[1] for preprocessing and automatic analysis. For each of the MCs, different tools were applied or integrated into self-written software. For MC 1, Tableau[2] was used for initial data understanding. For more specific analysis, an interactive visual analytics tool was written integrating text and geo-related data along the time dimension (see Figure 1). Apache Lucene[3] provided text indexing and querying capabilities. The Java imaging library nicejava[4] was used for image processing. To explore selected Tweet texts, we used the IBM Word-Cloud Generator[5]. Wireshark[6] was applied to analyze PCAP data in MC 2. Analysis of network relations was performed using the graph visualization tool Gephi[7]. Discovery of interesting events was supported by a self-written tool based on a MySQL[8] database (see Figure 2). Discovered events were further analyzed in detail using Tableau. For MC 1 and 3, both Apache Lucene and the IBM Word-Cloud Generator were used to provide word clouds with full-text search capabilities for overview and exploration (see Figure 3). The Stanford Named Entity Recognizer[9] was utilized to tag entities. Mallet[10] helped to discover topics. Complementary analyses used the visual text analysis system Jigsaw[11].

*Author e-mail addresses are: *firstname.lastname*@uni-konstanz.de

[1] KNIME (Konstanz Information Miner) - http://www.knime.org/
[2] Tableau - http://www.tableausoftware.com/
[3] Apache Lucene - http://lucene.apache.org/
[4] Java imaging library - https://github.com/santazhang/jpaint/
[5] IBM Word-Cloud Gen. - http://www.alphaworks.ibm.com/tech/wordcloud
[6] Wireshark - http://www.wireshark.org/
[7] Gephi - http://www.gephi.org/
[8] MySQL - http://www.mysql.com/
[9] Stanford NER - http://nlp.stanford.edu/software/CRF-NER.shtml
[10] Mallet - http://mallet.cs.umass.edu/topics.php
[11] Jigsaw - http://www.cc.gatech.edu/gvu/ii/jigsaw/index.html

Group Organization

For solving the Grand Challenge we formed three teams working to the individual MC tasks. Regular reasoning meetings among all MC teams were held, in which theories about cross-connections among the MC findings were discussed. Each meeting started by each group presenting their recent advances, problems, and findings. These meetings were used for all groups to jointly discuss their next steps, and at the same time brought the opportunity to identify possible relations between the MCs. For identifying such relations, initially we focused on temporal correlations, but later on came up with further findings.

Data Analysis Pipeline

Initially, we used general data analysis tools that could readily be applied to get an overview of the data and perform first preprocessing steps and analyses. The knowledge gained from this initial phase was used to further define more specific analysis tasks. We continued applying more specialized data analysis tools providing the corresponding capabilities. We tried to get as far as possible with existing methods. Where deemed necessary, we developed own visual analytics software in each of the MCs. When the amount of data was too large to be analyzed in our desktop office environment, we made use of the Konstanz Powerwall[12]. The large-size display (see Figure 3) allowed for high-resolution visualization, and in particular seemed to stimulate creative reasoning processes among the group.

3 Obtained Analysis Results

We introduce our self-written software for the different MCs, and describe obtained results. For sake of brevity, we discuss findings for MC 1 only[13]. Figure 1 shows our tool developed to solve MC 1. Shown are the critical days from May 18th to 20th. The core of the tool is the map display in the middle where the geo-location of each Tweet in the user-selected time-interval is displayed as a red dot. On the upper left, the weather conditions for each day are displayed. Further to the left analysists see the current date and have several options for configuration and filtering. At the bottom of each display, the volume of Twitter data is displayed over time. There, a play button starts and stops an animation for the location developments over time. When playing the animation over all days contained in the data set, the three displayed days stick out at a glance. For particular user selected time ranges, more detailed analyses can be performed. For example, a region can be selected in the map and all Tweets authored in that region in the selected time range are displayed in a separate window, where also keyword searches can be performed. An additional summary of all these Tweets can be generated on demand as a word cloud visualization. According to keywords contained in Tweets, we trained a

[12] Konstanz Powerwall - http://www.informatik.uni-konstanz.de/arbeitsgruppen/infovis/powerwall/
[13] Please see the accompanying video for all obtained results.

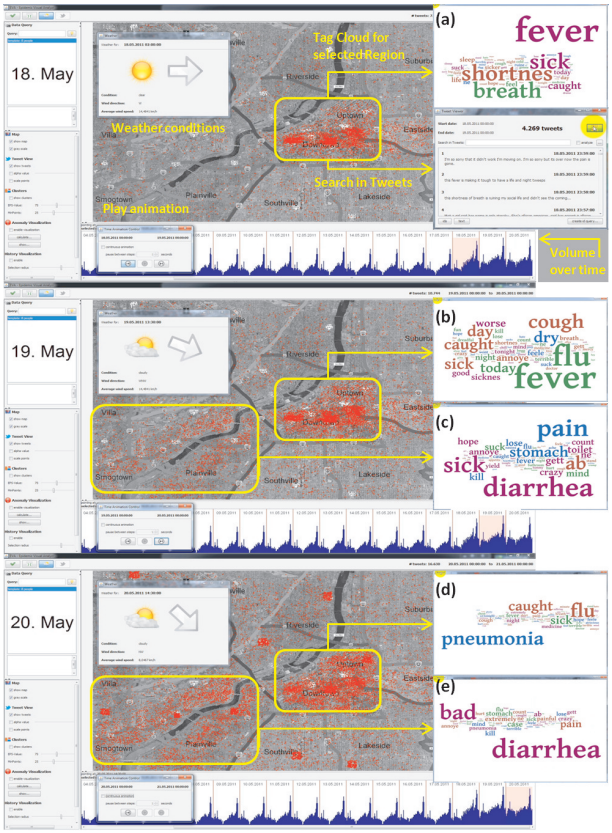

Figure 1: Tool designed for the solution of MC 1 showing geo-locations and text content of Tweets for different time ranges.

classifier distinguishing whether they report about sickness or not. In the screenshots of Figure 1 only Tweets reporting about sickness are displayed as red dots in the map. It can easily be seen that on May 18th, only the downtown area is affected, while on May 19th also the south-western part of the city along the river is affected. On May 20th, additionally a lot of smaller dense clusters show up; these correspond to the locations of hospitals. The word clouds to the right reveal further findings about development of characteristic symptoms. For the downtown area, the main symptoms developed from fever and shortness of breath on May 18th *(a)*, to fever, flu and cough on May 19th *(b)*. Finally, on May 20th pneumonia is reported *(d)*. Interestingly, the symptoms in the south western part of the town are different. On both May 19th *(c)* and May 20th *(e)* diarrhea is the dominating symptom that people report about. Both diseases, however, may have their origin at the same location right in between the marked regions. An algorithm we designed to discover anomalies in the geo-spatial distribution of Tweets identified a cluster right at that spot on May 17th. The Tweets talk about a truck accident with the truck spilling its cargo into the river. It is reasonable to hypothesize that the cargo was the cause of the disease and the spread happened both waterborne along the river and airborne in wind direction to downtown.

A screenshot of the situational awareness tool developed to solve MC 2 is provided in Figure 2. The tool contains a matrix visualization where the rows represent all the machines connected to the network and the columns correspond to the servers. Each cell represents the connections over time going from a specific machine to a specific server within a user-selected time window. Cells are vertically split into two parts. The upper one shows connections logged at the firewall in a green color tone, while the lower one shows IDS alerts in a red color tone. The details on connections can be easily accessed by mouse-click.

Figure 3 shows the interactive word cloud visualization tool developed to solve MC 3, running on the Konstanz Powerwall display. Word clouds for different days and based on different term scoring methods can be interactively searched and compared, fostering detection and understanding of relevant news topics.

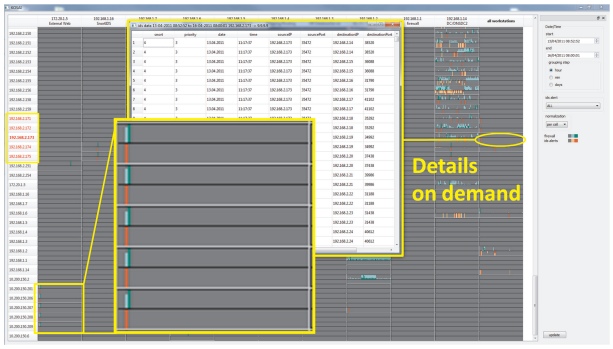

Figure 2: Tool designed for the solution of MC 2 showing network connections over time, separated into firewall logs and IDS alerts. Here, a pattern can be easily detected that shows several hosts behaving similar at the same point in time.

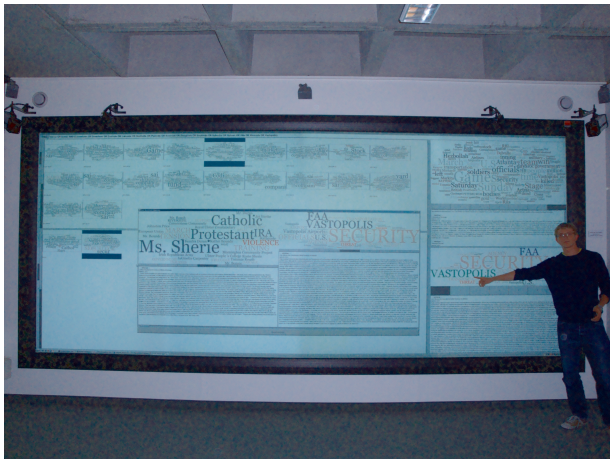

Figure 3: Comparative analysis of text clouds in MC 3 at the Konstanz Powerwall display.

4 Team Aspect in the Analysis Process

We believe that obtaining and comprehensively presenting our results was very much fostered by having regular team meetings. All teams presenting their intermediate analytic findings and technology-related experience in the group provided, as we believe, for an environment of creativity, cooperation, and motivating competition. At several points, we noticed that "the dots connected" to form a bigger picture with respect to the Grand Challenge. Repetitive consideration of the Challenge goals lead to high problem awareness among the teams, which eventually helped a lot in assembling a consistent final report.

Acknowledgements

We thank Curran Kelleher for discussion of intermediate results and great help with production of the final report video.