Jigsaw to Save Vastopolis

VAST 2011 Mini Challenge 3 Award: "Good Use of the Analytic Process"

Elizabeth Braunstein* Mercyhurst College
Carsten Görg† Univ. of Colorado Denver
Zhicheng Liu‡ Georgia Tech
John Stasko§ Georgia Tech

ABSTRACT

This article describes our analytic process and experience of using the Jigsaw system in working on the VAST 2011 Mini Challenge 3. We describe how we extracted and worked with entities from the documents, and how Jigsaw's computational analysis capabilities and visualizations scaffolded the investigation. Based on our experiences, we discuss desirable features that would enhance the analytic power of Jigsaw.

Keywords: Visual analytics, investigative analysis, information visualization, data ingestion, evidence marshalling

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;

1 INTRODUCTION

We worked on the VAST 2011 Mini Challenge 3 using Jigsaw, a visual analytics system designed for investigative analysis of text documents. Jigsaw provides multiple coordinated visualizations of the relationships between entities and documents, and integrates computational techniques such as analyses of document clusters, document similarity, and document sentiment. For details on these capabilities of Jigsaw, please refer to the articles [3, 2] and the project website [1]. In this article, we discuss our analytic process and experience of using Jigsaw to solve the mini challenge.

2 ANALYTIC PROCESS

We imported the 4,744 Mini Challenge text documents into Jigsaw and used the Illinois Named Entity Tagger (LBJ) within the system to identify people, organizations, locations etc., in the documents. This process resulted in a huge number of entities (e.g. 21,866 people and 19,184 organizations). To make the exploratory analysis more manageable, we discarded entities that occurred in at most one or two documents, assuming that these entities were either false positives, or they likely were not important to the central plot. This operation decreased the number of person entities by almost a factor of 10, resulting in a much more manageable set.

We further "cleaned" the set of entities by manually removing or correcting many incorrectly identified entities, by adding entities that were missed in the identification process, and by merging different entities (strings) under one alias if we believed that they referred to the same real world entity. We spent about a week on this process, using Jigsaw's Document and List Views to assist.

Assuming that only a small subset of the entire document collection would be relevant to the central plot, we tried to use Jigsaw's computational analysis capabilities to find a viable lead. We ran Jigsaw's text analyses on the documents to compute summary sentences for each document, similarities across documents, and clusters of related documents.

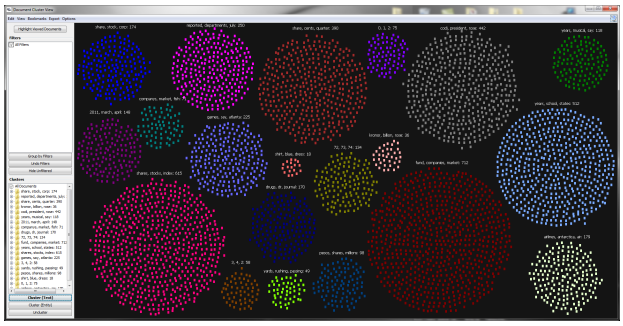

Figure 1: Document Cluster View showing sets of related documents.

Unfortunately, the computed clusters (Figure 1) did not directly lead us to a smaller subset of interesting documents to investigate in-depth. We thus used the visualizations in Jigsaw to help explore and read potentially relevant documents. In particular, we used the List View (Figure 2) and the Document View (Figure 3) extensively. The List View allows the analyst to sort different categories of entities alphabetically or by the number of documents in which they appear. Selecting an entity in the view highlights other entities that appear in the same documents. By tracing seemingly interesting entities and their connections, we loaded the relevant documents into Jigsaw's Document View for more detailed analysis. Reading the actual documents and understanding the events was very important, and we did come across documents hinting towards potential terrorism-related and criminal plots.

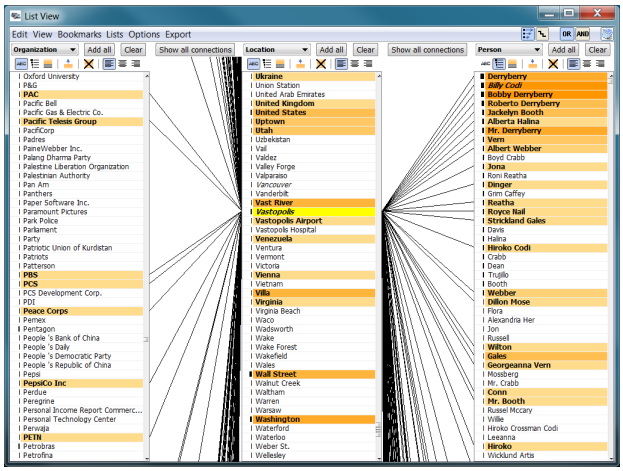

Figure 2: List View showing entities connected to Vastopolis.

As we explored and read more documents, we began to notice that the majority of the documents in the collection were modified versions of actual news articles from the 1990's with key entity

*e-mail: ebraun65@lakers.mercyhurst.edu
†e-mail: Carsten.Goerg@ucdenver.edu
‡e-mail: zliu6@gatech.edu
§e-mail: stasko@cc.gatech.edu

names changed. Ultimately, we believed that these documents were not related to the embedded challenge plot. Other interesting and potentially relevant documents, however, were typically shorter and seemed to center around recent activities at a fictitious city called Vastopolis.

At this point, we decided to examine all of the documents in the collection to find the ones fitting this shorter, Vastopolis-related pattern. Using the Document Grid View in Jigsaw, we sorted all the documents by their length, but too many other short non-relevant documents existed for this method to be useful. Consequently, we loaded all the documents into Jigsaw's Document View (Figure 3) and performed a rapid triage-style pass through them, looking for documents meeting this suspicious profile. The Document View provides a user interface for doing this task very quickly and we were able to go through the entire collection in about three to four hours. This process resulted in approximately 60 "suspicious" documents to examine in more detail.

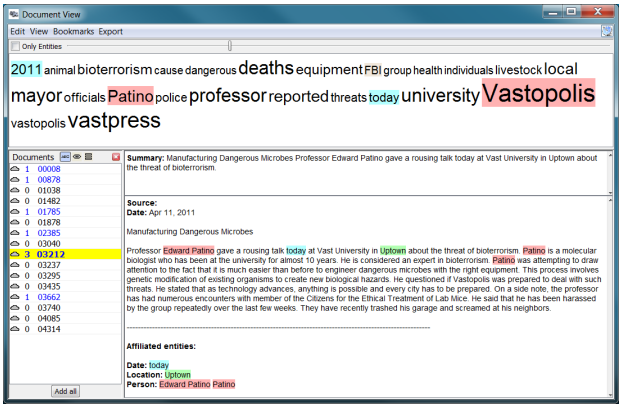

Figure 3: Document View showing a number of the documents crucial to the plot.

We loaded these documents into Jigsaw's Calendar View (Figure 4) to see their chronology and we read the documents more closely in this temporal order. In parallel, we used the Tablet window in Jigsaw as an evidence marshalling tool to organize significant events in these documents into timelines, to log and connect related people and organizations, and to gradually build up hypotheses about the crimes and terrorist activities. A more coherent story of the events within the documents then began to emerge.

3 Reflections and Future Directions

For this Mini Challenge, Jigsaw was most useful to us as aid for rapid triage on the documents, helping to determine their potential relevance to the plot. It provided multiple analytical perspectives on the documents' text.

Although we managed to winnow the large document collection down to a smaller set of documents which had the central plot embedded, it should be evident from the previous section that this process was non-trivial. Considerable manual work was needed to clean extracted entities, to read documents to discover patterns, and to identify a subset of relevant documents. Jigsaw's visualizations are most useful for following potential leads and identifying possible connections between important people and organizations. This challenge, however, had no good leads for starting. Jigsaw's newer computational analyses, specifically document clustering, was designed to help with initial exploration, but was not so helpful here. The clustering was not sophisticated enough to group small sets of meaningful, related documents nor could we steer or direct the clustering in useful directions. We envision that a more semantic analysis or clustering could be very beneficial instead. For example, one

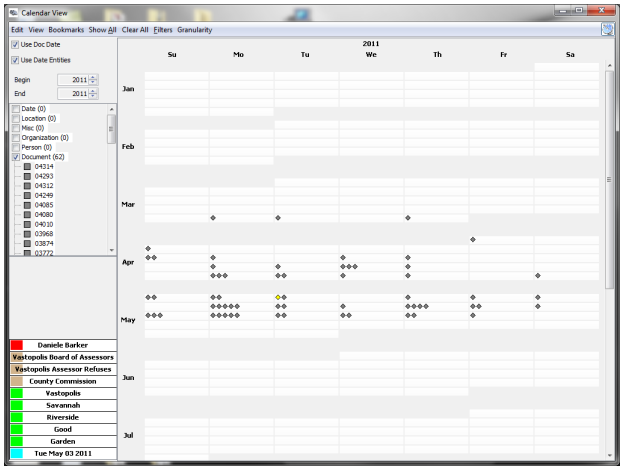

Figure 4: Calendar View showing the suspicious documents as gray diamonds, organized by the date of document creation.

should be able to issue queries such as "Show me all the documents containing activities about food contamination." The computational analysis required should go beyond search based on lexical strings and incorporate information retrieval based on semantically related keywords.

When submitting our final solution we were not sure if we had identified the correct plot because the narrative we had constructed did not seem complete. It turned out that we had missed a key document that helped explain certain events that we did not manage to put together. Interestingly, this particular document was actually included in the initial set of 60 "suspicious" documents we idenitifed. During the evidence marshalling stage, however, we failed to notice key information within it and relate it to the plot that we had constructed from other documents. The ability to suggest connections that may not initially be obvious would be another valuable computational analysis capability within Jigsaw.

4 Conclusion

A primary task in the VAST Challenge is to "connect the dots" or "put the pieces together." Jigsaw has been useful to us in such a role. In order to complete this task, however, the dots or the pieces must first be identified. That process of identifying good potential leads or angles to explore in more depth is still quite challenging, requiring extensive manual effort and browsing. Potentially valuable future research directions should include the ability to provide semantic, contextual search of documents and techniques to rectify potential human cognitive limitations.

Acknowledgements

This work is supported in part by the National Science Foundation under awards CCF-0808863 and IIS-0915788 and the VACCINE Center, a Department of Homeland Security's Center of Excellence in Command, Control and Interoperability.

References

[1] Jigsaw project. http://www.gvu.gatech.edu/ii/jigsaw/.

[2] C. Görg, J. Kihm, J. Choo, Z. Liu, S. Muthiah, H. Park, and J. Stasko. Combining computational analyses and interactive visualization to enhance information retrieval. In *Fourth Workshop on Human-Computer Interaction and Information Retrieval*, 2010.

[3] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118132, 2008.