Analysis Experiences Using Information Visualization

Elizabeth Hetzler and Alan Turner
Pacific Northwest National Laboratory

We present lessons learned from an observational study of the application of the In-Spire visually oriented text exploitation system in an operational analysis environment.

Professional analysts deal with a high volume of information and must constantly work to separate out the valuable data.[1] However, analysts have difficulty determining what data is useful without reading or skimming almost all returned documents from a search. This presents them with a difficult tradeoff. Searching information broadly returns hundreds or thousands of documents. Analysts spend a lot of time refining their queries, adding or excluding specific words or phrases, or using proximity searches. It takes time to work out syntax, and although search results are getting smaller—and hence more manageable—due to these query refinements, excluding valuable documents along with the useless ones is still a risk. Patterson et al.[2] document this behavior in a study of professional intelligence analysts who were asked to answer a specific set of questions in a limited amount of time. Many of the analysts excluded some or all of the high-value documents, as they progressively reduced their search results to a number of documents manageable enough to read.

This need to read all or most all documents has additional implications. As well as investigating specific questions and assignments, most analysts follow a particular issue over time. To help track these ongoing accounts, they often set up profiles or standing queries that constantly reflect the latest information available on that topic. Of course, the tradeoff between completeness and size still exists, and many analysts need rather complex profiles to help keep the daily results to a manageable size. This need to keep current with relevant matter leads to a tyranny of the inbox. Users have so many documents to read or skim as a matter of course that they have less time to step back and engage in more strategic activities. Analysts often don't have time to get an overview of their material, look for emerging patterns, and employ various alternative analysis techniques.

Under strong time constraints, analysts might adopt a strategy of satisficing. For information analysts, the documents are a means toward the goal of identifying, testing, and supporting most-likely hypotheses. The implication of satisficing is that analysts might focus prematurely on a likely hypothesis, which in turn could lead them to unduly weight evidence in support of that hypothesis and downplay evidence that contradicts it.[3]

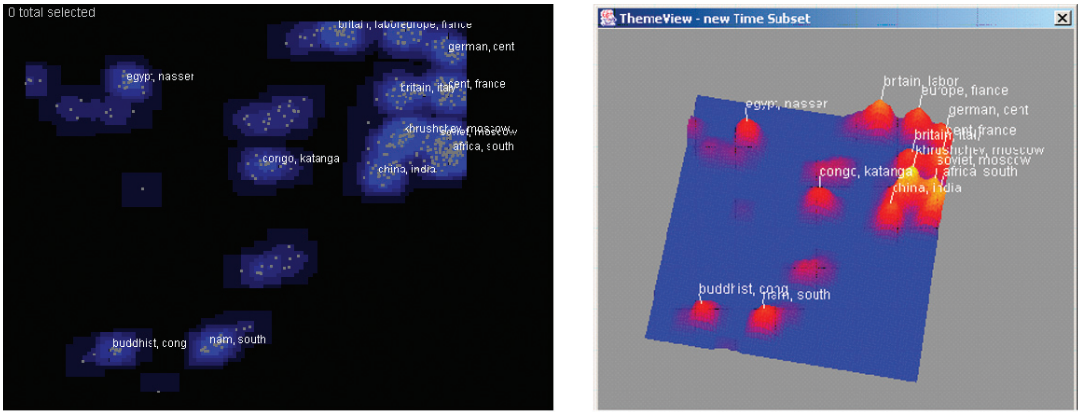
Approach

Information visualization has been defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition."[4] Many intriguing visualizations help deal with the data overload problem. Examples include visual tools based on similarity, relevance, explicit entity relationships, patterns in citations, and patterns in metadata (many excellent examples can be found elsewhere[4]). The literature describing creative visual approaches and prototype implementations is large and growing. However, little has been published on usage experiences with such tools.

The In-Spire visualization tool (based on the previous Spire system[5]) uses statistical word patterns to characterize documents based on their text content. Among its tools are two visualizations that show document themes, similarities, and differences: document-centric Galaxy uses a dot plot metaphor, and collection-centric ThemeView uses a landscape metaphor (see Figures 1a and 1b). We targeted several additional In-Spire analytic features explicitly at the user community.

Our goal in developing In-Spire was to provide a data visualization tool that let users deal with many documents, reducing the amount of time spent crafting queries and the chance of eliminating useful material before the user can see it. It also needed capabilities that let users understand a collection of material at an overview level first, instead of launching immediately into reading individual documents. The tool needed to support analytic demands for both convergent (focusing on a particular set of information, relationships, or thesis in depth) and divergent (seeking alternative

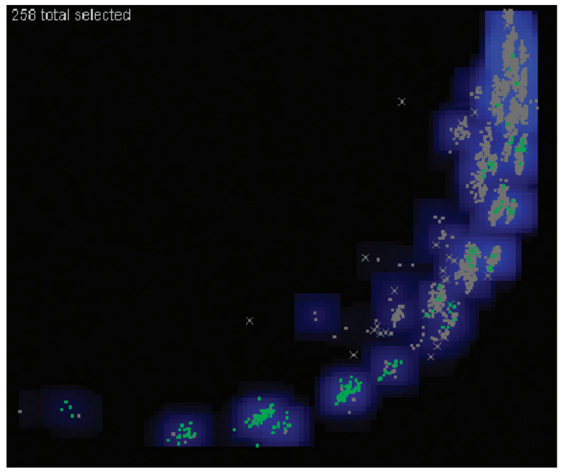
 Published by the IEEE Computer Society



(a)

(b)

1 (a) Galaxy uses position to show similarity. (b) ThemeView uses the same layout and adds a landscape metaphor to show density of text content.



2 Galaxy can be used to test effects of Boolean exclusion clauses.

explanations and remaining open to other potentially interesting information or relationships) understanding methods.

Figure 2 shows an application of In-Spire's approach to a user task. The results collection is a set of news documents from a search on Pakistan. The clumps of dots represent documents that are similar in content; clumps close to each other also have similar content. The clumps in the lower left of the figure are largely about the cricket championship matches in that region (we turned off labels for visual clarity).

An analyst using a traditional search tool might become distracted by the cricket documents and construct an exclusion clause to eliminate them—something like "not (cricket or wicket or champion)." The green dots in the image show which documents such a clause would exclude; not all of the cricket documents would go, but several of the noncricket ones would be lost as well, because of tangential mention of the particular words in the exclusion clause. This kind of visual interaction helped us in the design of In-Spire, providing an understanding of the impacts of query changes as well as some of the pitfalls of a Boolean search.

General observations

Our study involved more than 24 users representative of the overall user community, highly skilled and well trained in their disciplines. The users covered a range of ages, demographics, career experience, and specialized subdisciplines. We structured the user group to have three to four analysts from each of several distinct disciplines. Thus we heard several viewpoints of how our tool might apply to each discipline. We provided two 2-hour sessions of training on In-Spire, including how to apply it in analysis. More training time might have reduced some of the issues users experienced; however, some could only spare a limited amount of time for the study. The analysts used the visualization system to further their normal analytic tasks on their production issues and on the data of particular interest to them. They had a printed user guide, online help, and phone and email support options.

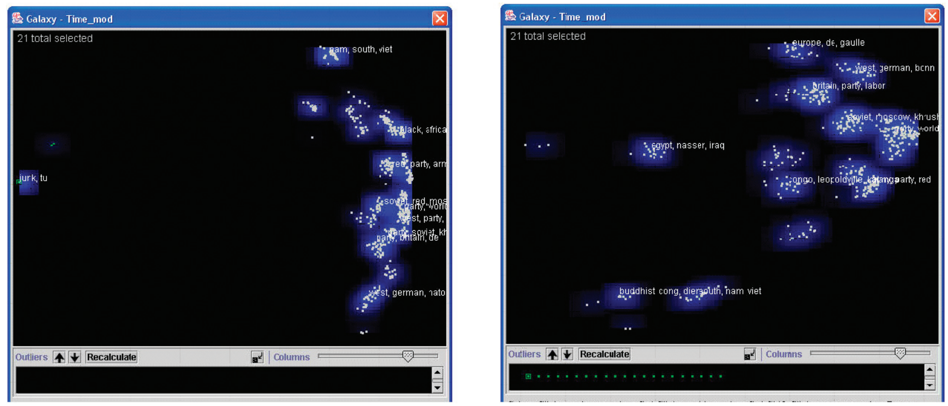
We asked participants to focus on whether or how the tool might provide analytical value, although we also encouraged usability feedback. The results led to the introduction of several new capabilities to In-Spire that more directly support the analytical process. We based our findings on a composite of direct observation in working with about half of the users, and on the results of interviews, questionnaires, and reported comments from all users.

In general, users understood the visual metaphors well. The concept of grouping by similarity made sense to them and provided a capability not present in their current tools. Several told us that the similarity of the clumps helped them more easily find relevant information. One used the clumps to organize her daily information so that she could read the documents more effectively. Some particularly liked the landscape metaphor of ThemeView for getting an overview of a document collection; again, other current tools provide no overview capability.

For some users, In-Spire acted as a mechanism to follow themes over time, a key element of many analytic tasks. The tool's time slicer capability lets users see how particular themes grow or shrink over time, and can show how the mix of themes in the Galaxy changes over time. Users felt that quickly seeing and exploring this timing information was particularly useful.

Several told us that the tool helped find documents they might have missed. One analyst experimented with a data set of more than 30,000 documents, much more than he would otherwise have attempted, and found several useful documents that he would not have found

3 (a) Users can temporarily move aside less interesting documents, such as the clump on the left, and (b) the remaining documents spread out to reveal new themes and associations.



(a) (b)

otherwise. He used the clustering capability, as well as the various query tools within In-Spire. One user tried an ambitious analysis, using more than 70,000 documents, experimenting with cluster and spatial locations, metadata encoding, and time trends. While the analysis did not yield the results expected, he still found it encouraging to have the ability to explore so many documents at once.

The tool seems to encourage different methods of interacting with information. Users were accustomed to a workflow of a complex search, sorting of results, refining the search, and then quick successive reads. In-Spire's visual clusters, theme abstractions, and searching and exploring within a tangible context encouraged different approaches. Some found this a strength. They told us that thinking about an issue in new and different ways was critical to some analytical tasks, and In-Spire supported this need better than current tools. Some found the change confusing and weren't sure how to apply the tool to their task. As a result, we've made several changes to the tool and the training method to help ease the adoption stage.

Differences in user data and focus area affected the benefit users experienced. Some had focused, well-understood tasks and data well suited to their current tools, hence they did not really need the additional capabilities provided by In-Spire. The messiness of production data—that is, real data containing formatting errors, grammatical errors, and so on—affected the tool's performance, making the similarity calculations less useful for some users. Adapting to the variety of data needed by users took considerable tuning, and has spawned significant follow-up research.

Analytical improvements

We added several major capabilities to the In-Spire tool, based on feedback from participants and as we learned more about the analytical process.

While analysts agreed with the risks of exclusionary search clauses, they still wanted a way to eliminate or move aside documents that probably weren't relevant to their issue. In a Galaxy view, such outlier groups sometimes tend to dominate the space, compacting too far the more relevant clumps. Further, analysts often use a progression of convergent and divergent methods. They might pursue a line of thought for a while, then back up and pursue an alternate line. Based on this need, we added capabilities to In-Spire that let users temporarily move aside documents, holding them for later consideration. Conceptually, we can compare this approach to the one described in Pirolli et al., which provides a textual list of clusters that change to match user convergent and divergent exploration.[6] Figures 3a and 3b illustrate the effects we incorporated into In-Spire. The user can select sets of documents, move them into an isolation area, then recalculate the Galaxy based on the remaining documents.

Algorithmically, the software reclusters and reprojects the remaining documents but does not change their underlying representation. The effect for the user is that the remaining documents spread out, with new themes emerging, and a faster and more fluid process than creating a subset. The user can still search and read set-aside documents and move them back into the Galaxy at any time. This capability makes the visualization more alive, gives the user a sense of more control over the visualization, better reflects analyst progress and current focus, and serves to adapt the information presentation to convergent thinking patterns.

To encourage analysts to broaden their queries, we added a capability to show users which documents would have resulted from their original narrow query and which were extra. For this, we needed an exact comparison. Rather than duplicate the external search capabilities in In-Spire, we leveraged other, preexisting tools. Users could enter both a narrow and a broad query into an outside search tool and send both results to In-Spire. In-Spire can show document membership in these result lists at any time through color codes (see Figure 4). This capability lets users easily find new documents thematically close to ones found in their familiar searches. This new feature was not so much an improvement to In-Spire as it was a feature of the interface between In-Spire and the other tools that analysts used.

Some analysts told us that only 5 percent or less of their search hits are useful. Further, a paragraph or two of useful information might be embedded in a document that covers different subjects overall. In such circumstances, visual representations based on dominant themes might not reveal themes and clusters relevant to the users. It became clear that users needed to specify particular concepts and themes of interest that should be given more weight in the visualization than they would otherwise. We have started research on this need.

Research to deployed tool

Some of the lessons learned from this experience highlight mismatches between current research and user needs that make it difficult for systems to bridge the gap to usable and deployable systems. A recent book[7] and a keynote address, "Crossing the Information Visualization Chasm: From Innovation to Adoption," by Ben Shneiderman at the 1999 Information Visualization Symposium have discussed variations on this theme. Information exploitation tools for information analysts

- should fit the analysts' methods for using data,
- should support more of the analytical flow than current tools, and
- might need considerable tuning when applied to real data (for those tools developed using research data).

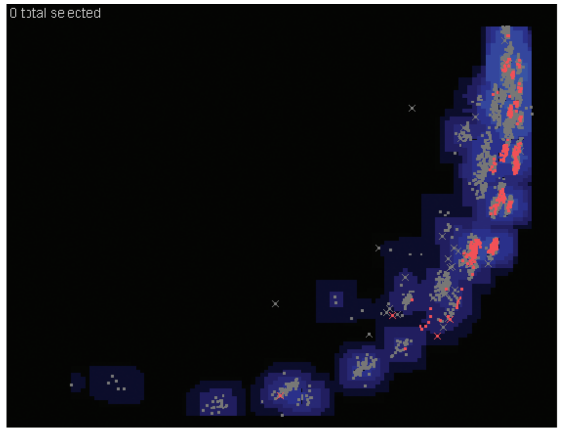
Analysts use data in particular ways. Many research tools operate on a fixed set of data. A static set is often necessary as a starting place for common evaluation or for duplication of results. However, this can lead to a focus on bucket-of-data approaches. Professional analysts follow an issue over a period of time. A search tomorrow will yield additional documents—perhaps only a handful, perhaps hundreds more. In either case, users need to easily add the new documents to an ongoing exploration. If the visualization must be recomputed each time, they lose valuable work and time. In addition, the users' perspective on the issue might change over time, meaning that their search criteria will evolve along with their decisions about which documents are relevant. These tools must be usable on such evolving collections.

Main tasks in the analytic flow for analysts include

- Research or data monitoring. They need to find relevant information.
- Analysis. They employ various techniques along with expert judgment to reach relevant conclusions.
- Draft and edit written reports. They must document conclusions and supporting evidence.

One primary goal of many exploitation tools is to make the research task more effective and efficient. This was also a main goal for In-Spire. However, as we worked with users, it became apparent that such tools could provide even more value by helping users transition their work into analysis and reporting tasks. Support for the analytic thinking and reporting phases—such as problem organization, printing, evaluative summaries, or exporting key documents into other tools—might seem less central to information exploitation researchers, but are key to acceptance into a user community and an excellent opportunity for complementary research agendas.

Analysts typically use data from a variety of sources to improve the quality of their analysis and products.[1] Text documents might contain a variety of field formats.



4 System interface between In-Spire and the query tool lets users see the effects of broadening a query.

Metadata might have XML, field tag, position, or syntax marks. These might appear at the beginning or the end or be interspersed throughout the text content. Normal grammar and style rules might not apply. Misspellings or alternate spellings are common. Acronyms and abbreviations abound. Further, analysts might quickly need to access new data sources, with minimal time for source-specific tuning. Useful tools must be flexible and agile when applied to real data and must minimize the task of ingesting production data.

Conclusions

Information visualization can provide significant benefits for professional analysts. New helpful capabilities might result from new tool features and from tuning the interfaces between tools. However, the leap from an intriguing research system to a useful, deployed tool is difficult to achieve. Developers must address a variety of pragmatic challenges, such as data idiosyncrasies. They need to be sensitive to user work patterns and timing. Capabilities such as finely tuned bridges to other user tools and support for other portions of the analytic process might prove essential to user acceptance, even though these are not the tool's technical focus. ■

Acknowledgments

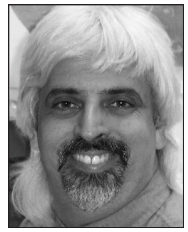
We thank Russ Rose for his guidance and vision and the In-Spire pilot team for their help and feedback. We gratefully acknowledge the Pacific Northwest National Laboratory visualization team. The Battelle Memorial Institute manages the Pacific Northwest National Laboratory for the US Department of Energy under contract DE-AC06-76RL1830. The US government supported this work.

References

1. M.M. Lowenthal, *Intelligence: From Secrets to Policy*, CQ Press, 2000.
2. E.S. Patterson et al., *Aiding the Intelligence Analyst in Situations of Data Overload: From Problem Definition to Design Concept Exploration*, tech. report ERGO-CSEL 01-TR-01, Inst. for Ergonomics/Cognitive Systems Engineering Lab., Mar. 2001.

3. R.A. Heuer, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, 1999.
4. S. Card, J. MacKinlay, and B. Shneiderman, eds., *Readings in Information Visualization*, Morgan Kaufmann, 1999.
5. J. Wise et al., "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," *Proc. IEEE Symp. Information Visualization* (InfoVis), 1995, IEEE CS Press, pp. 51-58.
6. P. Pirolli et al., "Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection," *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems* (CHI), ACM Press, 1996, pp. 213-220.
7. G. Moore, *Crossing the Chasm*, Harper Collins, 2002.

***Elizabeth Hetzler** is a chief scientist and project manager in the Information Analytics group at the Pacific Northwest National Laboratory. Her research interests include visual interaction, text and information analysis, human–computer interaction, and homeland security. Hetzler has a BA in math and French from Vanderbilt University and an MS in computer science from the Washington University in St. Louis.*



***Alan Turner** is a chief scientist at Pacific Northwest National Laboratory. His research interests include visual information exploitation, rich interaction environments, multimedia storytelling, and multithreaded narrative. Turner has an MA in computer science from the University of Missouri, Rolla. He is an IEEE and an ACM member and serves on the* CG&A *editorial board.*

Readers may contact Elizabeth Hetzler at Information Sciences and Engineering, Pacific Northwest Nat'l Lab., P.O. Box 999, Richland, WA 99338; beth.hetzler@pnl.gov.

For further information on this or any other computing topic, please visit our Digital Library at http://www.computer.org/publications/dlib.