Jigsaw: Investigative Analysis on Text Document Collections through Visualization

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

ABSTRACT
This article describes the Jigsaw system for helping investigative analysis across collections of text documents. Jigsaw provides multiple visualizations of the documents and the entities within them to help investigators discern embedded stories and plots. Our early focus within Jigsaw has not been on legal documents and E-discovery, but we feel that the system may have potential in these areas as well. This article illustrates Jigsaw's views and operations using Enron email archives as example documents.

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

INTRODUCTION
We have been developing a system called Jigsaw to help investigative analysts explore and make sense of collections of text documents. In particular, we have designed Jigsaw to help investigators uncover stories, plots, and threats embedded across the documents. While our focus has not been on legal documents or E-discovery, we are curious to explore whether Jigsaw might be useful in these areas as well. This article provides a brief overview of Jigsaw and its capabilities, using a subset of the Enron email archive as an example document collection.

Jigsaw has been developed to help people with sensemaking, exploration, and analysis activities on collections of unstructured, plain text documents, in particular, relatively short documents (approximately 1-10 paragraphs) in loose narrative form. Examples of such documents include police case reports, short news articles, or email notes. While Jigsaw can process longer documents, its utility degrades in these cases (reasons will be illustrated later in the article). The email shown below, taken from the Enron data set, is a good example of the kind of document ideal for Jigsaw.

Email 114844

```
Message-ID: <22094025.1075842958662.
JavaMail.evans@thyme>
Date:  Fri, 1 Sep 2000 00:43:00 -0700 (PDT)
From:  steven.kean@enron.com
To:  jeff.dasovich@enron.com,
susan.mara@enron.com, mona.petrochko@enron.com,
tim.belden@enron.com, mary.hain@enron.com
Subject:
Cc:  paul.kaufman@enron.com,
richard.shapiro@enron.com,
james.steffes@enron.com
Mime-Version:  1.0
Content-Type:  text/plain; charset=us-ascii
Content-Transfer-Encoding:  7bit
Bcc:  paul.kaufman@enron.com,
richard.shapiro@enron.com,
james.steffes@enron.com
X-From:  Steven J Kean
X-To:  Jeff Dasovich, Susan J Mara,
Mona L Petrochko, Tim Belden, Mary Hain
X-cc:  Paul Kaufman, Richard Shapiro,
James D Steffes
X-bcc:
X-Folder:  \Jeff_Dasovich_Dec2000\Notes Folders
\All documents
X-Origin:  DASOVICH-J
X-FileName:  jdasovic.nsf

When we have described the problems and
solutions for California we have focussed on
generation siting and flexibility to hedge.  We
have stayed away from transmission issues on the
assumption that California, with its ISO and PX,
does not suffer from the same discrimination
issues as other parts of the country.  Is this
true?  Does California's system layer in
priorities for utility use of the system -- eg
doesn't PG&E control "path 15"?  Does that
control provide advantageous access to PG&E? Are
there other examples and are there links between
these "preferences" and the current problems in
California?  As we are trying to convert
reliability and pricing concerns into FERC
action these would be helpful arguments to have
available to us.
```

Investigators may seek to connect the individual facts and events described in specific documents into a larger, more coherent thread or story. Putting the pieces together in this way can lead to a better understanding of the broader, more general notions and implications of the document collection.

Jigsaw's particular focus is on illuminating connections between the entities in the documents—the people, organizations, places, and so on. Jigsaw visualizes the documents

and the entities within them in a number of different representations, each one specifically created to communicate some different aspect of the data. For instance, Jigsaw can help to understand social networks of people, connections between people and places, and the evolution of events in time.

Within our research communities, these types of activities are know as sensemaking [3, 4, 5]. Investigative reporters, law enforcement officials, and intelligence analysts all routinely perform these types of activities. Clearly, as the number of documents being examined grows, the sensemaking activities become more challenging.

A variety of approaches to support 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 investigations. 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, we are more likely to help the analyst discover the unknown connections that weave a larger narrative together.

Thus, Jigsaw espouses an *information visualization* [1, 6] approach to investigative and exploratory data analysis. More specifically, when visual approaches like this are combined with computational techniques to manage and filter the extremely large data sets that may be present, the resulting system illustrates *visual analytics* [8] principles.

Clearly, many different types of investigations occur within the legal community. Realistically, we speculate that Jigsaw's utility is limited for typical E-discovery type tasks that may involve millions of documents. Instead, Jigsaw's value likely rises when a document collection has been narrowed down to a few thousand documents and an investigator wants to understand how the people, organizations, and events in those documents interact to reveal the "big picture."

JIGSAW

Jigsaw [7] 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 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 documents. 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 document if it appears in that document (and vice versa) and two entities are connected if they appear in at least one document together. Entities that appear in more than one document 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 document 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 document which typically displays a new set of entities and documents 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 configu-

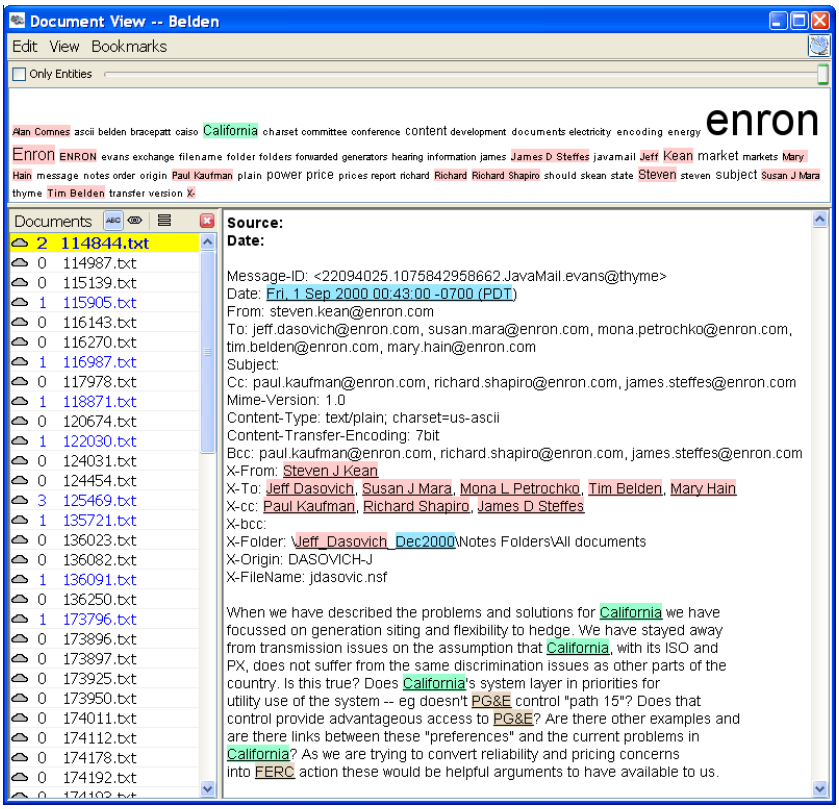

Figure 1. The Document View showing an example report.

ration 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 document cluster view showing all documents, a calendar view for examining temporal patterns, and a fundamental document view showing document text with highlighted entities. Below we describe some of these views in more detail.

In Figure 1 the Document View shows the example document mentioned in the introduction. To facilitate fast scanning of text documents, entities are highlighted according to their type. The tag cloud at the top of the view describes the contents of the marked documents in the document list.

Figure 2 shows the connections of "Tim Belden" in the List View. For each of the lists an entity type can be selected and the lists can be sorted either by frequency, alphabetically, or by the strength of the connection. The bars on the left 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 rectangles represent documents, 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 documents respectively, the analyst can explore the network step by step.

Figure 4 shows the Calendar View. Documents and entities from the data set are displayed in the context of a familiar calendar showing years, months, weeks and days. The small diamond items drawn on a particular day represent documents (colored gray) or entities (colored according to its type) in the context of the date(s) noted in documents in which they appear. When the user moves the mouse pointer over a document-representation diamond drawn in the calendar, all the entities appearing in that document are shown on the left.

We have found that the system is more useful when a set of views can be laid out and easily examined without window flipping and reordering. Due to the large amount of screen real estate required to display its views, Jigsaw ideally should be run on a computer with multiple and/or high-resolution monitors.

More details about Jigsaw can be found in [7] and at the project website:
`http://www.cc.gatech.edu/gvu/ii/jigsaw`.

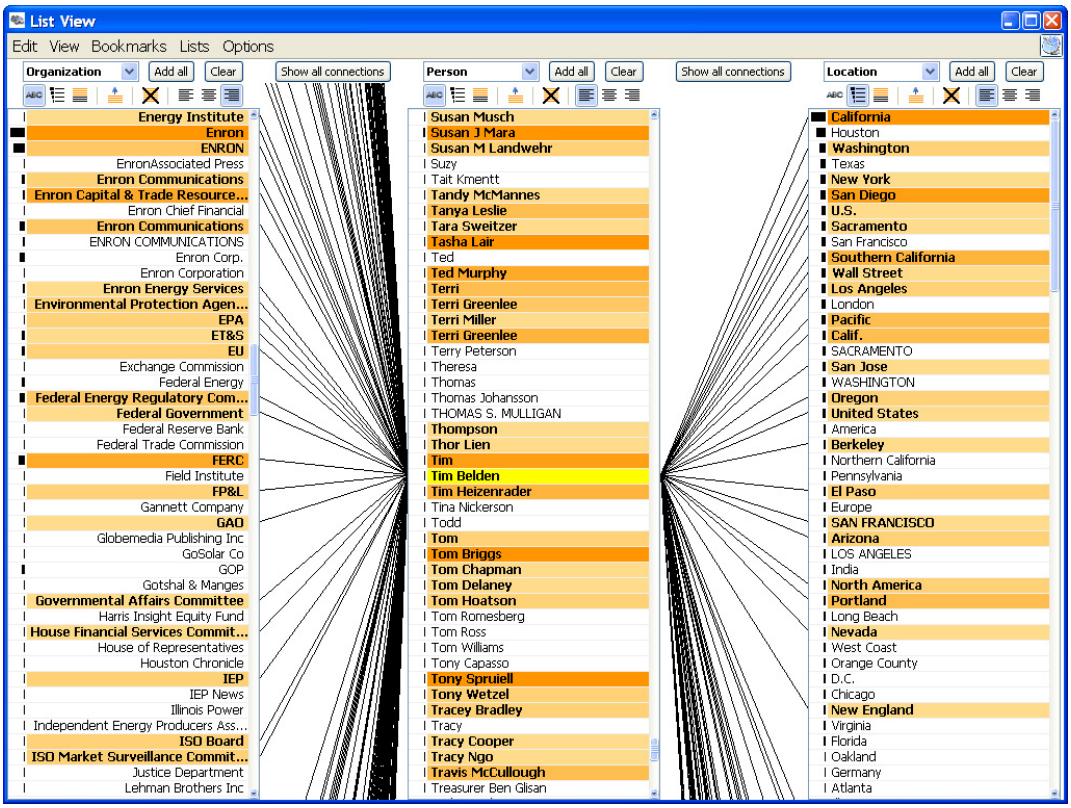

Figure 2. The List View showing connections of "Tim Belden".

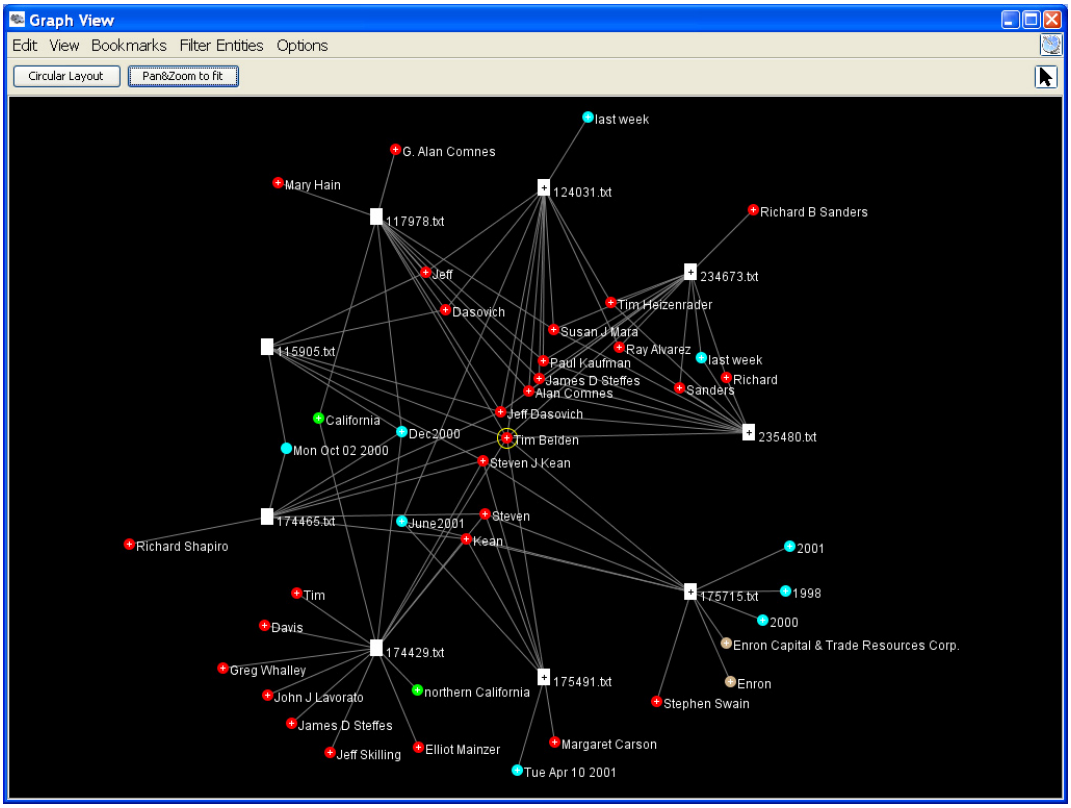

Figure 3. The Graph View after exploring some connections of "Tim Belden".

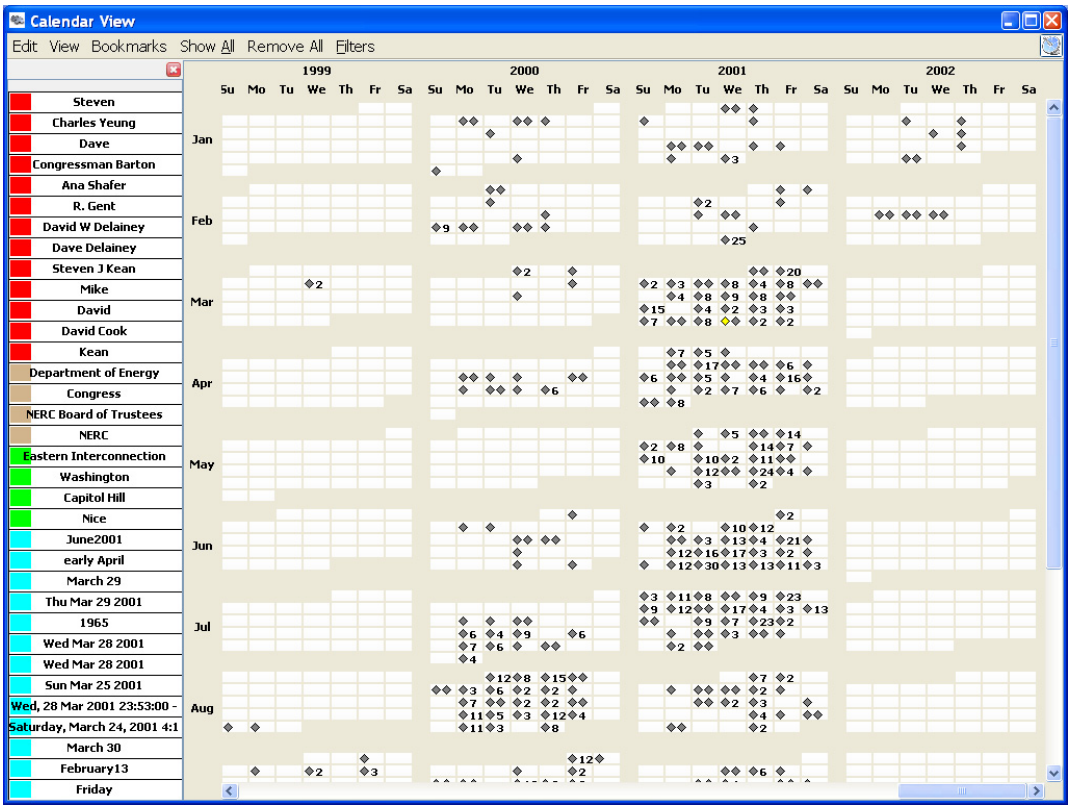

Figure 4. The Calendar View showing when emails were sent.

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 like a visual index onto the document collection, helping to provide fast, contextualized access to the individual entities and documents that an analyst is studying.

Fundamentally, an analyst must read documents 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 [4], 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 5 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 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:

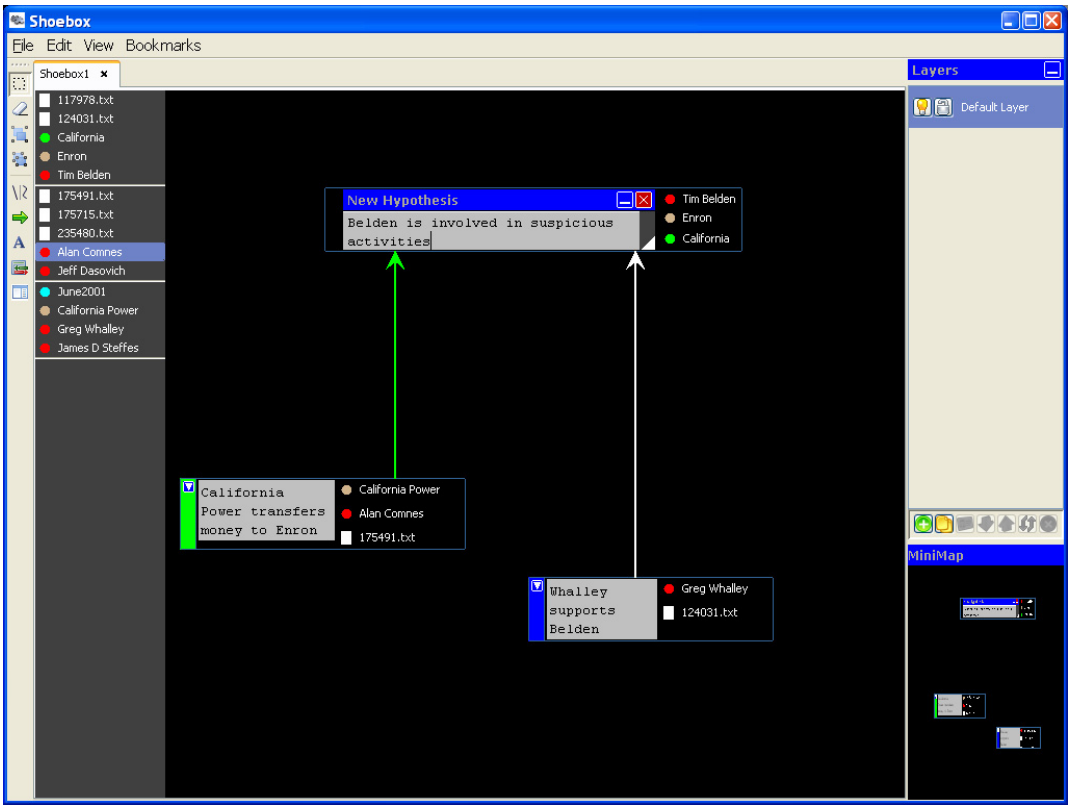

Figure 5. The Shoebox View showing the hypothesis, group, and link feature.

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.

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. We believe that this type of exploration may be useful in legal activities where sensemaking and analysis occur.

Jigsaw provides multiple visualizations of documents and the entities within them, as well as the connections that exist between entities and/or documents. 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.

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 [2].

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.

REFERENCES

1. CARD, S. K., MACKINLAY, J., AND SHNEIDERMAN, B. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

2. 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.

3. KLEIN, G., MOON, B., AND HOFFMAN, R. Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems 21*, 4 (July 2006), 70–73.

4. PIROLLI, P., AND CARD, S. Sensemaking processes of intelligence analysts and possible leverage points as identified through cognitive task analysis. In *2005 International Conference on Intelligence Analysis* (May 2005).

5. 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.

6. SPENCE, R. *Information Visualization*. ACM Press, 2001.

7. 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.

8. THOMAS, J. J., AND COOK, K. A. *Illuminating the Path*. IEEE Computer Society, 2005.