Visual Analytics of a Pandemic Spread

VAST 2010 Mini Challenge 2 Award: Thorough Description of Analytic Process

Andrada Astefanoaie* University of Konstanz
Rodica Bozianu† University of Konstanz
Marc Broghammer‡ University of Konstanz
Roland Jungnickel§ University of Konstanz
Christian Rohrdantz¶ University of Konstanz
Juergen Schniertshauer‖ University of Konstanz
David Spretke** University of Konstanz
Peter Bak†† IBM Haifa Research Lab Israel

1 Introduction

The task of the VAST 2010 Mini Challenge 2 was to characterize the spread of an epidemic outbreak. The analysis should take into consideration symptoms, mortality rates and temporal patterns of the disease. Finally, the outbreak should be compared across different locations searching for anomalies.

For the preprocessing and the automated analysis of the data we used the Konstanz Information Miner (KINME)[1], which is a modular data exploration platform that enables the user to visually create dataflows, the R[2] software for statistical computing and some self-written Java programs for data preprocessing. For conducting the visual investigations we applied Many Eyes[3] and Protovis[4], which compose scalable and customized views for datasets of interest.

2 Analytic Process

To find the answers for both questions of the Mini Challenge, an analytic pipeline was designed which combines automatic and semi-automatic data analysis with interactive visual explorations. The analytic pipeline (see Figure 1) is divided into three parts: preparation of data (Section 2.1), information extraction and visual analysis (Section 2.2) and knowledge extraction and results' refinement (Section 2.3).

2.1 Preparation of data

The first part of the analytic pipeline, common to both questions of Mini Challenge 2, is comprised of data preprocessing and the initial analysis. The raw input data consists of hospitalization and death records for 11 locations. An initial statistical analysis helped us to gain general knowledge about the data, like the average mortality rate and the fact that males and females were affected to the same extent by the disease.

For further automatic and visual analyses, the different data sources about the patients hospitalization and death had to be merged first. Afterwards, the symptom descriptions for patients had to be processed in multiple steps, from which the removal of duplicate symptoms and their cleansing took major efforts (several hours).

2.2 Information Extraction and Visual Analysis

The number of occurrences of each symptom on each day for both hospitalizations and deaths, as well as the mortality rate for each

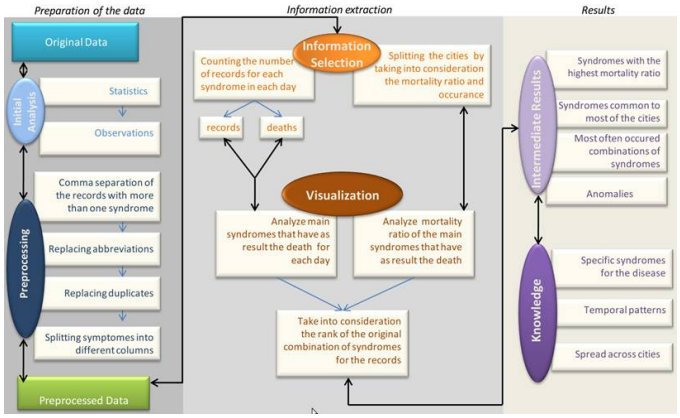

Figure 1: The analytic pipeline represents the workflow that was used to extract the information needed to answer the Mini Challenge.

symptom were calculated automatically. As a result, we gained some indication which symptoms relate to the disease.

In a confirmatory step, the correlation of the symptoms was calculated and visualized with the arc graph presented in Figure 2. The visualization shows that some symptoms are strongly and others are not correlated with the main symptoms.

The small multiples visualization from Figure 3 enables to see further temporal aspects of the data. The chart reveals the evolution of deaths for each location - taking into consideration the day of hospitalization and the day of death (left column), the total number of infected patients (middle column) and the temporal evolution of infected people that died (right column). It is clearly visible that the locations Turkey and Thailand pattern differently and were probably not affected by the disease. More details about which symptoms were involved in which location can be derived from the temporal heatmap in Figure 4. The chart shows the change of mortality rates over time. We expected to see a simple correlation between symptoms and mortality rate. But, it seems that mortality rates do not follow a clear pattern as a function of symptom or time. However, an anomaly of the mortality rate peaking towards the end of the recovery period is perceivable for several countries.

The world map in Figure 5 was created to explore potential correlations between the location of a country and different data attributes, as well as to compare the spread across countries. The most affected country is Syria, being represented by Aleppo. The next one is Kenya (represented by Nairobi) with a marginal difference.

*e-mail: andrada.astefanoaie@uni-konstanz.de
†e-mail: rodica.bozianu@uni-konstanz.de
‡e-mail: marc.broghammer@uni-konstanz.de
§e-mail: roland.jungnickel@uni-konstanz.de
¶e-mail: christian.rohrdantz@uni-konstanz.de
‖e-mail: juergen.schniertshauer@uni-konstanz.de
**e-mail: david.spretke@uni-konstanz.de
††e-mail: peterba@il.ibm.com

[1] KNIME (Konstanz Information Miner) - http://www.knime.org/
[2] R - http://www.r-project.org/
[3] Many Eyes - http://manyeyes.alphaworks.ibm.com/manyeyes/
[4] Protovis - http://vis.stanford.edu/protovis/

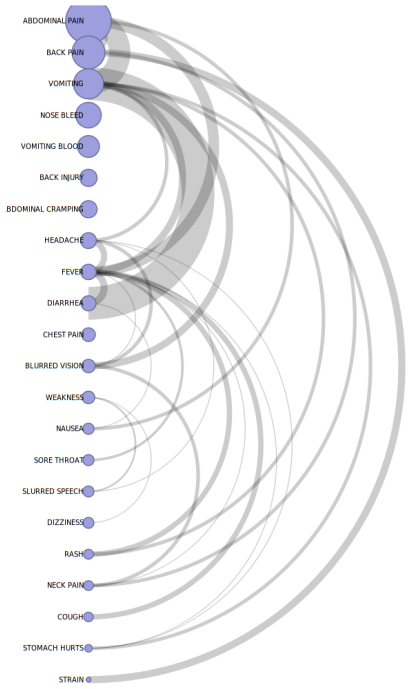

Figure 2: The Arc Graph shows the co-occurrences between the most common symptoms. The size of each node represents the number of occurrences of the symptoms and the width of the arc represents the number of records with both symptoms.

Figure 3: The Small Multiples Chart.

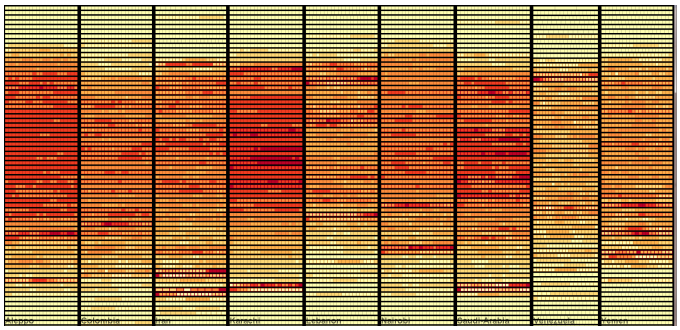

Figure 4: Heatmap visualizing the change of mortality rates over time: Every column represents a country. Within each country time is mapped from top to bottom. Symptoms are mapped from left to right. The more red, the higher the mortality rate is.

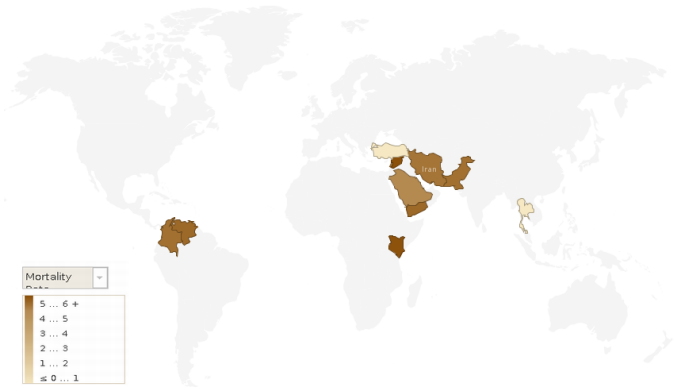

Figure 5: The World Map shows the average mortality rate per country in the geospatial context.

2.3 Results

Based on all the aforementioned and further visualizations several interesting findings were made. First, both major symptoms (Vomiting, Diarrhea, Nose Bleeding, Abdominal Pain, Back Pain) and minor symptoms (Conjunctivitis Red, Encephalitis, Facial Swelling, Hearing Loss, Proteinuria, Tremor) of the disease were detected. Nairobi could be identified as the first location to be affected and to reach the peak. The peaks of the disease were reached on May 16th in Kenya (Nairobi), May 17th in Syria (Aleppo), May 19th in Lebanon, Yemen and Pakistan (Karachi), May 20th in Venezuela, Saudi Arabia and Iran and May 21st in Colombia. The temporal developments of mortality rates showed that the severity of the disease was not the same for each country. Turkey and Thailand might be considered as anomalies because they were not affected. Finally, it became obvious that the spread of the disease was at least partly caused by infected people flying from one country to another: Not all affected countries border each other and the time spans between the outbreaks were very short.

3 Lessons Learned

The clearly defined analysis pipeline enabled us to solve the analysis task of Mini Challenge 2 by mainly applying existing tools. Whereas the preprocessing required some own code writing and took a considerable amount of time, once the data was in a structured and clean format it could be easily visualized. Quite a number of different visualizations could be tested without too much additional effort and some of them immediately revealed interesting additional insights, while others helped to confirm previously generated hypotheses.