VAST Challenge MC1: An Off the Shelf Approach to Messy Data

Fintan McGee* Bertjan Broeksema† Benoît Otjacques‡

Centre de Recherche Public Gabriel Lippmann

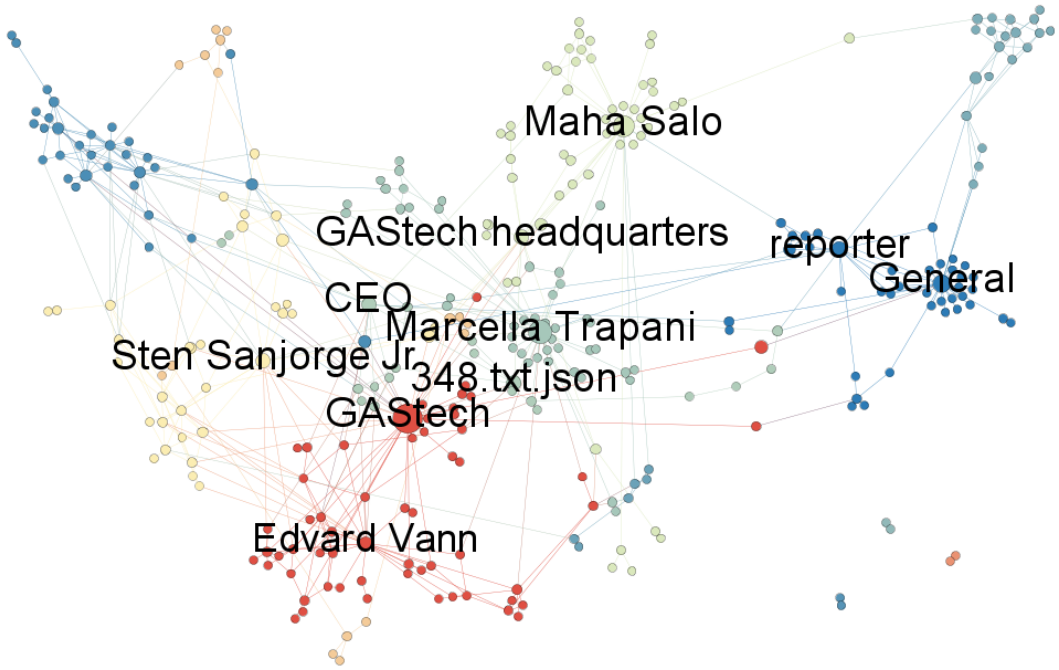

Figure 1: A graph of MC1 challenge articles of January 20th and extracted entities.

ABSTRACT

We describe our approach to the VAST challenge using off the shelf software, for analysis and visualization. This allowed us to gain insight into the data within a short time-frame. We discuss the merits of our approach and set out directions for future research based on our experience.

1 INTRODUCTION

Intelligence operations often occur under time pressure and deal with messy data. Consequently, tools for analyzing data under such circumstances must be both versatile and flexible.

We have been looking at the VAST 2014 challenge MC1, which involved the analysis of a collection of text documents. The goal of this challenge was to retrieve background information about an (fictive) organization that was supposedly involved in the disappearance of employees of a gas drilling company. Additionally, a series of events surrounding the disappearance had to be characterized.

Due to practical constraints we could only spend limited time on this project. We were two people who both spent one week full-time on analyzing the data and summarizing our findings. We took this constraint as part of the challenge. Due to this we also could only use of-the-shelf tools to dive into the data, as opposed to developing new tools.

*e-mail:mcgee@lippmann.lu

†e-mail:broeksem@lippmann.lu

‡e-mail:otjacque@lippmann.lu

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

2 DATA PROCESSING

Both analysis and visualization of data require the data to have a well-defined structure. Typically, it should be in a tabular format, where each column represents one variable and each row one observation. Transforming data into a tidy format is discussed extensively by Wickham [4]. As in many cases, the data provided for the challenge had a high degree of messiness. It came in different formats (pdf, word documents, plain text) and the degree of structure in the documents differed considerably as well. Additionally, the quality of the information itself had also a high variability. For example, a list of employees of the gas company was highly structured and had (apparently) a low amount of erroneous information in it. The text documents on the other hand where somewhat structured, but not consistently the same and many documents appeared to come from automated translation services. As a result many of them had strange translations and many misspellings of names.

A related issue to cleaning up data is reproducibility. Initially the goal of the analysis is to find the information leading to the liberation of the hostages. However, in later stages, for legal reasons, it will be important as well to be able to reconstruct how certain pieces of information were found.

Given these consideration we chose R[2], to process the data at hand. It provides a highly flexible and interactive environment to process raw data files and for quick generation of plots. The

scripts work on the raw data and can be re-run to reproduce the cleaning we did in order to find the results. Changes to these scripts, where tracked using version control for further reproducibility of the analysis.

2.1 Data Analysis Approach

RStudio was used to explore the cleaned data and to create various plots during our exploration phase to gain insight in the data. Plots that gave some important insights in the data, and the code that generated them where stored as well. This resulted in an iterative process that lead to further refinements to the cleaning scripts. Our approach enabled: quick creation of plots of parts of the data, easy sub-setting and visually inspection parts of the data, and update cleaning code as the analysis process went on. These benefits where key for a quick understanding of the data, without spending too much time upfront cleaning the data.

We highlight one part of the analysis here, which is the analysis of the text documents. These came as separate files, which had some structure, but not exactly the same structure in all files. Our cleaning scripts took these files, and turned them into a simple table. Each row in this table has the fields: file, date, source, subject and text. Simple histograms and NA-counts where used to check for possible anomalies in our cleaning code and/or in the data.

This approach proved to be very versatile and lightweight, enabling quick exploration of the data from various angles. One of the documents provided, described the history of the POK until 2009. So, to gain insight in what happened in 2009 and later, we subsetted our articles table with respect to Date. Next we generated a histogram based on the dates to get a sense of when important events might have happened, shown in Fig. 2. As can be seen there are various peaks, one them, the one in 2012, reflects a series of articles about people remembering the death of Elian Karel. It was unknown to us that Elian had died, so we used this as a starting point for further investigation of what had happened.

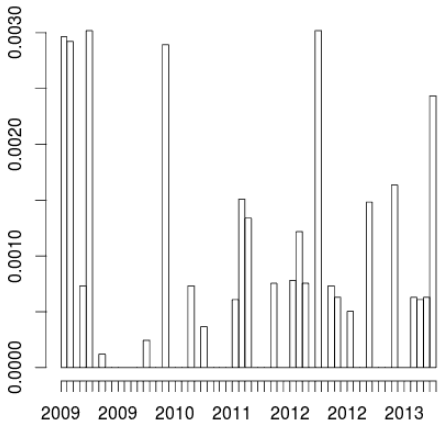

Figure 2: Article count per month in 2009-2013

2.2 Graph Visualization

Once the source data had been cleaned and homogenized, we used the OpenCalais[3] Named Entity Recognition web-service to extract entities form the documents. The JSON data contain the extracted entities were converted into a bimodal graph, where entities were connected by edges to the articles that referenced them. For the longer documents, such as the historical analysis of the POK, entities were extracted on a per paragraph basis. We utilized Gephi [1] to provide visualization of the related entities in the documentation. Gephi allowed for simple visual transformations and graph clustering. It also allows for graph operations such as inducing sub-graphs based on the ego network of a node of interest. This approach allowed for rapid investigation into events, such as the deaths of the Edris family, to determine whether or not they linked to the kidnapping event. Figure 3 shows a graph produced by this approach, when investigating the current leadership of the POK. The email data also provided a bi-modal network that allowed us to visually inspect a subset of employees who were related to POK member and the message shared between this group.

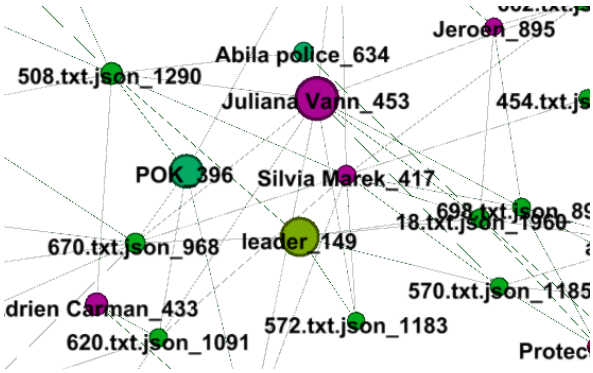

Figure 3: An image showing part of the union of the ego-networks of Elian Karel and the term “Leader”. Document 670 connects the terms “POK”, “leader”, and “Sivia Marek”. It’s contents confirm she is the current leader of the POK

3 Discussion

Our analysis approach enabled us to provide possible explanations of the events on Kronos. The textual analysis using R provided the most effective approach to determine the events of January 20th to 21st. Visualizing a network extracted using these dates (figure 1) helped to further improve the text analysis approach, by highlighting the people who were important to the events on these dates. In general the graph visualization provided feedback for the R textual analysis. However, it should be also noted that the process of cleaning the data and analysis it using R, preempted some of the discoveries. That is, some important events were stumbled upon during purely textual analysis. This is most likely an artifact of having a limited data set size. The graph visualization did allow us to quickly validate or reject theories, for example it helped us reject the idea of links between the Edris family and the POK.

4 Conclusions

Our experience of the VAST challenge gave us many insights into dealing with large messy data sets. The use of off the shelf tools allowed us to gain insight with minimal effort. From the beginning we were able to analyze data instead of developing a new visual analytics system. It is clear that we would have benefited from the integration of visual analytics into the initial data cleaning at the very beginning of our analysis.

The lack of structure in the competition data added to the challenge. Many visual analytics approaches assume a structure, even as simple as a table. Designing approaches that are independent of initial structural assumptions, but which try to combine forming structure and performinganalysis, might be an interesting future research direction.

References

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[2] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[3] T. Reuters. Open calais. `http://www.opencalais.com/`, 2008. Accessed: 2014-08-13.

[4] H. Wickham. Tidy data. *Under review*, 2014.