Safeguarding Abila: Discovering Evolving Activist Networks
VAST 2014 Mini Challenge 1: Unstructured Text and Network Data Analysis

Parang Saraf* Patrick Butler† Naren Ramakrishnan‡

Discovery Analytics Center
Department of Computer Science
Virginia Tech

ABSTRACT

We introduce a system for visual analysis of news articles and emails. This system was developed in response to VAST Mini-Challenge 1 and comprises different interfaces for mining textual data and network data.

Index Terms: H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Interaction styles (e.g., commands, menus, forms, direct manipulation)

1 INTRODUCTION AND PROBLEM OVERVIEW

The VAST 2014 Mini Challenge 1 describes a hypothetical scenario where some of the employees of an imaginary organization, GAStech have gone missing and it is speculated that an environmental activist group, Protectors of Kronos (POK) is responsible behind the disappearance. The provided dataset includes a set of GAStech and POK related news articles from various news sources, resumes of selected GAStech employees, organizational structure, and email headers of the emails exchanged between employees for two weeks leading to disappearance. The challenge requires uncovering the current organizational structure of POK and its evolution over time. Additionally, a detailed timeline of key events on the day of disappearance and the following day is also required.

2 SYSTEM DESIGN

We developed a web-based visual analytics system for analyzing unstructured textual data and network data. The system provides several widgets that empower an analyst with analytical tools required for uncovering hidden entities and their temporal distribution in textual data, as well as identification of significant, co-occurring nodes in email network data.

2.1 News Analyzer Interface

The news analyzer interface provides mechanisms for keyword based querying of articles, comparison of keyword trends over time and identification of entities relevant to searched keywords. The search interface makes use of a Python based search engine, Whoosh that allows for text indexing, parsing of logical queries using operators such as AND, OR, NOT and scoring of search results based on different algorithms. For the top five articles in the returned search results, the system identifies three most similar news articles and display them as related items (see Figure 1). Similarity is based on a vector of TFIDF scores. The interface further allows for filtering of search results by news sources and date range.

In order to identify keyword trends in news articles, a frequency over time plot is used that compares number of occurrences of

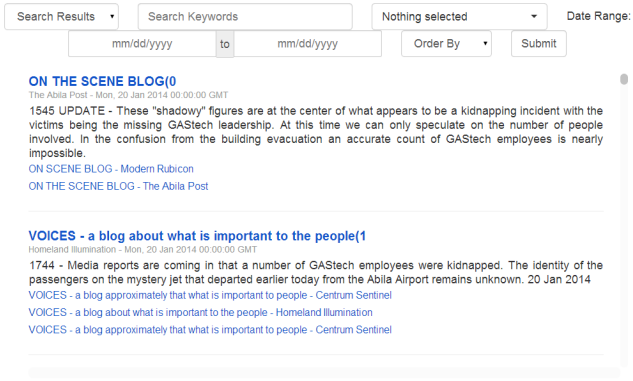

Figure 1: News Search Interface

searched keywords. The plot allows an analyst to understand the popularity of keywords at different time intervals and also identify correlated keywords by comparing individual time series. The plot shown in Figure 2 compares the keywords *POK* and *leader* and identifies all the instances where they peak simultaneously. The plot also provides a view-finder functionality that can be used to zoom-in and visualize only the selected time duration.

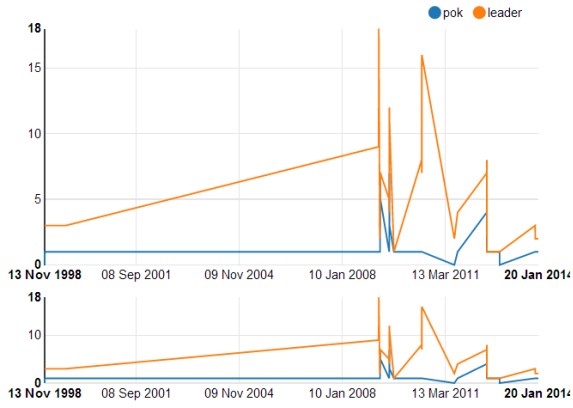

Figure 2: Comparing Keywords Frequency over Time

One of the salient features of the news analyzer interface is the dynamic generation of named entity word clouds. The Stanford NER parser [2] is used on resultant searched articles and three classes of named entities - person, location and organization are identified. The top 50 entities in each of the three classes are shown in separate word clouds. This analysis helps in uncovering hidden entities that are relevant to the searched keywords. Since the corpus had many hypothetical entities that were not in the training set, classification was not 100%. However, results were sufficient to

*e-mail: parang@cs.vt.edu
†e-mail: pabutler@vt.edu
‡e-mail: naren@cs.vt.edu

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

Figure 3: Dynamically generated word cloud of NER class-type *Person* for different date ranges for the query: *POK AND leader*

provide a satisfactory idea of the common feature space. Figure 3 presents word clouds of class-type person generated for three different date ranges for the same search query: *POK AND leader*. The prominent keywords in these word clouds help in identification of POK leaders during these periods.

2.2 Email Analyzer Interface

The email analyzer interface provides a search mechanism for querying email headers either by subject or by participants. The resulting emails can be filtered by department or date and are displayed graphically using a radial chart (see Figure 4) and textually as email threads by grouping emails with same subject and participants together. Further, the resultant emails are classified into three categories - *Company-wide emails* that are sent to most employees in the company, *Department-wide emails* that are sent to most employees in the department and *Personal emails* that are exchanged between selected employees. The checkboxes beneath the radial chart are used to select particular email types. Senders and receivers of emails to/from a particular employee are displayed in different colors in order to easily identify the underlying email network.

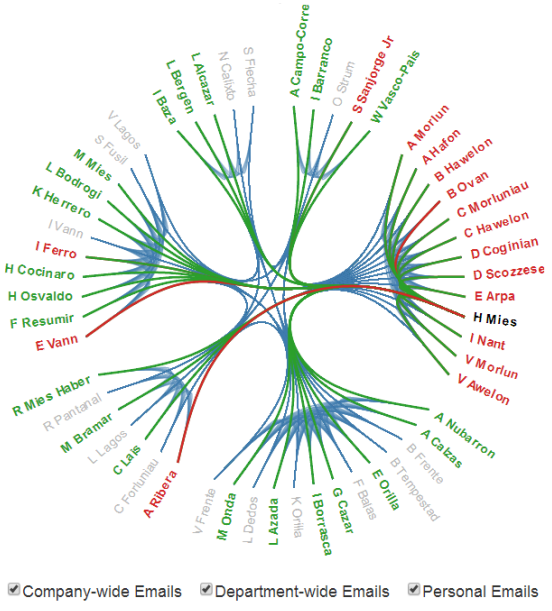

Figure 4: Network of Emails Exchanged

In order to identify social groups of employees that frequently exchange emails among themselves, a co-occurrence matrix representation (Figure 5) was used. The matrix employs the spectral co-clustering algorithm [1] that groups elements together whose values are larger than the corresponding rows and columns. The co-clustering algorithm was implemented using the Python based scikit-learn package [3].

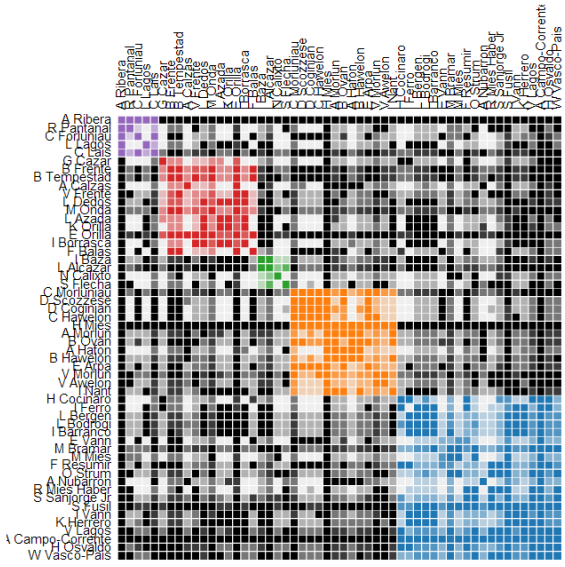

Figure 5: Co-occurrence matrix between email participants

Acknowledgements

We would like to thank Ritika Dokania for her creative inputs and feedback on the visualization, as well as for lending her voice to the explanatory video that describes the system. This work is partially supported by US NSF Grant CCF-0937133.

References

[1] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

[2] T. G. Jenny Rose Finkel and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.