Event-Based Text Visual Analytics

Ji Wang, Lauren Bradel, Chris North

Abstract—We present an event-based approach for solving a directed sensemaking task in which we combine powerful information foraging tools with intuitive synthesis spaces to solve the VAST Challenge 2014 Mini-Challenge 1. A combination of student-created and commericially available software are used to solve various aspects of the scenario. In addition to applying entitiy extraction and topic modelling, we enable the user to explore a large dataset using multi-model semantic interaction, which infers analytical reasoning from user actions to augment the data spatialization and determine what information should be presented and suggested to the user. Additionally, we visualize extracted topics using Tableau to construct a timeline of events surrounding the questions posed by the challenge.

Index Terms—Sensemaking, topic modelling, semantic interaction, event extraction.

INTRODUCTION

Entity extraction, topic modelling, and event extraction are powerful text processing techniques that have the potential to greatly improve the sensemaking process by foraging for potentially useful nuggets of data. While these techniques are useful on their own, they can be enhanced through integration with synthesis-oriented sensemaking tools. The coupling of these techniques with visual analytics tools, particularly ones that can learn from user interactions, enables users to extract information from large text corpora while filtering out less relevant information.

In our attempt to solve the VAST Challenge 2014 Mini-Challenge 1, we leveraged two visual analytics tools, StarSPIRE and Tableau, in order to determine the organizational structure of the Protectors of Kronos organization and identify the events surrounding the GASTech employee kidnappings, respectively. These techniques demonstrate how powerful foraging techniques can be harnessed in a synthesis-oriented space to facilitate knowledge extraction from a large and diverse text corpora.

1 UNDERSTANDING THE STRUCTURE OF POK

In order to determine the structure of Protectors of Kronos as well as understand how the structure had evolved, we utilized StarSPIRE, a visual analytics tool prototype specializing in unstructured text analysis. StarSPIRE [Figure 1] is a multi-model semantic interaction tool, which couples foraging and synthesis activities through learning from user interactions to adjust a model of the user's interests within the dataset. This model is then used to rearrange data, represented as a node-link diagram, in the workspace as well as determine what data should be displayed and what should be stored out of sight. This push-and-pull nature of disclosing only a portion of the dataset that exceeds a variable relevance threshold allows the user to explore the dataset incrementally and according to their current interests. Visual encodings such as node size, color, and saturation are used to denote the perceived relevance of the document in order to guide the user through the documents. StarSPIRE is best designed for directed sensemaking tasks, as it presents the user with a blank screen and relies on a search to gather initial data onto the workspace. Future iterations of this tool prototype may include providing the user with an overview of the dataset through topic modelling or clustering. However, this was not necessary for this task due to the directed nature of the VAST Challenge questions. More on the specific interactions afforded by StarSPIRE can be found in [3].

In the past, StarSPIRE had been used only on plain text documents. However, this year's VAST Challenge presented the challenge of integrating data from multiple file types. These files included plaintext news articles, Microsoft Word documents for historical summaries and employee resumes, and Microsoft Excel spreadsheets detailing employee information as well as email headers. We were tasked with converting this data into meaningful documents that could be parsed by an entity extractor. We created Python scripts to automate the parsing process. For the employee resumes, we stripped out the text and discarded the text markup to create plaintext files labelled with the name of each corresponding employee. For the historical documents provided, each sub-header was given its own separate document in order to create smaller files that would be better focused around a single topic. The spreadsheet data was parsed by row in order to convert the tabular data into plaintext form while preserving the column titles for easy reference. This process resulted in approximately 2,100 separate text files, which were then processed an entity extractor [1] resulting in approximately 2,200 entities. The text files and the associated entities were then loaded into StarSPIRE, which took under ten seconds.

Analysis began with a search for "Protectors of Kronos." This search yielded historical documents as well as news articles. Through actions such as highlighting names of prominent members, overlapping highly relevant documents, and adding annotations to documents, the workspace evolved to depict the history of the organization as well as current activities. We positioned documents in the workspace to reflect different topics, such as the foundation of POK, the alleged murder of a POK leader, and POK's relation to the greater Kronos region. With each interaction, StarSPIRE adjusted the user's interest model, fetched or removed documents, and

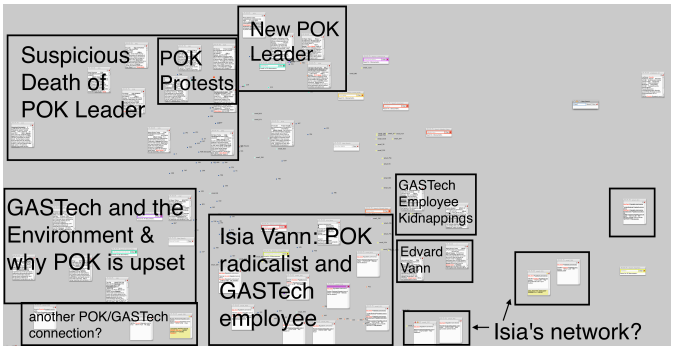

Figure 1. StarSPIRE workspace indicating loose clusters which were jointly constructed by the analyst and the underlying algorithms.

• *Ji Wang, Lauren Bradel, and Chris North are with Virginia Tech and can be reached at {wji, lbradel1, north}@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

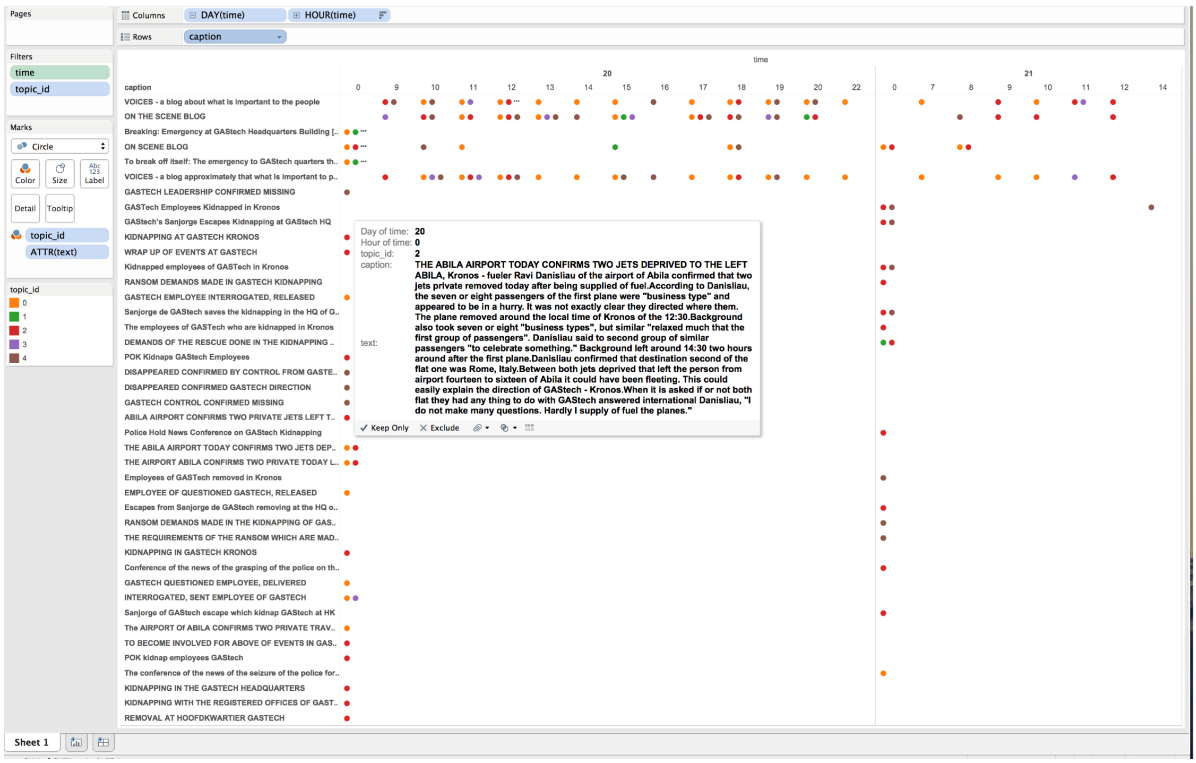

Figure 2. Tableau scatterplot view visualizing documents and their associated extracted topics along a timeline. Different colors indicate different topics extracted using Latent Dirichlet Allocation (LDA).

repositioned the spatial layout. The quick interaction-feedback loop enabled the analysis to proceed smoothly without unnecessary downtime between interactions.

2 Event Extraction

To determine the events concerning the kidnapping of GASTech employees, we applied entity extraction [4] and topic modelling [2] algorithms, which enabled us to extract events from the dataset in a specific timeframe that were subsequently visualized [Figure 2]. Dates were extracted from the news articles, allowing us to construct a subset of documents that pertained to the dates specified in the mini-challenge. Additionally, entities such as location and persons were identified. After obtaining a subset of news articles, we ran a topic modelling algorithm in order to extract the main themes of the articles and hone in on events. We leveraged Latent Dirichlet Allocation (LDA), which represents documents as a probabilistic mixture of topics [2]. We chose to set the number of topics at five. We experimented with altering this parameter but found that for this particular task, five topics were adequate to extract the relevant events. This resulting data was loaded into Tableau and visualized along a timeline through the scatterplot view [5].

By checking the different topic's news articles in scatter plots for the detailed content, we summarized the major events of the dates in question efficiently. Through this analysis, we were able to construct a timeline of the events surrounding the kidnapping of several GASTech employees as well as construct two hypotheses regarding the kidnappings.

3 Conclusion

Through our analysis, we were able to establish the history and evolution of an organization, identify ties to a local business, and extract events occurring within a given timeframe. By combining text processing techniques with visual analytics capabilities, we were able to construct two plausible scenarios as a solution to VAST Challenge 2014 Mini-Challenge 1. We leveraged appropriate tools (StarSPIRE and Tableau) to answer questions that were best suited to their strengths (directed sensemaking and sequential event tracking, respectively). We constructed two competing hypotheses regarding the kidnapping of GASTech employees. The evidence obtained through our analysis of the data supported both hypotheses. While this evidence was not sufficient for solving the overall scenario, we believe that our techniques demonstrated their usefulness in directed sensemaking sand timeline construction.

Future work includes investigating additional visual encodings and interaction techniques to enhance directed and undirected sensemaking tasks. Additionally, we intend to integrate topic modelling into StarSPIRE and add a temporal arrangement feature in order to be able to use a single tool to complete scenarios such as the one posed in the VAST Challenge 2014 Mini-Challenge 1.

Acknowledgments

This work was supported in part by NSF grants IIS-1218346 and SES-1111239.

References

[1] Baldwin, B. and Carpenter, B. LingPipe. *Available from World Wide Web: http://alias-i.com/lingpipe*.

[2] Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. *the Journal of machine Learning research*, *3*. 993-1022.

[3] Bradel, L., North, C., House, L. and Leman, S. Multi-Model Semantic Interaction for Text Analytics *Visual Analytics Science and Technology 2014*, IEEE, Paris, France, 2014.

[4] Finkel, J.R., Grenager, T. and Manning, C., Incorporating non-local information into information extraction systems by gibbs sampling. in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (2005), Association for Computational Linguistics, 363-370.

[5] Hanrahan, P., Stolte, C. and Mackinlay, J. Tableau Software, Visual Analysis for Everyone, 2007.