ScatterBlogs: Geo-Spatial Document Analysis

VAST 2011 Mini Challenge 1 Award: “Unique Integration of Tag Clouds in Geo-spatial Visualizations”

Harald Bosch∗ Dennis Thom∗ Michael Wörner∗ Steffen Koch∗ Edwin Püttmann†
Dominik Jäckle† Thomas Ertl∗

Institute for Visualization and Interactive Systems, Universität Stuttgart

Figure 1: The map of Vastopolis with overlays for message densities (gray colored areas with three bright hotspots), event labels of spatiotemporal term anomalies (text colored by the relevant time interval), and the highlighted relevant areas of the truck event (orange).

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—GUI; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Search process

1 INTRODUCTION

The VAST 2011 Mini-Challenge 1 (MC1) dataset featured roughly one million georeferenced micro blog messages during a flu-like epidemic outbreak in the fictitious city of Vastopolis. The challenge task was to identify the outbreak’s ground zero location, means of transmission, and the affected regions.

By providing a system for interacting with the data in a visual, direct, and scalable fashion, as well as offering diverse views and data management components, it was possible to solve the mini challenge correctly. Our system supports an analysis process allowing an iterative integration of qualitative as well as quantitative findings and it provides sophisticated hypothesis validation mechanisms. Visualizations for showing terms with space-time density anomalies, performing selective key term exploration, and managed clustering makes an interactive analysis possible.

2 ANALYTICAL PROCESS

Within the ScatterBlogs desktop analysts create and incrementally reformulate data sets for free exploration or validation of hypotheses. During this process the system supports the analysts in three basic ways: Firstly, it features overviews that suggest interesting artifacts in the dataset to be examined; Secondly, the currently examined sets of messages are shown in different views, thereby visualizing trends and patterns; And thirdly, it supports the quick creation of new sets, the reuse of existing sets for different purposes, and the easy recombination of sets – all through simple interaction.

At any time the analysts are in full control over the analysis work flow while the system keeps track of the different sets of data that were created. For the VAST Challenge we distinguish two phases of analysis.

In the first phase of the analytical process an overview of the situation in Vastopolis is obtained by textually querying all messages containing any of the symptoms mentioned in the challenge’s task description. Hereby, a small subset of the dataset is extracted and established as context for the subsequent analysis that can be performed using selection and highlighting techniques. Nonetheless, the complete dataset is still accessible and can always be included in the context by modifying the query. To explore a context the spatial density of messages can be shown on the city map for any given time-frame using a hierarchical time-range slider (cf. Fig. 2). For the given dataset this method instantly reveals that symptom related messages concentrate mainly in two dense spatiotemporal
clusters (cf. Fig. 1) – the first one covering a wedge-shaped area over Downtown and Eastside and the second one covering the riverside of Plainville, Smogtown, and Westside. Also, it can be noticed that at the end of the time period there are clusters of increased symptom mentions at the city’s hospital locations.

To get a quick impression of what is posted in the noticeable clusters we developed a tool that shows the most frequent terms of the messages captured by a movable lens. The terms are scaled according to their frequency and placed in a small tag cloud around the lens. By moving the lens over the map, it can be observed that people in the two wedge shaped clusters each talk about different kinds of symptoms. The symptoms observed along the river seem to be more gastrointestinal in nature and do not spread across other areas of the map, while the ones over Downtown and Eastside are mostly fever-related and can also be found throughout the city.

After having identified the affected regions, we look for possible causes for the epidemic outspread. For such tasks, we developed a tool supporting the fast retrieval of event-related key terms by means of geolocated tag clouds. Based on the hypothesis that local events result in high concentrations of similar term usage in dense spatiotemporal areas, the dataset is searched for such anomalies. The relevance for each possible key term is determined by the significance of its anomalies. Consequently, tags are placed on the map near the location of the term’s most significant anomaly and the size as well as the order of placement is determined by the term’s relevance. If the optimal location is already occupied by a more significant term, the tag is placed in the nearest vicinity using an approach similar to Luboschik et al. [2]. Once generated, the tag overlay allows an interactive exploration of events by hovering over the tags to highlight the regions affected by related anomalies. By selecting a term it can be added to the current query thereby including all of its messages into the analysis context.

We use the key term overlay of the whole dataset to locate a truck accident on a bridge right between the two affected regions. Selecting the messages of this event with the lens and browsing the message contents reveals that this accident involved fire and spill cargo. At this point the hypothesis can be formulated that the truck accident resulted in a spread of infectious materials both airborne over the downtown area and waterborne down the river.

In the second phase of the analysis process, we validate this hypothesis by excluding other possible means of infection and by checking the spatiotemporal correlation of the truck accident and the illness. The spatiotemporal correlation of the truck accident and the outbreak can be validated using a three-dimensional representation of the messages over space and time as depicted in Fig. 2. Here, it can be observed that the messages concerning the truck accident are in close spatiotemporal proximity to the messages mentioning flu-like and gastrointestinal symptoms which supports the hypothesis that these events are somehow correlated. To rule out the possibility of person-to-person transmission, we check whether all hospitalized persons, who reported symptoms, were exposed in the affected region. We adopted a selection management component from PatViz [1] to formulate the complex queries required. Here, one can interactively define and combine selections using basic set operations. We create one set containing all hospitalized users and another set of all exposed users through textual and spatiotemporal selections. Intersecting them by user ids shows that almost all hospitalized users were directly exposed in the flu-symptom area. We consider this as an indication that there is no person-to-person infection. Similarly, it can be noticed that none of the persons infected in the gastrointestinal outbreak receives treatment in hospitals.

3 Conclusion
We presented a system for microblog analysis that seamlessly integrates search backend and visual frontend. It provides powerful, automatic algorithms for detecting spatio-temporal ‘anoma-

Fig. 2: The ScatterBlogs analysis desktop comprises amongst others: textual querying, a time-range slider including a histogram and weather information, blog scatterplots on the map and in a timeframe cube, a map lens with a local tag cloud, a subset management component combining and tagging selections, and a table of the messages. The current message set is manually tagged to emphasize the two affected areas (green and yellow) and the truck accident (blue). Note the temporal and spatial proximity.

REFERENCES

ACKNOWLEDGEMENTS
We would like to thank the IEEE VAST Challenge 2011 team and all reviewers. This work was partially funded by the the German Federal Ministry for Education and Research (BMBF), the German Science Foundation (DFG), and the European Commission as part of the VASA project, the SPP 1335, the FP7-project PESCaDO (FP7-248594), and the Graduate School of Excellence for advanced Manufacturing Engineering (GSaME).