

Analyst's Workspace: Protecting Vastopolis

VAST 2011 Mini-Challenge 3 Award: "Novel Use of Large Screen Workspace to Support Analysis"

Christopher Andrews, M. Shahriar Hossain, Samah Gad, Naren Ramakrishnan, and Chris North

Virginia Tech

ABSTRACT

Analyst's Workspace is a sensemaking environment designed specifically for use of large, high-resolution displays. It employs a spatial workspace to integrate foraging and synthesis activities into a unified process. In this paper we describe how Analyst's Workspace solved the VAST 2011 mini-challenge #3 and discuss some of the unique features of the environment.

KEYWORDS: Visual analytics, large, high-resolution displays, intelligence analysis, space

INDEX TERMS: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces

1 OVERVIEW

Analyst's Workspace (AW) is a spatially oriented sensemaking environment designed specifically for use on a large, high-resolution display (Figure 1). The primary goal of AW is to provide a spatial workspace in which foraging and synthesis activities can be integrated into a single, fluid investigative process.

The primary visual objects in AW are documents, in which, due to the screen real estate provided by the large, high-resolution display, the full text is visible. Using a combination of manual and automatic layout, the analyst can arrange the documents spatially to schematize and otherwise externalize aspects of his or her synthesis of the information. Entities are another important aspect of the environment, and automatic named entity extraction is performed on the documents to generate the set of entities, which is editable by the analyst. The identified entities appear in several different guises within the interface (Figure 2).

Within documents, identified entities are underlined, colored by type (person, location, etc). If an entity seems important, the analyst clicks on it, and it opens into the workspace as a labeled list of all of the documents containing it. The analyst can then spider out from this point, opening these documents in turn. The list can also be iconified, providing a more abstract representation for the entity that can then be used to label document collections, or to create entity level structures such as social networks.

While the freeform workspace allows the analyst to use spatial relationships to encode arbitrary information, visual links are used to show computationally based relationships. Clicking on an entity, for example, shows links to every document that mentions that entity. The links are multi-scale, and can link to objects within documents, similar to the cross-application links described by Waldner [1]. Thus a link from an entity links not just to the document, but also to the actual mentions of the entity within. In order to keep the workspace uncluttered, only the links from the currently selected object are shown. Other facilities of the environment are textual highlighting, colored "sticky" notes, and

{cpa, msh, samah, naren, north}@vt.edu



Figure 1 Analyst's Workspace running on our 10,240 x 3200 large, high-resolution display.

full text search.

In preparation for this challenge, we also incorporated some data mining tools in AW. One of these tools was a heuristic based storytelling algorithm that we described in [2] to find connections between any pair of documents. Using tf-idf vector space representations of documents, the tool uses constraints on document similarities and neighborhoods to find a path between documents identified by the analyst. The second tool provides access to nearest neighbor documents through a cover tree data structure [3]. A third tool ("similarity network exploration") allows the analyst to find connections between entities based on co-occurrence, either at the sentence, or the document level.

2 ANALYSIS

In order to give a feel for working with AW, we will walk through some particulars of our analysis of the data. One of the significant challenges of a dataset of this size is identifying a place to start. As AW does not provide any capabilities for getting an overview of the dataset, we began by classifying the documents using the AlchemyAPI [4], which categorizes documents into news categories such as politics and sports. We then added a simple browser to AW to allow us to filter the documents by these categories.

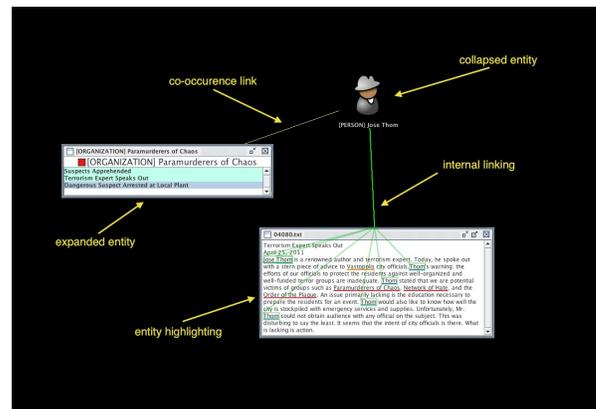


Figure 2 Three entity representations in AW: document underlines, list of documents (expanded), and icon (collapsed).

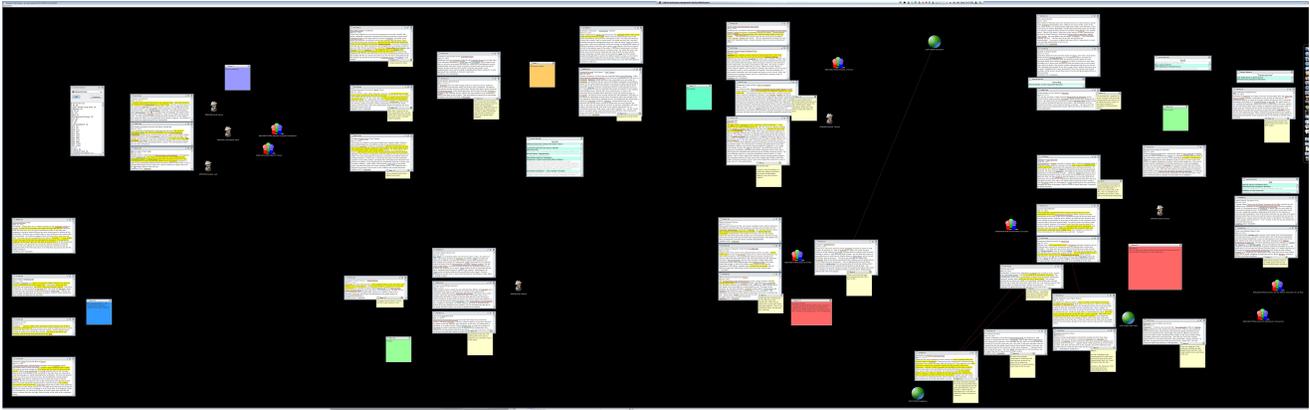


Figure 3 The final state of the workspace at the end of the investigation. [zoom in for details if viewing electronically]

We began by looking through the documents that were classified as “law and crime” or were considered “unclassifiable”, which effectively reduced the initial 4,474 documents down to 122. Using the browser, which provides a quick preview of the currently selected document, we rapidly paged through the documents in these categories, yielding eleven potentially interesting documents that we opened into the workspace for further investigation as we encountered them.

The most interesting of these documents was a report on local terrorist groups. We distributed the other documents to peripheral locations around the space where they would remind us to follow up on them later, as well as serving as seeds for new clusters if we found something relevant before we returned to them.

Clicking on the entities for the terrorist groups to open them into the workspace, we found that the only one we had any serious reports on was the Paramurders of Chaos. We opened the two associated reports, ordered them temporally in space, and labeled the new cluster with the Paramurders of Chaos entity. These documents offered no additional entities to follow, so we set the cluster to one side for more in depth investigation after we completed our initial trawl for suspicious events.

After collecting some more leads, we returned to this threat and began looking for a threat to the food supply, uncovering information about dying livestock and a warning about bioterrorism, which led to information about a bioweapon expert’s laboratory being robbed. Putting this into the timeline with the Paramurders documents makes it plausible that this is the same lab equipment found in the raid. Working on the theory that the food supply may be contaminated, we scanned the documents dated after the trespassing incident for any possible early indicators of something wrong. This led to information about the early flu season and dead fish in the Vast River. Opening the Vast River entity, we find an accident involving a food truck preceding the dying fish, forming another potential connection.

This process was repeated for each of the other events that we identified. We expanded all of the entities, performed relevant searches and used the neighborhood tool for further suggestions. Each sequence of events was ordered in columns with a network of the associated entities placed nearby. Events that shared an entity would be placed closer in the space. Each set of events was also labeled with current theories about the actual threat or avenues of investigation that would need to be explored.

Once we were confident that we had exhausted all avenues of investigation, the space was reordered again at the cluster level to arrive at the state seen in Figure 3. This is ordered roughly by threat level, with the primary bioterrorism threat on the right, with the threat level decreasing to the left. In addition, all material related to bombs have been separated out and are together at the

top of the workspace. In total, we spent about two and a half hours identifying the primary plots and laying them out and another two hours making sure there were no more leads to follow.

3 DISCUSSION

Interestingly, after the initial more than 30-fold pruning afforded by the Alchemy classifier, most of AW’s key foraging tools were of limited use for this dataset. While there were many entities in the dataset (we identified approximately 55,000), the actual threat documents contained very few, and involved minimal connectivity. Further, any connections between documents were easily ascertained using just the nearest neighbor tool, or were determined to be non-existent.

However, the large expanse of space afforded by AW was quite valuable in that it allowed us to marshal evidence for multiple simultaneous investigations as fragments of information were found. The presence of the full text served as external memory that could be rapidly consulted for connections when new evidence arose, and served as reminders of the unexplored avenues of investigation.

Another key point to make here is that the size of the workspace allows us to display all of our investigations – we did not close any when they seemed to lead to a dead end or a resolution. This served two key purposes. First, new evidence could be found that invalidated the initial conclusion (as is the case with the dead livestock). Second, the space now serves as a record of the investigative process. We can record not just what we suspect, but also what we found and rejected, which can play an important role in reporting of analytic results.

ACKNOWLEDGMENTS

This work is supported in part by US NSF grant CCF-0937133 and the Institute for Critical Technology and Applied Science (ICTAS), Virginia Tech.

REFERENCES

- [1] Waldner, M., Puff, W., Lex, A., Streit, M., and Schmalstieg, D.: Visual links across applications. In *GI 2010*, pp. 129-136, 2010.
- [2] Hossain, M.S., Andrews, C., Ramakrishnan, N., and North, C.: Helping Intelligence Analysts Make Connections. In *Proceedings of the AAAI '11 Workshop on Scalable Integration of Analytics and Visualization*, 2011. [to appear]
- [3] Beygelzimer, A., Kakade, S., and Langford, J.: Cover Trees for Nearest Neighbor. In *ICML '06* pp. 97-104, 2006.
- [4] AlchemyAPI. URL <http://www.alchemyapi.com>