Data Ingestion and Evidence Marshalling in Jigsaw
VAST 2010 Mini Challenge 1 Award: Good Support for Data Ingest

Zhicheng Liu*, Carsten Görg*, Jaeyeon Kihm[§], Hanseung Lee[§], Jaegul Choo*, Haesun Park*, John Stasko*

Georgia Institute of Technology

ABSTRACT

This article describes the sense-making process we applied to solve the VAST 2010 Mini Challenge 1 using the visual analytics system Jigsaw. We focus on Jigsaw's data ingest and evidence marshalling features and discuss how they are beneficial for a holistic sense-making experience.

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 used the Jigsaw system [3] to solve the VAST 2010 Mini Challenge 1. Jigsaw has been under continuous development since 2006, and we already used it for our award-winning entry in the VAST 2007 contest [2]. Since the original Jigsaw publication [4], we have enhanced the system with additional features and views. This article focuses on the data ingest capability and the Tablet component for evidence marshalling, and we discuss how we used these features while working on the Mini Challenge.

2 DATA INGESTION

Jigsaw supports an XML based data file format (.jig file) that describes the attributes of documents and the entities within them. Since the documents in the Mini Challenge data set were in the Microsoft Word format, we first saved them as plain text files. Next, we simply used the Emacs text editor and its macro capability to convert each of these text files into a .jig file with the appropriate XML tags for the documents' ID, date, source, body text, and so on. The macros helped us to repeatedly search for a text string and then output the relevant surrounding XML tag.

After generating proper Jigsaw data files, we imported them into the system and identified entities via Jigsaw's embedded entity identification (EI) capabilities. (Jigsaw includes a few different open source EI packages for analysis, and we used one from Dan Roth's group at the Univ. of Illinois in the Challenge.) We identified people, places, and organizations using the Illinois statistical entity identification package and we identified dates, phone numbers, and URLs using a pattern-based identification process that we created.

2.1 Entity Cleanup

Since Jigsaw focuses on analyzing and visualizing entities and their relationships, it is important that entities are identified

---

* e-mail: {zcliu, goerg, joyfull, hpark, stasko}@cc.gatech.edu
§ e-mail: {jkihm3, hanseung.lee}@gatech.edu

correctly. Through direct manipulation in its Document View, Jigsaw supports manually adding entities that were missed by the entity identification process, changing the type of, or altogether deleting wrongly identified entities. Applying these entity cleanup features allowed us to work with a clean and consistent set of entities and their connections.

Additionally, the same logical entity may be identified by different strings in different documents. Jigsaw provides an operation that allows analysts to merge different entities (strings) under one alias (see Figure 1). After assigning a primary identifier to the merged entities, that identifier represents all the initially different entities in Jigsaw's visualizations. Jigsaw uses italics to indicate entities with aliases.

We performed aliasing early in our sense-making process to clean up obvious misspellings and name variations. When it was less certain that two similar names actually referred to the same entity, we read the documents more carefully and created aliases whenever necessary during the exploration.
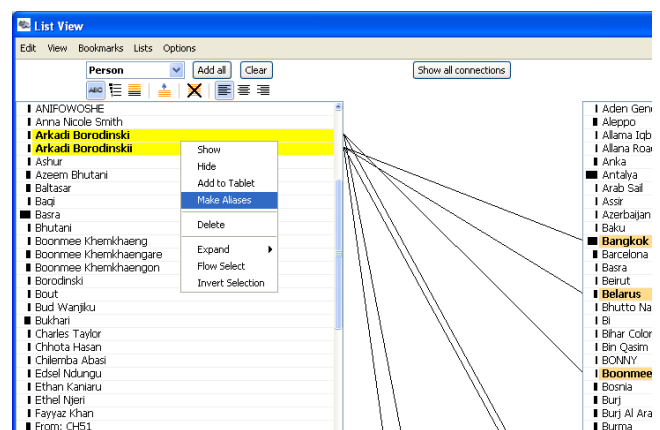

Figure 1. Selected entities can be merged via the "Make Aliases" command in Jigsaw.

2.2 Computational Analyses

To better assist analysts in browsing and understanding text documents in a more structured manner, we have coupled the interactive visualizations in Jigsaw with automated computational analysis capabilities such as analyses of document similarity, document sentiment, document clusters, and document summarization [1]. After importing the documents and identifying and cleaning the entities, we performed these automated analyses on the Challenge data. The computational analyses, in particular the document clustering, proved to be very useful in guiding the process of reading and making sense of the documents.

We started our exploration by examining the high-frequency entities and their connections in the List View and the Graph View. This enabled us to directly focus our attention on important people and places in the data set. Showing document clusters grouped by topics in the Document Cluster View helped us to

keep track of the different threads of the stories embedded in the data set; in addition, this view indicated which documents we had already read and explored.

3 EVIDENCE MARSHALLING

Since the number of documents in the data set was relatively small (just over a hundred compared to more than a thousand in the 2007 contest), we were able to quickly familiarize ourselves with most of the documents using the views in Jigsaw. We soon realized that unlike in the 2007 contest data where only a small subset of the documents was relevant to the final solution, most documents in this Mini Challenge seemed to contribute to a larger story. To take notes about interesting entities and events, formulate hypotheses and organize findings, we used Jigsaw's new Tablet functionality, an environment that serves as an evidence marshalling and sense-making tool.

3.1 The Tablet Interface

The Tablet adopts a minimalistic design, intending to offer greatest flexibility for visual thinking and sense-making. Entities in Jigsaw's views can be directly added to the Tablet via popup menu commands. The added entities retain their original color-coding according to their types. Analysts also can create their own items representing customized entities or events. Any two items or entities can be linked and the links can be labeled. Additional information about the items can be represented as post-it-notes (on a yellow background). Analysts can also create timelines and link entities or items to specific points on the timeline. All the visual items in the Tablet can be freely moved around and repositioned. Figure 2 shows multiple timelines we created for important people in the data set. Figure 3 shows a social network that we built during our investigation using some of the Tablet's functionality.
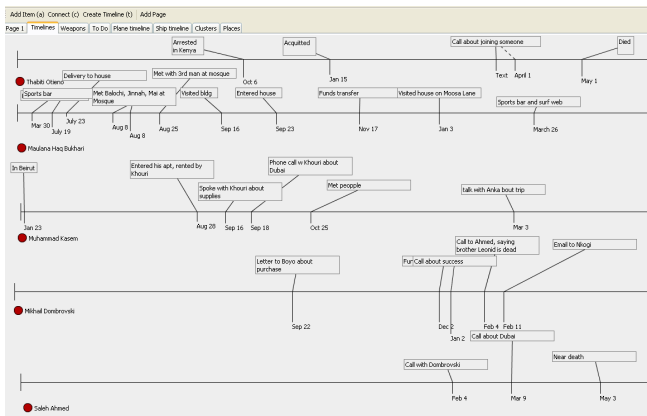

Figure 2. Multiple timelines in the Tablet in Jigsaw. Each timeline records the activities of a specific person.

3.2 Sense-making with the Tablet

We created multiple pages in the Tablet, each represented by a tab. The pages organized our findings and thinking processes in terms of different perspectives and themes including social networks, timelines, specific topics such as weapon and fund transfers, and geographically connected people and events. We iteratively modified and refined the hypotheses and findings represented in the Tablet as we read the documents in greater depth and discovered connections between interesting entities.

Jigsaw also provides the functionality to save the entire workspace including the pages in the Tablet and the states of
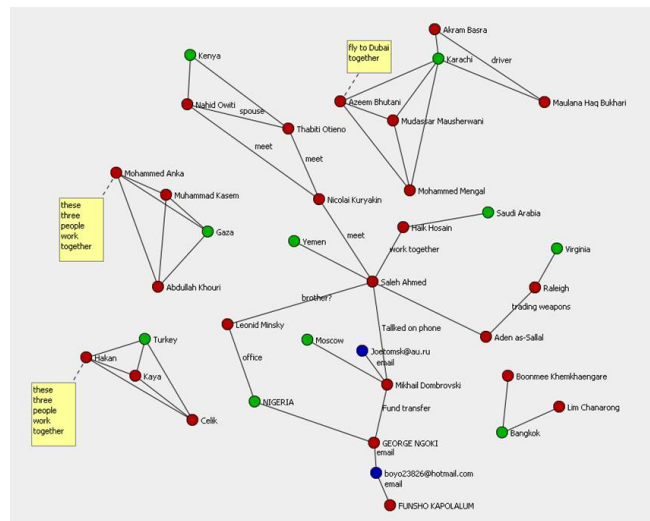

Figure 3. A social network in the Tablet in Jigsaw. The red nodes represent people and the green nodes represent places. A person is connected to a place if he/she is based in that location. Two persons can be connected in various ways, and the semantics of the connection is annotated as the label on the link. Additional information about the entities can be added in yellow notes.

every view that have been created and modified. Since our investigation spanned across multiple sessions, this feature was especially useful.

4 CONCLUSION

The Jigsaw system proved to be very useful for investigating the events here in Mini Challenge 1. Unlike the text document-focused VAST Contest of 2007 where one needed to find the "needle in the haystack", here many different documents contributed to a complex, multifaceted storyline. Jigsaw's flexible document import and entity identification capabilities coupled with the new Tablet sense-making environment were particularly helpful in our investigation. For further details on Jigsaw, we refer the readers to the Jigsaw website [3] where several videos showing different aspects of the system are available.

5 ACKNOWLEDGMENTS

REFERENCES

[1] 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.

[2] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. Stasko. Jigsaw meets Blue Iguanodon - The VAST 2007 Contest. In *IEEE Symposium on Visual Analytics Science and Technology*, pp. 235-236, 2007.

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

[4] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In *Information Visualization*, 7(2):118-132, 2008.