Sensemaking across Text Documents: Human-Centered, Visual Exploration with Jigsaw

John Stasko, Carsten Görg, Zhicheng Liu
School of Interactive Computing & GVU Center
Georgia Institute of Technology
Atlanta, GA 30332
{stasko,goerg,zcliu}@cc.gatech.edu

ABSTRACT

In this article we describe our research on providing techniques and tools for a particular type of sensemaking — finding embedded narratives or plots across large collections of relatively brief text documents. Our approach is decidedly human-centered and we have created the Jigsaw system as an embodiment of these principles. Jigsaw provides multiple visualizations depicting the documents, the entities within them, and connections between the two. Analysts interactively explore the entities and documents as a way to guide document selection for reading, which ultimately is necessary to synthesize the embedded plot. This article describes the challenges in this particular style of sensemaking as well as the ways in which Jigsaw is attempting to overcome those challenges.

Author Keywords

Sensemaking, investigative analysis, information foraging, information visualization, multiple views.

ACM Classification Keywords

H5.2. [Information Systems]: Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

The term "sensemaking" is a relatively general concept and can be interpreted in a number of different ways. Russell et al. describe sensemaking via Learning Loop Complex theory, made up of the components 1) search for representations, 2) instantiate representations, and 3) shift representations [4]. Particularly germane to the domain of our work, Pirolli and Card describe a notional model of the sensemaking loop in intelligence analysis involving a progression of activities with many back loops and two key subcomponents, foraging and sensemaking [3].

In the context of our research, sensemaking involves people working with a large collection of data and seeking to identify and understand different stories or narratives that are embedded throughout the collection. Let us clarify this notion by further describing the domain of the data and the particular activities that people are performing with/on it.

In our case, the data being examined is textual documents, each one in a loose narrative form and typically several paragraphs long. News stories or case reports are good examples of this type of document. Longer documents like multi-page articles or even books could be analyzed using our tools, but the process likely would not be as effective and is not really our focus. An example of the kind of text document our work applies to is shown below (taken from VAST 2007 Contest [6]):

Report 20040510-4_16: May 14, 2004

```
VANCOUVER, British Columbia - A Canadian immi-
gration panel is considering whether accused
environmental saboteur Tre Arrow can apply for
refugee status in Canada.

Arrow, 30, who is wanted for fire bombing log-
ging and cement trucks in Oregon, asked the
Canadian authorities to remain in Canada as a
political refugee at a hearing in Vancouver on
Tuesday.

A key issue will be whether Arrow is affiliated
with a terrorist group, which would immediately
disqualify him from receiving refugee status in
Canada, authorities said.

The Immigration and Refugee Board is scheduled
to decide by May 31 whether Arrow is affiliated
with the Earth Liberation Front, a group the
FBI considers a terrorist organization respon-
sible for scores of attacks on property over
the past dozen years.
```

Each document in the collection includes a number of facts and refers to people, places, objects, dates, actions, and so on. Individual documents are relatively focused, however, and typically describe a particular event or occurrence. So, one may think of each document as presenting a small bit of evidence or a nugget of information.

Thus, sensemaking in our context is the process of connecting a series of individual bits of evidence to construct a larger, broader story or narrative. Sensemaking is about understanding how the individual events and entities referred to in the documents relate to each other and, when composed together, reveal a larger plot.

For instance, consider police investigators who have a large collection of case reports. Suppose that one report describes a crime in which a blue Ford car was seen fleeing from the scene. Investigators may wonder if other case reports mention a blue Ford, and if so, what related activities are mentioned in those reports. Suppose that another such report exists and it describes a known criminal who used a blue Ford in a recent theft. Investigators might extend their search to examine activities of this individual, places he has been, known accomplices, and so on, keeping a careful watch for unexpected, serendipitous connections and coincidences that arise. Ultimately, investigators are likely seeking to uncover larger criminal plans and threats in order to either thwart planned crimes before they occur or to apprehend the criminals being sought.

This kind of scenario is precisely the focus of our research. More specifically, we seek to

- Better understand the investigative processes conducted in these types of scenarios
- Build technologies and systems that will aid the investigators in their work
- Assess whether our technologies and systems are effectively assisting the analysts

Clearly, this is not the only kind of sensemaking activity that exists, but we have included a description here to help the reader understand the focus of our research.

OUR APPROACH

A variety of approaches to helping people in sensemaking scenarios like the ones we describe do exist. Some use automated techniques and tools that examine a document collection without human intervention and report on discovered plots or narratives. These approaches typically use techniques and algorithms from the fields of artificial intelligence, data mining, and machine learning.

Our approach is quite different, instead involving human-centered investigations where we provide human analysts with computational tools to assist them while conducting investigations. Our tools seek to enable the powerful perceptual capabilities of people and bring those capabilities to bear throughout the sensemaking process. We firmly believe that human analysts harbor tremendous investigative skills, but the masses of data and documents typically present today can overwhelm the analysts' investigative capabilities. Thus, we provide visualization tools that transform the data (text documents in our case) into visual representations that can more easily be surveyed, scanned, examined, reviewed, and studied.

In order to facilitate the powerful exploratory, investigative skills of people, our tools are highly interactive and flexible. We seek to help analysts browse the document collection rapidly and to more deeply explore "interesting" avenues of investigation. Analysts must uncover whether the agents and events in question relate to potential plots being developed. Our approach also hinges upon multiple visual representations of the documents and entities within them. Any one visualization simply may not provide the right perspective onto the data to allow an analyst to perceive an important connection. By supplying multiple visual representations of the data, each providing a view onto some important characteristic or characteristics, we are more likely to help the analyst discover the unknown connections that weave a larger narrative together.

JIGSAW

Jigsaw [5] is a system for helping analysts with the kinds of investigative scenarios discussed above. It is a multi-view system, including a number of different visualizations of the documents in the collection and the entities (people, places, dates, organizations, etc.) within those documents. Accompanying the visualizations is a textual search query interface so that particular entities can be examined directly. When used in this way, Jigsaw acts much like a search engine that simply displays results through visualizations rather than text lists.

Jigsaw is much more than a search engine with visual results, however. Once views show documents and their entities, users can explore the collection by interactions with those objects. For instance, new entities can be displayed and explored by simple user interface operations in the views that expand the context of entities and reports. In fact, far more entities and documents are initially displayed via user interaction than by textual search queries. Search queries often serve to jump-start an exploration, but view interaction yields richer representations and exploration.

Most of the views in Jigsaw illustrate connections between documents and entities or between entities and other entities. Jigsaw uses a simple model of "connection" — an entity is connected to a report if it appears in that report (and vice versa) and two entities are connected if they appear in at least one report together. Entities that appear in more than one report together are considered to be more strongly connected with the connection value dependent on the total number of documents of co-occurrence. This simple model of connection is easy to implement, is easy for people to understand, and we have found it to be powerful for helping exploration of document collection.

The views in Jigsaw are linked so that actions in one view propagate to the other views whose visual state updates to reflect that action. For example, the most common operation on a view is to mouse-click on an entity or report which selects that object, and then the rendering of other objects in the view updates to reflect their relation to the selected object. In Jigsaw this action is propagated to other views which then also select that same object and update their displays appropriately. Another common operation is to "expand" an entity or report which typically displays a new set of entities and reports that are connected to this object. This operation is usually invoked by a double-click on an object or a click-activated menu.

The person using Jigsaw also can decouple a view from event listening so that its visual state only changes via explicit operations in that view. We have found this capability to be very useful when an analysis process yields a view configuration that is particularly enlightening and the user wants to keep the view as-is during subsequent exploration.

Jigsaw's views include list, graph and scatterplot-based representations of object connections, an overview-style report view showing all documents, a calendar view for examining temporal patterns, and a fundamental document view showing report text with highlighted entities. Below we describe some of these views in more detail.

In Figure 1 the Document View shows the example report mentioned in the introduction. To facilitate fast scanning of text documents, entities are highlighted according to their type.

Figure 2 shows the connections of "saboteur Tre Arrow" along with those of the "Immigration and Refugee Board" in the List View. For each of the lists an entity type can be selected and the lists can be sorted either by frequency or alphabetically. The bars on the right border of each list entry display the frequency across the whole document collection of the entry. Connections between entities are visualized in two different ways: items connected to a selected entity are marked in a shade of orange (the stronger the connection, the darker the shade of orange) and in neighboring lists connected entities are additionally joined by lines. Thus it is possible to see which entities are connected in case multiple items are selected.

Figure 3 shows the Graph View. The larger white circles represent reports, the smaller colored circles represent entities (colored according to their type). By expanding and collapsing nodes to either show or hide their connected entities or reports respectively, the analyst can explore the network step by step.

More details about Jigsaw can be found in [5] and at the project website:
`http://www.cc.gatech.edu/gvu/ii/jigsaw`.

We want to stress that Jigsaw does not seek to depict the main themes running throughout the document collection or the semantically meaningful concepts within it (although these are worthy goals for future work). Presently, Jigsaw acts much like a visual index onto the document collection, helping to provide fast, contextualized access to the individual entities and reports that an analyst is studying. Fundamentally, an analyst must read reports to understand the events occurring within them. As document collections grow larger and larger, finding the most fruitful documents to read becomes more challenging. Furthermore, traditional search technology is not as useful in this situation because the plots/ stories discovered often involve unexpected and serendipitous connections between entities which are best found following a trail of linked evidence.

SENSEMAKING ACTIVITIES

In terms of the sensemaking model proposed by Pirolli and Card [3], we feel that Jigsaw can help analysts with both the information foraging and sensemaking loops, but its utility is much stronger for foraging right now. As discussed above, Jigsaw helps people find small collections of potentially important documents to read and study, a fundamental activity in information foraging.

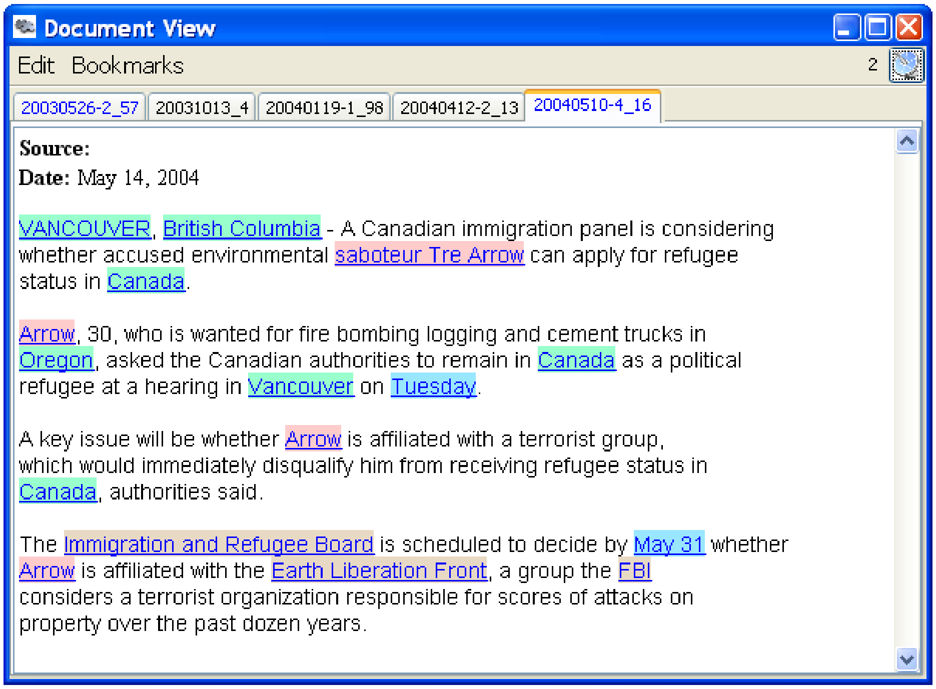
To support the evidence marshalling and sensemaking process, Jigsaw provides a special view called the Shoebox. The Shoebox helps the analyst to collect and organize items or information of interest that were revealed while exploring the document collection. Figure 4 shows an example of the Shoebox view.

The analyst can add items to the Shoebox from every view — they appear first in the 'inbox-area' on the left side of the Shoebox. Items added at the same time are grouped together and sorted by type. The Shoebox offers multiple ways to organize the items in the inbox and to join them to build sensemaking artifacts:

- Combining items to sentences (e.g. "Kim visited New York last spring") by adding comments and snapping entities together
- Grouping items according to a topic
- Forming hypotheses and using items as supporting or contradicting evidence
- Linking hypotheses, groups, sentences, and items.

These sensemaking artifacts support the analyst's thinking process in a visual way and reduce the amount of necessary text as much as possible. This is important since the analyst may already be overwhelmed with text documents. During an informal evaluation of Jigsaw with an analyst, a reoccurring statement was: "I don't want to read it, I want to see it."

While designing the marshalling support for Jigsaw, we envisioned two different approaches for collecting evidence: either augmenting the existing (data) views, or collecting evidence in a separate Shoebox view. Incorporating the marshalling process into the existing views would have the advantage of placing the information along with necessary comments right at the spot where it was discovered rather than duplicating information at another location. The disadvantage would be that evidence would be scattered across multiple views what would make it difficult to keep track of the collected information. Therefore, we decided to collect evidence in a separate Shoebox view. To address the problem of duplicating information, we added a hyperlink function to the Shoebox. This allows the analyst to connect views via links to bookmarks as evidence to the Shoebox.



Source:
Date: May 14, 2004

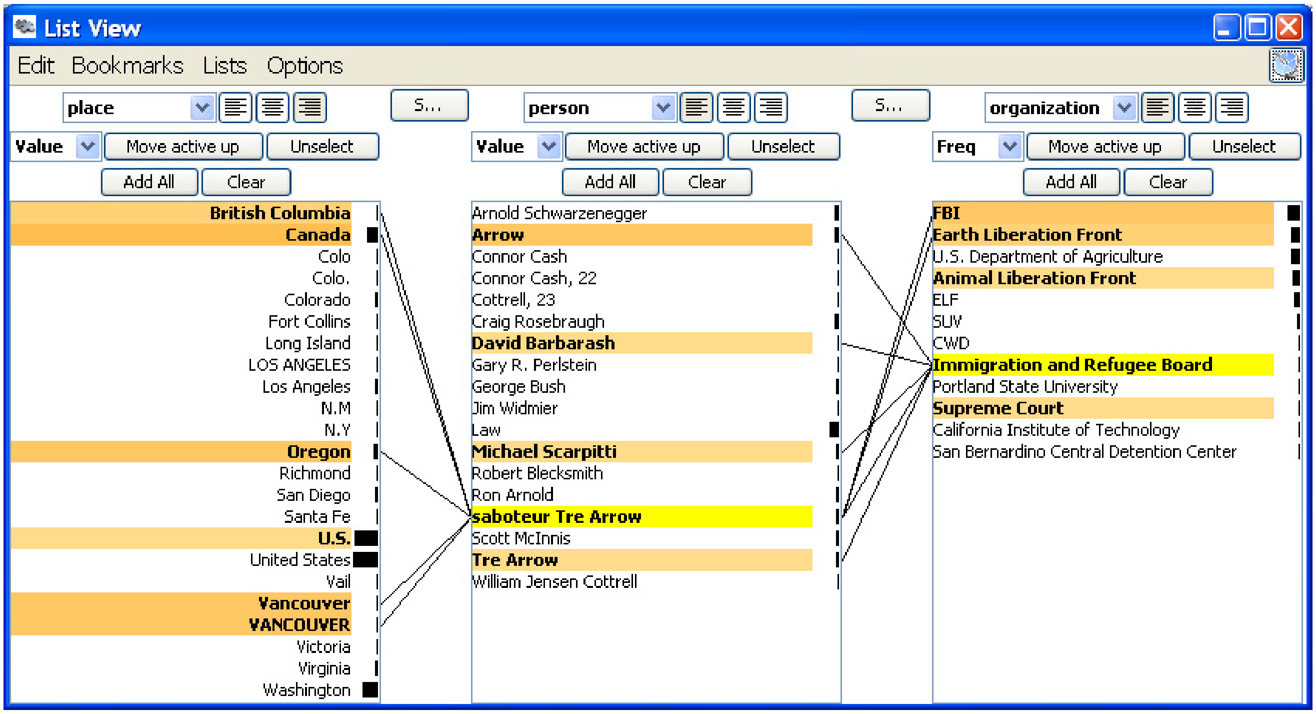
VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre Arrow can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

Figure 1. The Document View showing an example report.



British Columbia
Canada
Colo
Colo.
Colorado
Fort Collins
Long Island
LOS ANGELES
Los Angeles
N.M
N.Y
Oregon
Richmond
San Diego
Santa Fe
U.S.
United States
Vail
Vancouver
VANCOUVER
Victoria
Virginia
Washington

Arnold Schwarzenegger
Arrow
Connor Cash
Connor Cash, 22
Cottrell, 23
Craig Rosebraugh
David Barbarash
Gary R. Perlstein
George Bush
Jim Widmier
Law
Michael Scarpitti
Robert Blecksmith
Ron Arnold
saboteur Tre Arrow
Scott McInnis
Tre Arrow
William Jensen Cottrell

FBI
Earth Liberation Front
U.S. Department of Agriculture
Animal Liberation Front
ELF
SUV
CWD
Immigration and Refugee Board
Portland State University
Supreme Court
California Institute of Technology
San Bernardino Central Detention Center

Figure 2. The List View showing connections of “saboteur Tre Arrow” and the “Immigration and Refugee Board”.

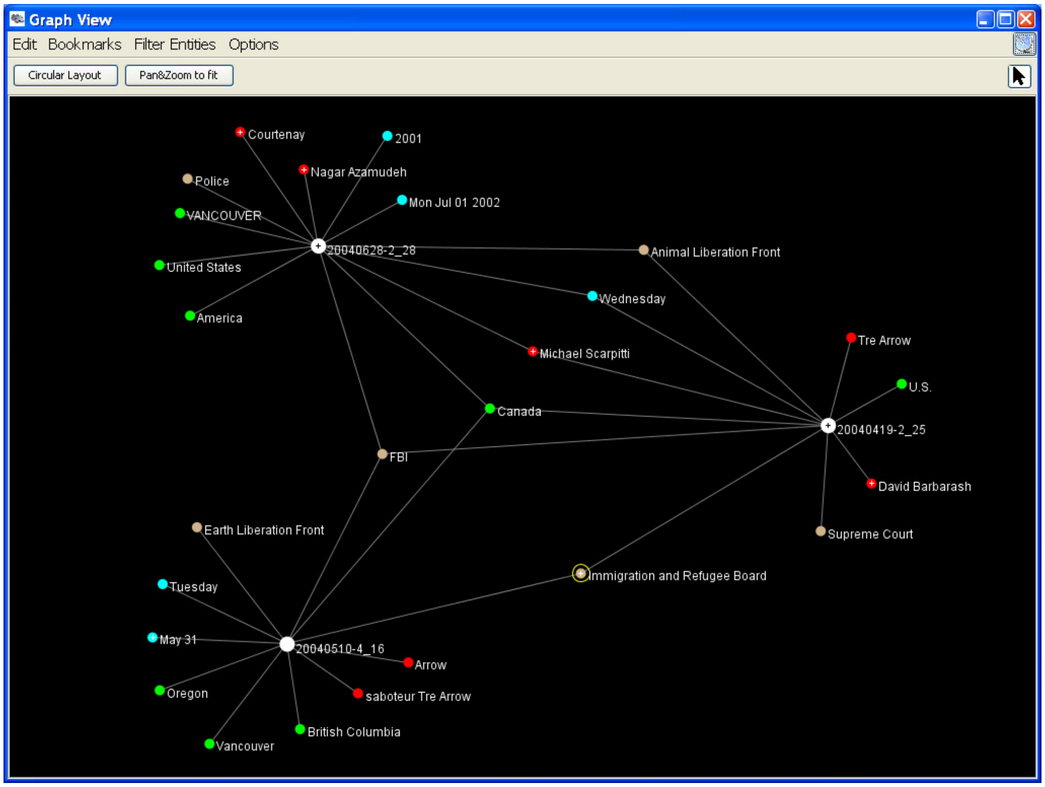


Figure 3. The Graph View after exploring connections of the “Immigration and Refugee Board”.

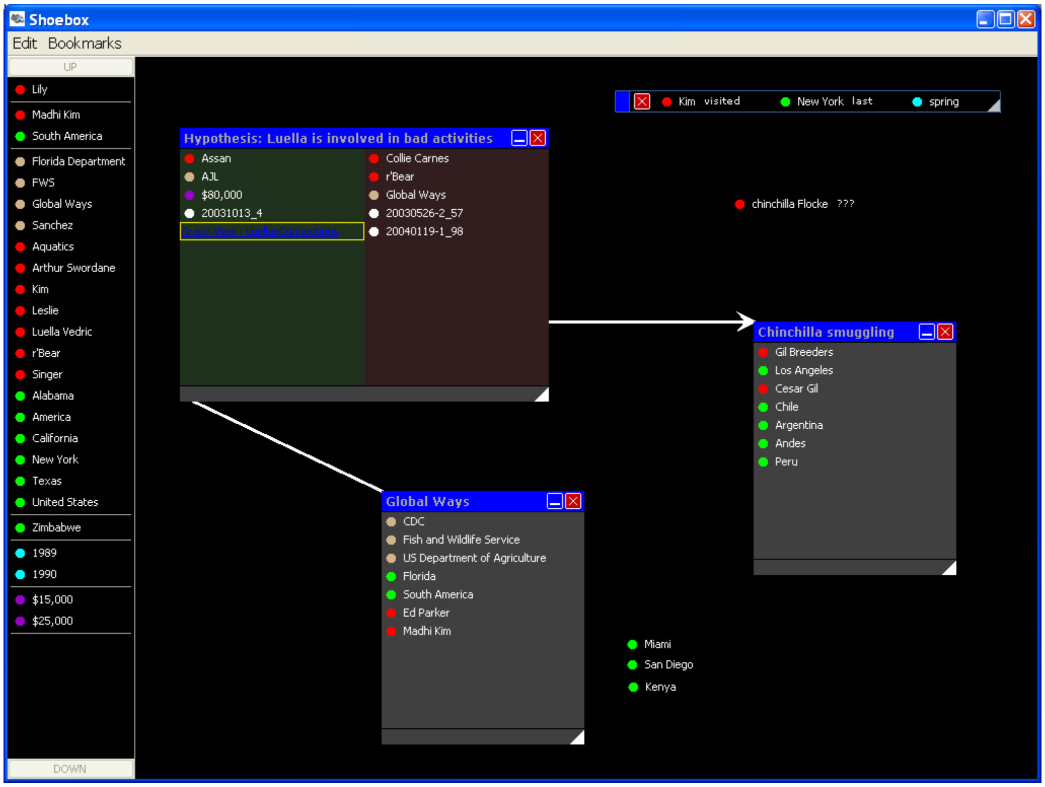


Figure 4. The Shoebox View showing the group, hypothesis, sentence, and link feature.

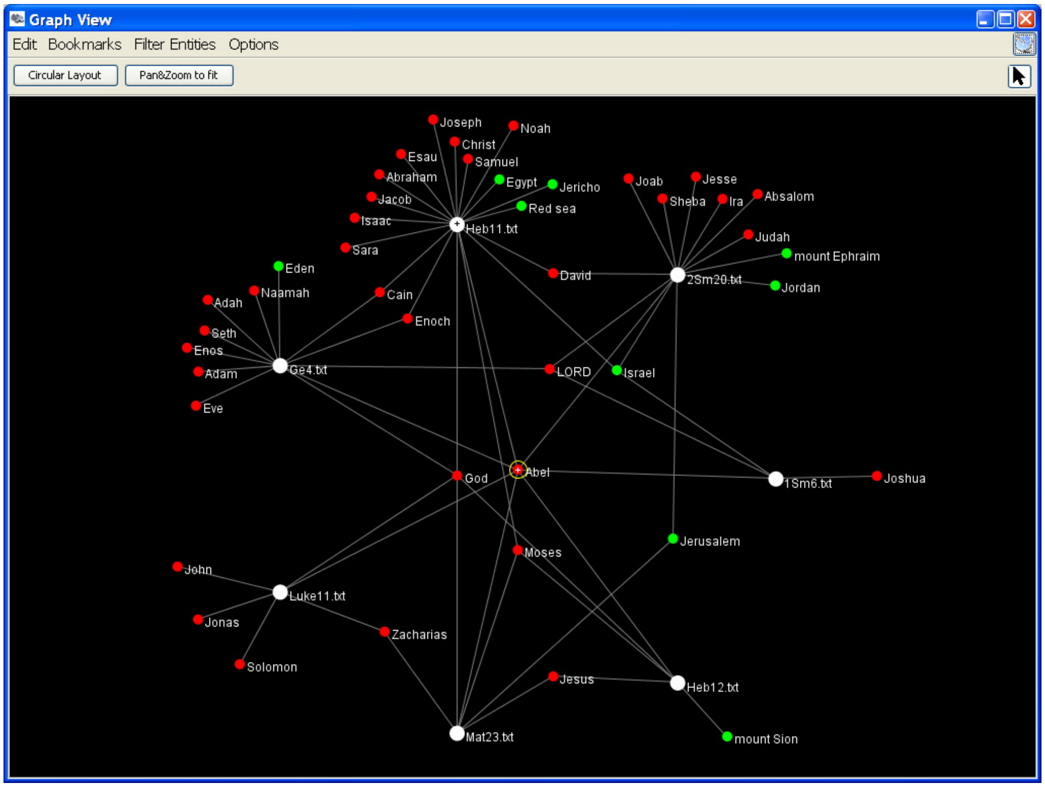


Figure 5. The Graph View showing "Abel's" social network in the Bible.

OTHER DOMAINS FOR POTENTIAL STUDY

While the primary focus of our work has been on investigative analysis for domains such as law enforcement and intelligence, one can imagine Jigsaw being used for analysis in other areas involving unstructured text data as well. We speculate that people other than trained analysts may benefit from its sensemaking capabilities as well.

One area that seems well-matched to Jigsaw's capabilities is academic research involving journal and conference papers. People new to a discipline may want to learn about the important researchers and papers in the area. Existing tools like PaperLens [2] address such scenarios already but trails of connected articles are not so easy to follow. In Jigsaw we can use metadata of papers such as authors, institutions, categories, etc., as entity types and then visualize and explore connections between these entities. A second scenario is to define entity types relevant for a specific research area. For example, in biology papers about research on genes, the different genes could map to a set of entities and biologists could benefit by learning about connections of genes across all relevant publications.

Another possibility would be to use Jigsaw to explore online articles such as web news reports and blogs on topics like political debates. We could imagine a web-crawling process that produces a large number of loosely related documents about a particular topic or person, and then an analyst (or someone simply seeking to explore and learn) using Jigsaw to find connections between entities among the documents which would help filter the collection to a smaller number of documents to actually read.

We already have used Jigsaw to explore connections between people, places, and organizations in the Bible. We built a set of documents by dividing the Bible into its chapters, and then we extracted the entities and used that data set as input for Jigsaw. The Graph View in Figure 5 shows the social network of "Abel" along with places mentioned in chapters in which he is present.

CONCLUSION

In this article we have described the Jigsaw system that has been designed to help investigative analysts find embedded plots or stories in large document collections. While this type of exploration is only a specific style of sensemaking activity, we feel that it can be a potentially important one to certain key organizations.

Jigsaw provides multiple visualizations of reports and the entities within them, as well as the connections that exist between entities and/or reports. Jigsaw provides a decidedly human-centered approach to sensemaking by allowing people to interact with the views and explore possible new avenues of examination. Presently, the system provides more information foraging utility than schema/hypothesis generation utility, but we are exploring how these latter capabilities could be added to the system too. This article provided a look at our initial attempts.

Evaluation of Jigsaw is an ongoing activity as well. Presently, we are conducting experiments to examine whether people can use individual views to answer the kinds of analytic queries common to the domains we study (e.g., Do these two people share any common acquaintances? Has this person ever been to that city?) Our next evaluation phase will involve more holistic study of the system to see if it does benefit analysis as compared with investigations using more common aids such as search engines and authoring/organizational tools. To do that, an analysis activity may have to be conducted over days rather than minutes. Finally, the utility of Jigsaw was illustrated at least informally by our use of the system to win the university component of the 2007 IEEE VAST Symposium Contest [1].

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation via Award IIS-0414667 and the National Visualization and Analytics Center (NVAC^{TM}), a U.S. Department of Homeland Security Program, under the auspices of the Southeast Regional Visualization and Analytics Center. Carsten Görg was supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD).

REFERENCES

1. GÖRG, C., LIU, Z., PAREKH, N., SINGHAL, K., AND STASKO, J. Jigsaw meets Blue Iguanodon - The VAST 2007 Contest. In *IEEE Symposium on Visual Analytics Science and Technology* (2007), pp. 201–202.

2. LEE, B., CZERWINSKI, M., ROBERTSON, G., AND BEDERSON, B. B. Understanding Research Trends in Conferences using PaperLens. In *CHI '05: CHI '05 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2005), ACM, pp. 1969–1972.

3. PIROLLI, P., AND CARD, S. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. In *Proceedings of 1st International Conference on Intelligence Analysis* (2005).

4. RUSSELL, D. M., STEFIK, M. J., PIROLLI, P., AND CARD, S. K. The Cost Structure of Sensemaking. In *CHI '93: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (New York, NY, USA, 1993), ACM, pp. 269–276.

5. STASKO, J., GÖRG, C., LIU, Z., AND SINGHAL, K. Supporting Investigative Analysis through Interactive Visualization. In *IEEE Symposium on Visual Analytics Science and Technology* (2007), pp. 131–138.

6. IEEE VAST 2007 CONTEST. `http://www.cs.umd.edu/hcil/VASTcontest07`, 2007.